Analyzing Biomarker Discovery: Estimating the Reproducibility of Biomarker Sets

Amir Forouzandeh^{1,*}, Alex Rutar², Sunil V Kalmady^{1,3}, and Russell Greiner^{1,4}

¹Department of Computing Science, University of Alberta, Edmonton, T6G 2E8, Canada

²Department of Pure Math, University of Waterloo, Waterloo, ON N2L 3G1, Canada

³Canadian VIGOUR Centre, University of Alberta, Edmonton, T6G 2E1, Canada

⁴Alberta Machine Intelligence Institute, Edmonton, T6G 2E8, Canada

*forouzan@ualberta.ca

ABSTRACT

Many researchers try to understand a biological condition by identifying *biomarkers*. This is typically done using univariate hypothesis testing over a labeled dataset, declaring a feature to be a biomarker if there is a significant statistical difference between its values for the subjects with different outcomes. However, such sets of proposed biomarkers are often not reproducible – subsequent studies often fail to identify the same sets. Indeed, there is often a very small overlap between the biomarkers proposed in pairs of related studies that explore the same phenotypes over the same distribution of subjects.

This paper first defines the *Reproducibility Score* for a labeled dataset as a measure (taking values between 0 and 1) of the reproducibility of the results produced by an arbitrary fixed biomarker discovery process for a given distribution of subjects. We then provide ways to reliably estimate this score by defining algorithms that produce an over-bound and an under-bound for this score for a given dataset and biomarker discovery process, for the case of univariate hypothesis testing on dichotomous groups. We confirm that these approximations are meaningful by providing empirical results on a large number of datasets and show that these predictions match known reproducibility results. We have also created a publicly available website, hosted at <https://biomarker.shinyapps.io/BiomarkerReprod/>, that produces these *Reproducibility Score* approximations for any given dataset (with continuous or discrete features and binary class labels).

1 Introduction

Improved understanding of a disease can lead to better diagnosis and treatment. This often begins by finding “biomarkers”, which is a generic term referring 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”¹. Typically, these are individual features (*e.g.*, expression values of specific genes^{2,3}) that follow different distributions (*e.g.*, have different mean values) in diseased patients versus healthy controls.

Sometimes, biomedical researchers can identify candidates for biomarkers based on their existing knowledge of the disease etiology and/or cellular pathways. This is done by seeking features that are causally related to the disease (*e.g.*, phenylketonuria is caused by mutations in a single gene PAH⁴) or a symptom of it (*e.g.*, Hemoglobin A1C for monitoring the degree of glucose metabolism in diabetes⁵). This paper, however, focuses on the usage of statistical tools to discover and evaluate biomarkers. A typical dataset can be viewed as a matrix whose rows each correspond to a subject (say a person) and each column corresponds to a feature (*e.g.*, clinical measure, or the expression value of a gene), and with a final column providing the outcome (*e.g.*, a binary outcome distinguishing case versus control); see Figure 1.

These “biomarker discovery studies” (also known as “association studies”) then attempt to determine which of the features (columns) differ significantly among distinct outcomes. Two standard examples of such studies are the “Genome Wide Association Study” (GWAS), over a set of SNPs⁶, and the “Gene Signature Study”, over gene expression values⁷. Typically, this involves first computing a representative statistic for each feature (*e.g.*, for continuous entries, running a t-test based on the mean and variance of the case versus control), and then declaring a feature to be a biomarker if the resulting MCC-corrected *p*-value is below 0.05⁸. Here, we use MCC (Multiple Comparison Corrections) as a generic term, which includes both False Discovery Rate (FDR) correction, and Family-Wise Error (FWE) Correction. (Appendix A.2 discusses some of the subtleties here, especially with respect to features.) In some situations, the researchers then validate these biomarkers using a biological or medical process (*e.g.*, based on knock-out or amplification studies^{9,10}). Other studies validate the proposed biomarkers based on existing biological knowledge. A third class of projects instead use the potential biomarkers to create a computational model – perhaps to learn a classifier^{11–13} – and then measure that down-stream model (perhaps based on its accuracy on a held-out set) and declare the biomarkers to be useful if that model scores well.

Figure 1. Data matrix, showing t-test p-values for each (shown) feature for the GSE 7390 dataset¹⁴, with respect to the group outcome (here “Metastasis” for breast cancer); the circled features, with $p < 0.05$, are (purported) biomarkers. For notation: We will refer to each of the first r columns of the matrix as a “feature”; these are often called “(independent) variables”. We refer to the final column as a “outcome” – e.g., case versus control – these are often called “labels”, “dependent variables”, “groups”, “phenotypes” or “classes”. Finally, we will use “subject” to refer to each row of that matrix; these are sometimes called “instances” or “samples”.

A great many papers, however, simply publish the list of purported biomarkers without providing validation for this set; see Appendix ???. This paper focuses specifically on this case. We address this limitation by providing a falsifiable (statistical) claim about such sets of biomarkers, which suggests a way to validate the proposed biomarker sets.

While some biomarkers are causally related to the associated outcome, this can be difficult to establish (often requiring instrumented studies¹⁵); but fortunately, in many situations, 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 ...) alone supports that disease. This argues that a proposed biomarker is good if it was *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.*² (resp., Wang *et al.*³) reported signatures with 70 (resp., 76) genes, these two sets had only 3 genes in common. Ein-Dor *et al.*¹⁶ notes this in another situation: “Only 17 genes appeared in both the list of 456 genes of Sorlie *et al.*¹⁷ and the 231 genes of van’t Veer *et al.*²; merely 2 genes were shared between the sets of Sorlie *et al.* and Ramaswamy *et al.*¹⁸. Such disparity is not limited to breast cancer but characterizes other human disease datasets (Lossos *et al.*¹⁹) such as schizophrenia (Miklos and Maleszka²⁰)”. Many others^{21–23} report similar findings. Indeed, Begley and Ellis²⁴ report that only 6 of 53 published findings in cancer biology could be confirmed; which Wen *et al.*²⁵ notes is “a rate approaching an alarmingly low 10% of reproducibility”. Moreover, a 2016 Nature survey²⁶, of over 1500 scientists, found that 70% of researchers have tried but failed to reproduce another scientist’s experiments, and 52% thought there was a significant ‘crisis’ of reproducibility.

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 attempts to distinguish case from control, then the two sub-populations should only differ in a single characteristic. Unfortunately, matching cases and controls over all possible features is often not achievable.
2. A second issue is with the precise notion of what “*reproducible*” means. Is it a property of a *specific biomarker*, or of a *set of biomarkers*? There is no clear choice for an optimal objective measure. This is especially problematic when dealing with multi-factorial diseases, where the outcomes corresponds to a disjunction over many sub-diseases²⁷.
3. A final important issue is the impact of *sample size*. Many studies have a relatively low number of subjects, which increases the probability of finding both false negatives and false positives.

Our analysis assumes that the researchers have addressed (1) by running carefully designed, well-specified studies. Further, we also assume that there is no uncertainty in the outcomes 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 measurement varies with sample size (3).

In this paper, we assume that biomarkers are 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 enrichment²⁸) or networks of genes associated with high loadings of principal component and univariate Pearson correlation values (see PC-corr²⁹); but we are not considering *learning* combinations. By a *Biomarker Discovery* process $BD(\cdot)$, we mean a function that takes as input a data matrix of n subjects over a set of r features and identifies a subset of proposed biomarkers; see Figure 1. We will formally define a *Reproducibility Score*, $RS(D, BD)$, to quantify the “reproducibility” of the set of proposed biomarkers produced by the biomarker discovery process: *viz.*,

$$\begin{aligned} &\text{the average Jaccard score between these proposed biomarkers } BD(D), \\ &\text{and those produced by running the same } BD \text{ process} \\ &\text{over another comparable dataset drawn from the same distribution} \end{aligned} \quad (1)$$

where two datasets are comparable if they have the same number of subjects from each outcome. In order to estimate the reproducibility score in practice, we construct two approximations: an overbound and an underbound. We then provide empirical tests over many datasets, with a focus on t -tests as the main biomarker discovery process. We provide many examples for both microarray data (with continuous values) and SNP data (with discrete values) to provide practical evidence for the effectiveness of these approximations. Researchers can use this framework to estimate the reproducibility of the potential results of their biomarker discovery study. A low reproducibility score suggests that these biomarkers may not be accurate, potentially because the dataset used is too small, the dataset is too heterogeneous, or the biomarker discovery algorithm is not suitable for the dataset. We have also produced a publicly available website, hosted at <https://biomarker.shinyapps.io/BiomarkerReprod/>, that, given a labeled dataset, computes these estimates of the Reproducibility Score, with respect to any of a variety of biomarker discovery algorithms.

Outline: Section 2 formally defines the Reproducibility Score (RS) and describes the challenges of estimating this measure. We then define two approximations for RS: an overbound and an underbound. It also describes some of the standard biomarker discovery algorithms. Section 3 describes extensive empirical studies over many datasets – microarray and mRNAseq data (with continuous values) and SNP data (with discrete values), focusing on a standard Biomarker Discovery process $BD(\cdot)$, to confirm the effectiveness of these approximations. Section 4 summarizes some future work and the contributions of this paper. In the Supplementary Information, Appendix A discusses various notes: a glossary of the various technical terms used, how Biomarker Discovery differs from standard (supervised) Machine Learning, how a combination of a pair of features might be important for a predictive task, even if neither, by itself, is important (towards explaining why it can be so difficult to find biomarkers); etc. Appendix B presents results from other empirical studies, which explore how the RS varies with the type of MCC correction used (including “none”), the p -value threshold, the size of the dataset and the number of iterations of the approximation algorithms. Finally, the approximations we present are motivated by two heuristics (Heuristics 7 and 9). Appendix C presents arguments that motivate these heuristics, and also provide additional empirical evidence that support them.

We close this section by motivating the need for an objective measure for evaluating the quality of a set of biomarkers (Subsection 1.1), then overviews 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 (Subsection 1.2).

1.1 Motivation for Evaluating Biomarker Sets

To motivate the need for evaluating association studies, consider first *predictive studies*, which use a labeled dataset, like the one shown at the top of Figure 1, to produce a predictive model (perhaps a decision tree, or a linear classifier) that can be used to classify future subjects – here, into one of the two classes: Y or N. In addition to the learned classifier, the researchers will also compute a *meaningful estimate of its quality* – *i.e.*, of the accuracy (or AUROC, Kappa Score, etc.) of this classifier on an independent hold-out set³⁰, or the results of k -fold cross-validation over the training sample.

By contrast, many association studies report only a set of purported biomarkers, but provide no falsifiable claim about the accuracy of these biomarkers. Many meta-reviews claim that a set of biomarkers is problematic if they are not reproduced in subsequent studies^{16,31–33}. Given that biomarkers should be reproducible, we propose evaluating a biomarker set based on its reproducibility score. An accurate estimate of this score can help in at least the following three ways:

1. Researchers can compare various different “comparable” biomarker discovery algorithms to see which produces the biomarker set that is most reproducible. Here, “comparable” corresponds to the standard practice of only considering discovery tools that impose the same criterion, such as the same p -value, or only considering features that exhibit the same minimum fold-change. This type of analysis may help to determine errors within the biomarker discovery process. Moreover, we will see that MCC-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.

2. A low reproducibility score suggests that few of the proposed biomarkers will be found in another dataset, and highlights the potential that these proposed biomarkers may not be accurate. This could motivate researchers to consider a dataset that is larger, to focus on a more homogenous population, or perhaps consider another biomarker discovery technique.
3. Finally, there are many meta-reviews^{21,34,35} that note the lack of repeatability in many biomarker discovery papers, and question whether the techniques used are to blame. 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. 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 biomarker-discovery reproducibility score will inform the reader whether to expect another study, on a similar dataset, will find many of the same biomarkers. Note that we should view the reproducibility score as necessary for considering a proposed model, but not sufficient – *i.e.*, it might rule-out a proposed discovery model, but should not be enough to rule-in a model.

For these reasons, we provide an easy-to-use, publicly available webapp <https://biomarker.shinyapps.io/BiomarkerReprod/> that anyone can use to produce meaningful estimates of the reproducibility of a set of biomarkers.

1.2 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^{16,33,35}, and others exploring specific examples^{8,32}. These papers suggest different reasons for the problem, such as the heterogeneous biological variations in some datasets^{16,33} or problems in the methods used that may lead to non-reproducible results^{36,37}.

In particular, Zhang *et al.*³³ 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.*¹⁶ 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 outcomes, but the empirical correlations change for different (sub)samples of instances. These two papers motivate our need for tools that can effectively estimate the reproducibility – such as the ones presented here.

Several projects^{35–37} have attempted to formally analyse this problem. Ein-Dor *et al.*³⁵ 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 Subsection 3, especially Figure 4.

The goal of the MicroArray Quality Control (MAQC) project³⁶ 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 (especially with 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.*³⁷ found similar results by using the same procedures for another dataset. However, Klebanov *et al.*³⁸ 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 (think $p = 1.0$), is completely reproducible. 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 (sensitivity). 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.

Our approximation algorithms use a type of re-sampling to bound reproducibility. Below we summarize several other studies that similarly deal with re-sampling and biomarker discovery. Some studies provide ways to better estimate the true statistical significance, but do not provide a framework for evaluating empirical reproducibility – *e.g.*, Gagno *et al.*³⁹ used bootstrap resampling to estimate 95% confidence intervals and *p*-values for an internal assessment of their findings (related to breast cancer survival), Chitpin *et al.*⁴⁰ proposed a resampling-based method to better estimate the false discovery rate in chromatin immunoprecipitation experiments, and Pavelka *et al.*⁴¹ proposed a resampling-based hypothesis testing algorithm that provides a control of the false positive rate for identification of differentially expressed genes. Furthermore, Alshawaqfeh *et al.*⁴² and Zhao and Li⁴³ suggested methods for consistent biomarker detection in high-throughput datasets where evaluation was based on common biomarkers among the two resampled sets – *i.e.*, they are considering the false positives but not false negatives. (By contrast, our use of Jaccard score involves both.) Other studies had different goals – *e.g.*, Ma *et al.*⁴⁴ used resampling in a

permutation test, to evaluate the predictive power of the identified gene set based on the accuracy of down-stream classification task; recall however that our goal is to evaluate the reproducibility of the biomarkers directly (not downstream). Filosi *et al.*⁴⁵ propose methods for evaluating the stability of reconstructed biological networks in terms of inference variability due to data subsampling; we however are focusing on the reproducibility of the individual biomarkers. Hua *et al.*⁴⁶ conducted a simulation study to compare the *ranking performance* of several gene set enrichment methods; by contrast, our approach considers the SET of biomarkers, not the ranking, and is over several real-world datasets (not just simulated ones). Note that none of these used re-sampling techniques to bound the expected replicability of the set of biomarkers found by some discovery algorithm, nor to demonstrate the validity of those bounds.

2 Methods

2.1 Formal Description

As illustrated in Figure 1, a “Biomarker Discovery” algorithm, $\text{BD}(\cdot)$, takes as input a dataset D of n subjects, each described by r features $F = \{f_1, \dots, f_r\}$ and labeled with a binary outcome, and returns a subset $F' \subset F$ of purported biomarkers.

Typically, each $f \in F'$ differs in some significant way between each class. To be more precise, let x_i^j be the value of the i^{th} feature of the j^{th} subject, and $\ell^{(j)}$ be the outcome of the j^{th} subject (which is either + or -). Then a class difference means that the set $\{x_i^j \mid \ell^{(j)} = +\}$ of values of the i^{th} feature of the diseased individuals is significantly different from the values of that feature over the healthy individuals, $\{x_i^j \mid \ell^{(j)} = -\}$. For simplicity, 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 (such as gender, or the genotype of a SNP). Subsection 2.2 below will describe several such biomarker discovery algorithms.

As noted in Equation 1, the *Reproducibility Score* $\text{RS}(D, \text{BD})$ quantifies the “reproducibility” of the set of proposed biomarkers $\text{BD}(D)$ corresponding to running the biomarker discovery algorithm over the labelled dataset D . Here, we assume that the values of each feature x_i^j , for each outcome c , are generated independently from a fixed distribution (*i.e.*, “i.i.d.”)

$$p_{i,c}(v) = P(x_i^j = v \mid \ell^{(j)} = c).$$

Here and in general, we use $P(\cdot)$ to refer to either a probability density for continuous variables, or a probability mass for discrete variables. Note these are just the marginal distributions: we do not assume that the various features are independent from one another. – *i.e.*, this does not necessarily correspond to Naive Bayes³⁰. We will view $\vec{p}(\cdot) = [p_{i,c}(\cdot)]_{i,c}$ as the matrix of these $r \times 2$ different distributions, and let $\vec{p}^{[n_+, n_-]}(\cdot)$ be the distribution for sampling $n = n_+ + n_-$ instances independently from this distribution, where $n_+ \in \mathbb{Z}^+$ instances are drawn from the distribution $[p_{1,+}(\cdot), \dots, p_{r,+}(\cdot)]$ associated with positive outcomes, and similarly $n_- \in \mathbb{Z}^+$ instances from the distribution $[p_{1,-}(\cdot), \dots, p_{r,-}(\cdot)]$ associated with negative outcomes. Then for datasets $D', D'' \sim \vec{p}^{[n_+, n_-]}(\cdot)$ sampled independently, we define

$$\text{RS}^*(\vec{p}(\cdot), [n_+, n_-], \text{BD}(\cdot)) = \mathbb{E}[J(\text{BD}(D'), \text{BD}(D''))] \quad (2)$$

where the Jaccard score of two sets

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (3)$$

is the ratio of the intersection to the union of these sets – hence $J(A, B)$ ranges from 0 to 1, and is 1 if and only if $A = B \neq \{\}$, and is 0 if and only if these sets are disjoint. (We define this to be 0 if $A = B = \{\}$.) Note that the Jaccard score is only one possible measure to evaluate the degree of overlap of gene signatures; Appendix B.4 discusses a slightly different measure that is sometimes used to measure the reproducibility of a set of biomarkers. See also Shi *et al.*⁴⁷ for a comprehensive overview of these measures.

Of course, we do not know $\vec{p}(\cdot)$, and so we use an empirical distribution $\widehat{p}_D(\cdot)$, determined based on context from the dataset D , to produce the approximation

$$\text{RS}(D, \text{BD}) = \text{RS}^*(\widehat{p}_D(\cdot), \langle D \rangle, \text{BD}(\cdot)) \quad (4)$$

that estimates the reproducibility of the biomarker set $\text{BD}(D)$, where the notation $\langle D \rangle = [|D^+|, |D^-|]$ refers to the pair of sizes of the positive and negative subjects in D – corresponding to $[n_+, n_-]$. Note that this reproducibility score deals with the *sets* of biomarkers that are produced by the $\text{BD}(\cdot)$ function, and not any specific biomarker.

Of course, Equation 4 suggests the obvious bootstrap sampling algorithm⁴⁸. Empirically, however, we found that it did not perform well (see Appendix B.1) – motivating the algorithms described in Subsection 2.3.

2.2 Biomarker Discovery Algorithms: $BD(\cdot)$

We now discuss various approximation algorithms for biomarker discovery. As our goal is to illustrate the reproducibility issues with respect to the standard approach, we focus on that standard approach: where the biomarker discovery is based on independent two-sample t -tests, perhaps with family-wise error correction (see below). Note, however, that the methodology used here is completely general, and can easily extend to any other class of biomarker discovery process; see Limitations in Section 4.

Recall that we considering 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 $BD(\cdot)$ algorithms described here, to convert each categorical value to a real number: here converting 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. This allows us to view each such dataset as one with continuous values.

We assume that the real values of each feature, for each outcome, follows a normal distribution, which might be different for the different outcomes, and so we use an independent two-sample t -test for all of our empirical experiments. Recall that the test statistic is given by

$$t = \frac{\bar{X}_+ - \bar{X}_-}{\bar{s}_p \cdot \sqrt{\frac{1}{n_+} + \frac{1}{n_-}}} \quad \bar{s}_p = \sqrt{\frac{(n_+ - 1) \bar{s}_+^2 + (n_- - 1) \bar{s}_-^2}{n_+ + n_- - 2}}. \quad (5)$$

where n_+ and n_- are the number of instances with positive and negative outcomes, respectively, and with empirical means $\bar{X}_-$ and $\bar{X}_+$ and empirical variances $\bar{s}_+^2$ and $\bar{s}_-^2$.

Note the biomarker discovery process essentially performs a single statistical test for each feature. As the number of features is often large – often tens-of-thousands, or more – many projects sought ways to reduce the chance of false discoveries; a standard way to do this is through some MCC correction. We therefore consider biomarker discovery algorithms described as $BD_{t,p,\chi}(D)$, where the t in the subscript refers to the 2-sided t -test, the p for the p -value used, and χ to the MCC correction. Our canonical example is $BD_{t,0.05,BH}(D)$, with $p=0.05$, and $\chi=BH$ to the Benjamini/Hochberg correction⁴⁹. 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. See Appendix B.2 for more details.

2.3 Algorithms that Approximate the Reproducibility Score

As we have the dataset D with $n \approx [n_+, n_-]$ labeled instances, we can directly compute $BD(D)$. To compute $RS(D, BD(\cdot))$, however, we also need to produce one (or more) similar datasets D' , each with $[n_+, n_-]$ subjects drawn from the same (implicit) distribution $\bar{p}(\cdot)$ that generated D (with the same number of positive and negative instances), but which is presumably disjoint from D . While we do not have such D' 's, and so cannot directly compute the Reproducibility Score, we show below how to compute an overbound and an underbound of $RS(D, BD(\cdot))$.

2.3.1 Overbound: oRS

The $oRS(D, BD(\cdot), k)$ procedure produces (an estimate of) an overbound for $RS(D, BD(\cdot))$, by making it *easier* for a feature to be selected to be in both purported biomarker sets. In this algorithm, D is our given fixed dataset, $BD(\cdot)$ is the given biomarker discovery algorithm, and k is a parameter to determine the number of trials when computing the overbound. (Here, and below, see the Glossary in Appendix A.1 for a summary of the algorithms and their arguments.) The oRS algorithm first defines a size- $2n$ dataset DD that contains two copies of each subject in D , of course with the same outcome both times. It then randomly partitions this DD into two disjoint size- n datasets D_1 and D_2 , balanced by outcome. Here, each partition is with respect to the *list* of elements, so it will include duplicates. To insure that the resulting datasets are balanced, oRS first split DD into DD^+ and DD^- , where DD^+ are the cases and DD^- the controls. It then forms D_1^+ by randomly drawing $1/2$ of DD^+ , and D_1^- by randomly drawing $1/2$ of DD^- , then merges $D_1 = D_1^+ \cup D_1^-$; see Figure 2. The dataset D_2 is then formed from the remaining subjects of DD not included in D_1 . oRS then runs $BD(\cdot)$ on the datasets D_1 (resp., D_2) to produce two sets of biomarkers, and computes the Jaccard score for this pair of biomarker sets: $J(BD(D_1), BD(D_2))$. It then repeats this double-split-BD-Jaccard process k times, then returns the average of these k values:

$$oRS(D, BD(\cdot), k) = \frac{1}{k} \sum_{i=1}^k J(BD(D_{1r}), BD(D_{2r})) \quad (6)$$

where each dataset pair $[D_{1r}, D_{2r}]$ is created independently using the above procedure.

For each r , as we expect D_{1r} to overlap with D_{2r} , it is relatively likely that any D_{1r} -biomarker will also be a D_{2r} -biomarker (more likely than if D_{1r} was disjoint from D_{2r}), which means we expect the associated Jaccard score to be higher. This follows

Figure 2. Diagram showing oRS's process of generating pairs of subsets for a dataset D (k times), then using those to compute $\text{oRS}(D, \text{BD}(\cdot), k)$.

from the heuristic that, as two datasets have more common elements, we expect the number of biomarkers common to two datasets, to increase – *i.e.*,

$$\begin{aligned} &\text{if } A_1, A_2, B_1, B_2 \sim \bar{p}^{[n_+, n_-]}(\cdot) \quad \text{and} \\ &\quad |A_1 \cap A_2| \quad \text{is larger than} \quad |B_1 \cap B_2|, \\ &\text{then we expect } J(\text{BD}(A_1), \text{BD}(A_2)) \quad \text{will be larger than} \quad J(\text{BD}(B_1), \text{BD}(B_2)) \end{aligned} \quad (7)$$

ceteris paribus. Here, we view each of $\{A_1, A_2, B_1, B_2\}$ as a set of $n = n_+ + n_-$ r -dimensional instances. Note this relationship is simply a heuristic to motivate the algorithm – one that we expect to hold in practice. Appendix C provides some basic arguments, and empirical evidence, to support this claim.

We close with three quick observations:

1. **Expected Overlap:** We expect 50% of the instances to be duplicated in any given pair of datasets. See Lemma 2 in Appendix C.4.
2. **Relation to Bootstrap Samples:** Here, each subject occurs exactly twice in each pair of datasets $D1_r$ and $D2_r$. If we instead used bootstrap sampling (called bRS below), we expect many subjects would occur more often in the pair of datasets. (See Lemma 3 in Appendix C.4.) Given Heuristic 7, this means we expect bRS's Jaccard score here to be higher than for oRS's; as oRS is already an overbound for the true score (RS), this means bRS would be a worse bound, as $\text{bRS} \geq \text{oRS} \geq \text{RS}$. This is why we use our “doubling approach” oRS rather than bootstrap sampling bRS, as oRS produces values that are smaller, but still remains an overbound, as desired. See Figures 7 and 8 and Appendix B.1.
3. **Relation to RS*:** Our oRS approach is clearly related to RS* (Equation 2), as both compute the average Jaccard score of pairs of size- n datasets sampled from a distribution. They differ as (a) RS* uses the true distribution $\bar{p}(\cdot)$, while oRS uses only the estimate $\hat{p}_D(\cdot)$, (b) RS* is the *true* average while oRS is just the *empirical* average over k trials, and (c) RS* will draw *independent* datasets, but the oRS datasets will overlap.

2.3.2 Underbound: uRS

The $\text{uRS}(D, \text{BD}, k)$ procedure produces (an estimate of) an underbound for $\text{RS}(D, \text{BD}(\cdot))$ by making it *harder* for a feature to be selected to be in both purported biomarker sets. First, observe that as $[n_+, n_-]$ increases (keeping the n_+ -to- n_- ratio fixed, as we consider changing the size of the dataset), 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 a size- n subset $D^{(n)}$, versus from a size- $n/2$ subset $D^{(n/2)}$. Now consider two size- n datasets $D_1^{(n)}$ and $D_2^{(n)}$, and also two size- $n/2$ datasets $E_1^{(n/2)}$ and $E_2^{(n/2)}$. As $\text{BD}(D_1^{(n)})$ and $\text{BD}(D_2^{(n)})$ are each closer to F^* than $\text{BD}(E_1^{(n/2)})$ and

Figure 3. Diagram showing uRS's process of generating pairs of subsets for a dataset D (k times), then using those to compute $uRS(D, BD(\cdot), k)$.

$BD(E_2^{(n/2)})$, we expect $BD(D_1^{(n)})$ and $BD(D_2^{(n)})$ to be closer to each other, than $BD(E_1^{(n/2)})$ and $BD(E_2^{(n/2)})$, which means we expect that

$$J(BD(D_1^{(n)}), BD(D_2^{(n)})) \geq J(BD(E_1^{(n/2)}), BD(E_2^{(n/2)})).$$

In general, given that

$$\hat{E}^{(k)}[J(BD(D_1^{(s)}), BD(D_2^{(s)}))] \approx RS^*(\vec{p}(\cdot), s, BD(\cdot)) \quad (8)$$

(where $\hat{E}^{(k)}[\cdot]$ is the empirical average over k samples), this argues that the RS^* score should increase with the size s of the dataset – which suggests that

$$\begin{aligned} &\text{if } A_1, A_2 \sim \vec{p}^{sA}(\cdot), \quad B_1, B_2 \sim \vec{p}^{sB}(\cdot) \quad \text{and} \\ &\quad sA = |A_1| = |A_2| \quad \text{is larger than} \quad sB = |B_1| = |B_2|, \\ &\text{then we expect } J(BD(A_1), BD(A_2)) \quad \text{will be larger than} \quad J(BD(B_1), BD(B_2)) \end{aligned} \quad (9)$$

Figure 4(a) presents empirical evidence, over 5 datasets, supporting this claim – showing that the Jaccard score increases as we increase the size s of the datasets. Appendix C provides some arguments, and additional empirical evidence (over hundreds of simulations), that further support this heuristic.

This motivates our underbound algorithm $uRS(D, BD, k)$, which first partitions D into two disjoint size- $n/2$ subsets, E_1 and E_2 , with balanced outcomes. It then computes $J(BD(E_1), BD(E_2))$ which, assuming Heuristic 9, is an underbound in expectation for $RS(D, BD(\cdot))$. uRS does this partitioning k times, producing k different dataset pairs $\{[E_{1,r}, E_{2,r}]\}_{r=1, \dots, k}$, and returning the average Jaccard score, i.e.

$$uRS(D, BD(\cdot), k) = \frac{1}{k} \sum_{r=1}^k J(BD(E_{1,r}), BD(E_{2,r})). \quad (10)$$

See Figure 3.

2.4 Empirical study over Various Datasets

There are now many publicly-available datasets that have been used in association studies. Here, we use them to ...

- U1: Better understand what Jaccard scores are typical, for the standard $BD(\cdot)$ algorithms;
- U2: Determine whether our predictions match the results of earlier meta-analyses; and
- U3: Determine if our approximations are meaningful – i.e., if (for reasonable values of k):

$$uRS(D, BD, k) \leq RS(D, BD) \quad (11)$$

$$oRS(D, BD, k) \geq RS(D, BD) \quad (12)$$

Figure 4. (a) For each of the 5 datasets D , each point shows the average ($\pm$ sd) Jaccard $\hat{E}^{(k)}[J(\text{BD}(D_1^{(s)}), \text{BD}(D_2^{(s)}))]$ over $k = 20$ pairs $[D_1^{(s)}, D_2^{(s)}]$ of disjoint size- s subsets of D . Here, n is the size of the original dataset – note this can only go to $n/2$ – and we are using the standard $\text{BD}_{t,0.05,BH}$. (b) 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.

The next section will explicitly discuss all three issues. Of course, 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- $n/2$ (outcome-balanced) subset of D . In fact, $\text{uRS}(D, \text{BD}(\cdot), k)$ is a meaningful estimate of $\text{RS}(D^{(n/2)}, \text{BD}(\cdot))$; below we will use

$$\widehat{\text{RS}}(D^{(n/2)}, \text{BD}(\cdot), k) = \text{uRS}(D, \text{BD}(\cdot), k). \quad (13)$$

We will then compare this $\widehat{\text{RS}}(D^{(n/2)}, \text{BD}(\cdot), k)$ against $\text{uRS}(D^{(n/2)}, \text{BD}, k)$ and $\text{oRS}(D^{(n/2)}, \text{BD}, k)$ and to see whether the relations of Equations 11 and 12 both hold, with respect to various size- $n/2$ subsets $D^{(n/2)}$.

More generally, we can 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 outcome-balanced subsets $D', D'' \subset D$ where $|D'| = |D''| = s$ and $D' \cap D'' = \{\}$. For a fixed dataset D , and specified number $k \in \mathbb{Z}^{>0}$, we can then plot these $\widehat{\text{RS}}(D^{(s)}, \text{BD}(\cdot), k)$ values along with $\text{oRS}(D^{(s)}, \text{BD}, k)$ and $\text{uRS}(D^{(s)}, \text{BD}, k)$, as a function of s to see their behaviour; see Figure 4(b), for the Metabric dataset. Our website <https://biomarker.shinyapps.io/BiomarkerReprod/> also provides this visualization.

We explored our approximations over **25 different real-world datasets, including 16 microarray datasets and 2 RNAseq datasets with continuous data, and 7 SNP datasets with categorical data** (see Table 1). This first set includes 4 of the gene expression datasets discussed in the Zou *et al.*⁸ meta-analysis – each describing metastatic versus non-metastatic breast primary cancer subjects – to see if our method is consistent with their empirical results. We also included 11 other relatively-small gene expression datasets (with 19 to 187 subjects), focusing on human studies that had a binary class outcome from the GEO repository. To explore how our tools scale with size, we also included 3 other relatively large datasets, with 532 to 1654 subjects. As these were survival datasets, we set the binary outcome based on the median survival time (removing any subject that 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 (with 39 to 164 subjects), with discrete values, also selected from human studies with binary class outcome s . Figure 5 plots the number of features and biomarkers found, using the $\text{BD}_{t,0.05,BH}$ algorithm, for each dataset – both $D^{(n)}$ and $D^{(n/2)}$.

3 Results

We ran our suite of methods over the aforementioned 25 datasets, including 16 microarray datasets and 2 mRNAseq datasets, whose feature-values $\{x_i^j\}$ (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 mRNAseq datasets) and 7 were SNP datasets with categorical entries, *i.e.*, $x_i^j \in \{0, 1, 2\}$ is the number of minor alleles in the genotype for the i -th SNP for the j -th subject; see Table 1. Here, we use the standard $\text{BD}_{t,0.05,BH}(\cdot)$ biomarker discovery algorithm; see Section 2.2.

First, to address (U1) and (U2) in Section 2.4, we analyzed the 4 datasets mentioned in the Zou *et al.*⁸ meta-analysis (see the 4 “*” rows of Table 1) and computed the $\{\text{uRS}(D, \text{BD}_{t,0.05,BH}, 50), \text{oRS}(D, \text{BD}_{t,0.05,BH}, 50)\}$ values for each dataset D , as

Figure 5. Box+whiskers plots showing number of biomarkers found for $D^{(n/2)}$ when using $BD_{t,0.05,BH}$ over $k = 20$ iterations for various dataset, compared to the number of biomarkers for $D^{(n)}$, and to the number of features in each dataset. Note the y-axis is a log-scale. (Note we first changed all “0” values to “ 10^{-1} ”.) For details, see Tables 1 and 2.

Figure 6. Under-bound and over-bound for the 4 datasets from Zou *et al.* ⁸, as well as the true Jaccard score for each pair – 3 numbers for each dataset, shown by black circles.

well as the actual Jaccard score for biomarkers for each pair of datasets. (We were not able to replicate the results reported for the 5th dataset from that study, and so we excluded that one dataset from our analysis.) The results, in Figure 6, show that the Jaccard score for each pair is well within the bounds computed 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. Note that we verified that our $BD_{t,0.05,BH}$ algorithm matched the original study by verifying that the PO scores (Equation 15 in Appendix B.4) matched the ones that were originally published.

To address (U3), we also analyzed the other 14 continuous datasets D , and computed the oRS and uRS values using $k = 50$ repetitions; see Figure 7[left]. We see that the overbound oRS is consistently larger than the underbound uRS – *i.e.*, $uRS \leq oRS$ – as claimed by Equations 11-12. Figure 7[right] plots the corresponding values for the $D^{(n/2)}$ datasets, that use only 1/2 of the dataset, using the same $BD(\cdot)$ algorithm and $k = 50$. It also plots the “true” $\widehat{RS}(D^{(n/2)}, BD(\cdot), k)$ values for the datasets. Again, we see that $oRS \geq \widehat{RS} \geq uRS$.

Finally, similar to that experiment over the 18 continuous datasets, we examined the 7 discrete datasets and produced the reproducibility scores. Figure 8[left] shows the scores for each of the 7 SNP datasets, demonstrating that $oRS \geq uRS$ holds for the discrete cases as well. Figure 8[right] shows the scores for $D^{(n/2)}$ datasets, and performs the same verification. Those figures also allow us to see the Jaccard scores (U1 above) range from essentially 0 to around 0.475 for the $D^{(n/2)}$ datasets. In

Figure 7. Reproducibility scores (mean and standard deviation) for all 18 continuous datasets, both for complete datasets with n subjects (left) and for half-sized with $\frac{n}{2}$ subjects (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. Moreover, the right plot shows that the “truth” $\widehat{RS}$ is also between uRS and oRS. (The bRS lines in the plots are based on the Bootstrap Overbound method; see Appendix B.1.)

addition to the plots, Tables 1 and 2 present the relevant values.

Appendix B provides the results of many additional empirical studies, showing how the reproducibility scores change based on which (if any) MCC 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 .

4 Discussion

Limitations: This paper provides an effective way to estimate the reproducibility of the biomarkers found from a dataset, using a biomarker discovery algorithm. While the message of this paper is very general, the specific analyzes here all used the standard discovery algorithm $BD_{t,0.05,BH}$, to illustrate that the issues apply to the approaches commonly used. Appendix B explores some other discovery methods, that are based on t-tests. We anticipate the same issues would hold for other tests, such as Mann-Whitney, Wilcoxon or even multivariate analysis – but this is future work. All of our empirical studies 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 outcomes are binary and our use of t-test implicitly assumes the feature values, given each outcome, are Gaussian. 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 size-100 dataset, overlap with those from a size-300 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 “ $BD'(D, s)$ 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 MCC algorithm, or some other modification.)

Contributions: This paper has several contributions: (1) It motivates, then provides, a formal definition of reproducibility, that can help researchers “evaluate” a set of purported biomarkers; (2) It provides a pair of algorithms that can accurately bound this “reproducibility score” for a given dataset and biomarker discovery algorithm; (3) It provides empirical evaluation of these algorithms, over 25 different real-world datasets, to demonstrate that they work effectively; and (4) It introduces a freely-available website <https://biomarker.shinyapps.io/BiomarkerReprod/> that runs these algorithms on the dataset entered by a user, and biomarker discovery system, which will allow users to easily evaluate the quality of the biomarker set produced. Given how easy it is to use this tool, we hope that future researchers will automatically use it to quickly evaluate the quality of the purported biomarkers, then include these estimates when they publish their biomarkers. We anticipate this analysis may also lead to new biomarker discovery algorithms, to optimize this reproducibility score.

References

1. Strimbu, K. & Tavel, J. A. What are biomarkers? *Curr. Opin. HIV AIDS* 5, 463 (2010).

Figure 8. Reproducibility scores (mean and standard deviation) for 7 SNP datasets, both for complete datasets with n subjects $D^{(n)}$ (left) and for half-sized with $\frac{n}{2}$ subjects $D^{(n/2)}$ (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. Moreover, the right plot shows that the “truth” $\widehat{RS}$ is also within the range of oRS and uRS. (The bRS lines in the plots are based on the Bootstrap Overbound method; see Appendix B.1.)

2. Van't Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530 (2002).
3. Wang, Y. *et al.* Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet* **365**, 671–679 (2005).
4. Blau, N. Genetics of phenylketonuria: then and now. *Hum. mutation* **37**, 508–515 (2016).
5. Koenig, R. J. *et al.* Correlation of glucose regulation and hemoglobin a1c in diabetes mellitus. *New Engl. J. Medicine* **295**, 417–420 (1976).
6. Bush, W. S. & Moore, J. H. Genome-wide association studies. *PLoS computational biology* **8**, e1002822 (2012).
7. Boulesteix, A.-L. & Slawski, M. Stability and aggregation of ranked gene lists. *Briefings bioinformatics* **10**, 556–568 (2009).
8. Zou, J. *et al.* Revealing weak differential gene expressions and their reproducible functions associated with breast cancer metastasis. *Comput. biology chemistry* **39**, 1–5 (2012).
9. Dhami, R., Passini, M. A. & Schuchman, E. H. Identification of novel biomarkers for niemann–pick disease using gene expression analysis of acid sphingomyelinase knockout mice. *Mol. Ther.* **13**, 556–564 (2006).
10. Shlomi, T., Cabili, M. N. & Ruppin, E. Predicting metabolic biomarkers of human inborn errors of metabolism. *Mol. systems biology* **5**, 263 (2009).
11. Gormley, M., Dampier, W., Ertel, A., Karacali, B. & Tozeren, A. Prediction potential of candidate biomarker sets identified and validated on gene expression data from multiple datasets. *BMC bioinformatics* **8**, 415 (2007).
12. Zacharias, H. U. *et al.* Scale-invariant biomarker discovery in urine and plasma metabolite fingerprints. *J. proteome research* **16**, 3596–3605 (2017).
13. Abraham, A. *et al.* Deriving reproducible biomarkers from multi-site resting-state data: An autism-based example. *NeuroImage* **147**, 736–745 (2017).
14. Desmedt, C. *et al.* Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the transbig multicenter independent validation series. *Clin. cancer research* **13**, 3207–3214 (2007).
15. Pearl, J. *Causality* (Cambridge university press, 2009).
16. Ein-Dor, L., Kela, I., Getz, G., Givol, D. & Domany, E. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics* **21**, 171–178 (2004).
17. Sørli, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci.* **98**, 10869–10874 (2001).

Table 1. Results for all 25 datasets when using all the subjects, sorted by sample size (#subjects) – corresponding to Figure 5. The first 18 entries are Gene Expression datasets (including the 4 “*”ed entries, from the Zou *et al.* ⁸ meta-study), and the final 7 are SNP datasets. Reproducibility Scores are shown in the form of mean $\pm$ standard deviation. The “(Majority %)” values are the percentage of the subjects in the dataset with the more common outcome – e.g., 53% of the subjects in the GDS968 dataset are labeled “+” (for “long survival time”), and 82% of GSE7390 are labeled “–” (for “Non-Metastatic”).

Name	#subjects (Majority %)	#features	#biomarkers	uRS %	oRS %
GDS968 ⁵⁰	171 (53%)	5748	2506	47.5 $\pm$ 3.95	63.25 $\pm$ 0.76
GSE7390* ^{14,51}	198 (82%)	13245	18	0 $\pm$ 0	5.15 $\pm$ 2.81
GSE2034* ³	286 (67%)	13245	277	0 $\pm$ 0	12.6 $\pm$ 4.19
GSE1456* ⁵²	159 (78%)	13245	443	0 $\pm$ 0	13.6 $\pm$ 6.4
GSE11121* ⁵³	200 (86%)	13245	492	0.09 $\pm$ 0.2	13.4 $\pm$ 5.89
GDS2546 ^{54,55}	167 (54%)	12553	2965	30.8 $\pm$ 4.82	49.0 $\pm$ 0.55
GDS2545 ^{54,55}	171 (53%)	12558	4291	34.0 $\pm$ 4.09	54.58 $\pm$ 0.31
GDS2547 ^{54,55}	164 (54%)	12579	1810	23.7 $\pm$ 5.86	42.66 $\pm$ 0.70
KIPAN ⁵⁶	532 (81%)	18271	2782	12.3 $\pm$ 4.91	34.2 $\pm$ 4.74
BRCA ⁵⁶	552 (95%)	18320	2	0 $\pm$ 0	2.5 $\pm$ 2.1
GDS2771 ^{57,58}	187 (52%)	22215	1807	0.32 $\pm$ 0.64	31.68 $\pm$ 0.47
GDS3966 ⁵⁹	83 (63%)	22274	6554	31.7 $\pm$ 4.71	53.66 $\pm$ 0.27
GDS4185 ^{60,61}	67 (58%)	22283	6	0 $\pm$ 0	11.29 $\pm$ 0.83
Metabric ⁶²	1654 (57%)	24368	3675	18.5 $\pm$ 3.86	39.8 $\pm$ 3.56
GDS4431 ⁶³	146 (53%)	54613	140	0 $\pm$ 0	18.85 $\pm$ 0.34
GDS4719 ⁶⁴	19 (53%)	54675	1	0 $\pm$ 0	7.02 $\pm$ 0.71
GDS2737 ⁶⁵	37 (57%)	54675	4	0 $\pm$ 0	10.58 $\pm$ 0.5
GDS5218 ⁶⁶	110 (56%)	54675	10700	24.0 $\pm$ 5.09	46.06 $\pm$ 0.29
GSE13429 ⁶⁷	39 (79%)	262314	1267	1.66 $\pm$ 0.44	21.4 $\pm$ 3.76
GSE25103 ⁶⁸	122 (92%)	908512	325	0.27 $\pm$ 0.14	3.94 $\pm$ 2.22
GSE15097 ⁶⁹	68 (59%)	909456	108224	4.9 $\pm$ 2.9	34.2 $\pm$ 5.64
GSE15096 ⁶⁹	69 (58%)	909457	106482	5.37 $\pm$ 2.9	33.4 $\pm$ 5.28
GSE25104 ^{68,70}	122 (92%)	909547	326	0.27 $\pm$ 0.14	3.94 $\pm$ 2.22
GSE15826 ⁷¹	164 (54%)	909549	0	0 $\pm$ 0	2.5 $\pm$ 0.58
GSE18333 ⁷²	82 (54%)	909606	0	0 $\pm$ 0	0 $\pm$ 0

18. Ramaswamy, S., Ross, K. N., Lander, E. S. & Golub, T. R. A molecular signature of metastasis in primary solid tumors. *Nat. genetics* **33**, 49 (2002).
19. Alizadeh, A. A. *et al.* Prediction of survival in diffuse large b-cell lymphoma based on the expression of 2 genes reflecting tumor and microenvironment. *Blood* **118**, 1350–1358 (2011).
20. Miklos, G. L. G. & Maleszka, R. Microarray reality checks in the context of a complex disease. *Nat. biotechnology* **22**, 615 (2004).
21. Mischak, H. *et al.* Recommendations for biomarker identification and qualification in clinical proteomics. *Sci. translational medicine* **2**, 46ps42–46ps42 (2010).
22. Rossing, K. *et al.* Urinary proteomics in diabetes and ckd. *J. Am. Soc. Nephrol.* **19**, 1283–1290 (2008).
23. Haubitz, M. *et al.* Identification and validation of urinary biomarkers for differential diagnosis and evaluation of therapeutic intervention in anti-neutrophil cytoplasmic antibody-associated vasculitis. *Mol. & Cell. Proteomics* **8**, 2296–2307 (2009).
24. Begley, C. G. & Ellis, L. M. Raise standards for preclinical cancer research. *Nature* **483**, 531–533 (2012).
25. Wen, H., Wang, H.-Y., He, X. & Wu, C.-I. On the low reproducibility of cancer studies. *Natl. science review* **5**, 619–624 (2018).
26. Baker, M. Reproducibility crisis? *Nature* **533**, 353–66 (2016).
27. Holte, R. C., Acker, L., Porter, B. W. *et al.* Concept learning and the problem of small disjuncts. In *IJCAI*, vol. 89, 813–818 (1989).

Table 2. Results for all datasets when using half of the subjects – i.e., $D^{(n/2)}$. Reproducibility Scores and average number of biomarkers are shown in the form of mean $\pm$ standard deviation. (The caption for Table 1 describes the row ordering.)

Name	Average #biomarkers	uRS %	$\widehat{RS}$ %	oRS %
GDS968	1593.67 $\pm$ 173.35	26.0 $\pm$ 5.87	47.5 $\pm$ 3.95	48.6 $\pm$ 5.21
GSE7390*	1.38 $\pm$ 3.01	0 $\pm$ 0	0 $\pm$ 0	1.84 $\pm$ 2.1
GSE2034*	41.05 $\pm$ 144.8	0 $\pm$ 0	0 $\pm$ 0	6.39 $\pm$ 4.72
GSE1456*	58.38 $\pm$ 138.11	0 $\pm$ 0	0 $\pm$ 0	4.5 $\pm$ 4.41
GSE11121*	181.8 $\pm$ 265.5	0 $\pm$ 0	0.09 $\pm$ 0.2	9.48 $\pm$ 7.67
GDS2546	1266.61 $\pm$ 354.72	5.01 $\pm$ 7.15	30.8 $\pm$ 4.82	33.5 $\pm$ 6.3
GDS2545	2051.58 $\pm$ 494.77	6.38 $\pm$ 7.33	34.0 $\pm$ 4.09	38.1 $\pm$ 5.92
GDS2547	648.55 $\pm$ 255.09	1.82 $\pm$ 4.24	23.7 $\pm$ 5.86	27.3 $\pm$ 6.63
KIPAN	694.48 $\pm$ 398.88	0.88 $\pm$ 2	12.3 $\pm$ 4.91	19.3 $\pm$ 8.33
BRCA	36.75 $\pm$ 171.15	0 $\pm$ 0	0 $\pm$ 0	2.6 $\pm$ 4.2
GDS2771	457.17 $\pm$ 863.64	0.09 $\pm$ 0.62	0.32 $\pm$ 0.64	7.23 $\pm$ 8.48
GDS3966	2976.15 $\pm$ 811.92	10.6 $\pm$ 7.83	31.7 $\pm$ 4.71	37.9 $\pm$ 6.44
GDS4185	2.92 $\pm$ 40.20	0 $\pm$ 0	0 $\pm$ 0	0.701 $\pm$ 3.11
Metabric	1301.9 $\pm$ 504.80	1.97 $\pm$ 3.86	18.5 $\pm$ 3.86	21.6 $\pm$ 7.3
GDS4431	43.35 $\pm$ 284.56	0 $\pm$ 0	0 $\pm$ 0	2.55 $\pm$ 3.82
GDS4719	0.36 $\pm$ 1.88	0 $\pm$ 0	0 $\pm$ 0	2.8 $\pm$ 5.31
GDS2737	0.79 $\pm$ 7.33	0 $\pm$ 0	0 $\pm$ 0	1.73 $\pm$ 2.82
GDS5218	4207.29 $\pm$ 1589.79	7.96 $\pm$ 4.08	24.0 $\pm$ 5.09	31.5 $\pm$ 6.15
GSE13429	339.29 $\pm$ 170.11	1.73 $\pm$ 0.44	1.66 $\pm$ 0.44	14 $\pm$ 5.04
GSE25103	309.61 $\pm$ 78.24	0.13 $\pm$ 0.11	0.27 $\pm$ 0.14	6.63 $\pm$ 6.79
GSE15097	22788.62 $\pm$ 23342.37	0.77 $\pm$ 0.62	4.9 $\pm$ 2.99	16.8 $\pm$ 6.85
GSE15096	22393.87 $\pm$ 20624.09	0.74 $\pm$ 0.60	5.37 $\pm$ 2.9	16.1 $\pm$ 6.91
GSE25104	309.93 $\pm$ 78.14	0.13 $\pm$ 0.11	0.27 $\pm$ 0.14	6.64 $\pm$ 6.8
GSE15826	1.21 $\pm$ 5.41	0 $\pm$ 0	2.03 $\pm$ 0.06	1.21 $\pm$ 0.79
GSE18333	0.01 $\pm$ 0.1	0 $\pm$ 0	0 $\pm$ 0	0 $\pm$ 0

28. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545–15550 (2005).
29. Ciucci, S. *et al.* Enlightening discriminative network functional modules behind principal component analysis separation in differential-omic science studies. *Sci. reports* **7**, 43946 (2017).
30. Witten, I. H., Frank, E., Hall, M. A. & Pal, C. J. *Data Mining: Practical machine learning tools and techniques* (Morgan Kaufmann, 2016).
31. Frantz, S. An array of problems. *Nat. Rev. Drug Discov.* **4**, 362–363 (2005).
32. Li, M. *et al.* Identifying reproducible molecular biomarkers for gastric cancer metastasis with the aid of recurrence information. *Sci. reports* **6**, 24869 (2016).
33. Zhang, M. *et al.* Apparently low reproducibility of true differential expression discoveries in microarray studies. *Bioinformatics* **24**, 2057–2063 (2008).
34. Ioannidis, J. P. Biomarker failures. *Clin. chemistry* **59**, 202–204 (2013).
35. Ein-Dor, L., Zuk, O. & Domany, E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc. Natl. Acad. Sci.* **103**, 5923–5928 (2006).
36. Shi, L. *et al.* The microarray quality control (maqc) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nat. biotechnology* **24**, 1151 (2006).
37. Guo, L. *et al.* Rat toxicogenomic study reveals analytical consistency across microarray platforms. *Nat. biotechnology* **24**, 1162 (2006).
38. Klebanov, L., Qiu, X., Welle, S. & Yakovlev, A. Statistical methods and microarray data. *Nat. biotechnology* **25**, 25 (2007).
39. Gagno, S. *et al.* A new genetic risk score to predict the outcome of locally advanced or metastatic breast cancer patients treated with first-line exemestane: Results from a prospective study. *Clin. breast cancer* **19**, 137–145 (2019).

40. Chitpin, J. G., Awdeh, A. & Perkins, T. J. Recap reveals the true statistical significance of chip-seq peak calls. *Bioinformatics* **35**, 3592–3598 (2019).
41. Pavelka, N. *et al.* A power law global error model for the identification of differentially expressed genes in microarray data. *BMC bioinformatics* **5**, 203 (2004).
42. Alshawaqfeh, M., Bashaiah, A., Serpedin, E. & Suchodolski, J. Consistent metagenomic biomarker detection via robust pca. *Biol. direct* **12**, 4 (2017).
43. Zhao, S. D. & Li, Y. Score test variable screening. *Biometrics* **70**, 862–871 (2014).
44. Ma, S. *et al.* Identification of non-hodgkin's lymphoma prognosis signatures using the ctgdr method. *Bioinformatics* **26**, 15–21 (2010).
45. Filosi, M., Visintainer, R., Riccadonna, S., Jurman, G. & Furlanello, C. Stability indicators in network reconstruction. *PloS one* **9** (2014).
46. Hua, J., Bittner, M. L. & Dougherty, E. R. Evaluating gene set enrichment analysis via a hybrid data model. *Cancer informatics* **13**, CIN–S13305 (2014).
47. Shi, X., Yi, H. & Ma, S. Measures for the degree of overlap of gene signatures and applications to tcga. *Briefings bioinformatics* **16**, 735–744 (2015).
48. Efron, B. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, 569–593 (Springer, 1992).
49. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. royal statistical society. Ser. B (Methodological)* 289–300 (1995).
50. Rieger, K. E. *et al.* Toxicity from radiation therapy associated with abnormal transcriptional responses to dna damage. *Proc. Natl. Acad. Sci.* **101**, 6635–6640 (2004).
51. Patil, P., Bachant-Winner, P.-O., Haibe-Kains, B. & Leek, J. T. Test set bias affects reproducibility of gene signatures. *Bioinformatics* **31**, 2318–2323 (2015).
52. Pawitan, Y. *et al.* Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast cancer research* **7**, R953 (2005).
53. Schmidt, M. *et al.* The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer research* **68**, 5405–5413 (2008).
54. Chandran, U. R. *et al.* Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process. *BMC cancer* **7**, 64 (2007).
55. Yu, Y. P. *et al.* Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy. *J. clinical oncology* **22**, 2790–2799 (2004).
56. Data generated by the tcga research network: <http://cancergenome.nih.gov/>. Downloaded Feb 2019.
57. Spira, A. *et al.* Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nat. medicine* **13**, 361 (2007).
58. Gustafson, A. M. *et al.* Airway pi3k pathway activation is an early and reversible event in lung cancer development. *Sci. translational medicine* **2**, 26ra25–26ra25 (2010).
59. Xu, L. *et al.* Gene expression changes in an animal melanoma model correlate with aggressiveness of human melanoma metastases. *Mol. Cancer Res.* **6**, 760–769 (2008).
60. Hutcheson, J. *et al.* Combined deficiency of proapoptotic regulators bim and fas results in the early onset of systemic autoimmunity. *Immunity* **28**, 206–217 (2008).
61. Becker, A. M. *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* **8**, e67003 (2013).
62. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346 (2012).
63. Alter, M. D. *et al.* Autism and increased paternal age related changes in global levels of gene expression regulation. *PloS one* **6**, e16715 (2011).
64. Fernandez, D. R. *et al.* Activation of mammalian target of rapamycin controls the loss of tcr ζ in lupus t cells through hres-1/rab4-regulated lysosomal degradation. *The J. Immunol.* **182**, 2063–2073 (2009).

65. Burney, R. O. *et al.* Gene expression analysis of endometrium reveals progesterone resistance and candidate susceptibility genes in women with endometriosis. *Endocrinology* **148**, 3814–3826 (2007).
66. Raue, U. *et al.* Transcriptome signature of resistance exercise adaptations: mixed muscle and fiber type specific profiles in young and old adults. *Am. J. Physiol. Circ. Physiol.* (2012).
67. Venkatachalam, R. *et al.* Identification of candidate predisposing copy number variants in familial and early-onset colorectal cancer patients. *Int. journal cancer* **129**, 1635–1642 (2011).
68. Peng, C.-H. *et al.* A novel molecular signature identified by systems genetics approach predicts prognosis in oral squamous cell carcinoma. *PloS one* **6**, e23452 (2011).
69. Nrv, E. *et al.* High-resolution dna analysis of human embryonic stem cell lines reveals culture-induced copy number changes and loss of heterozygosity. *Nat. biotechnology* **28**, 371 (2010).
70. Lee, C.-H. *et al.* Il-1 promotes malignant transformation and tumor aggressiveness in oral cancer. *J. cellular physiology* **230**, 875–884 (2015).
71. Pamphlett, R. Affymetrix 6.0 study of sporadic motor neuron disease patients and controls., geo, v1. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE15826> (2010).
72. Mao, X. *et al.* Distinct genomic alterations in prostate cancers in chinese and western populations suggest alternative pathways of prostate carcinogenesis. *Cancer research* **70**, 5207–5212 (2010).
73. Satagopan, J. M., Venkatraman, E. & Begg, C. B. Two-stage designs for gene–disease association studies with sample size constraints. *Biometrics* **60**, 589–597 (2004).
74. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Annals statistics* 1165–1188 (2001).
75. Dunn, O. J. Multiple comparisons among means. *J. Am. statistical association* **56**, 52–64 (1961).
76. Hochberg, Y. A sharper bonferroni procedure for multiple tests of significance. *Biometrika* **75**, 800–802 (1988).
77. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. journal statistics* 65–70 (1979).

Acknowledgements

The authors all gratefully acknowledge the funding support from Canada’s NSERC (Natural Sciences and Engineer Research Council) and the Alberta Machine Intelligence Institute (Amii). We also benefited from the preliminary research work by Patrick Schwaferts and discussions with Sambasivarao Damaraju.

Author contributions statement

All 4 authors helped design the experiment, analyze the results, and write this manuscript. In addition, R.G. motivated and initiated this project; S.K. designed and implemented the overbound and underbound algorithms, and the web application; and A.F. and A.R. (re)implemented and ran the bulk of the experiments. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

A Notes from Text

This appendix provides various short notes related to material in the main text:

A.1 Glossary

Term	Description
$BD_{t,p,\chi}(D)$	Biomarker Discovery algorithm: given a labeled dataset D , return the set of features (associated with D), that exhibit a class difference; see Equation 1. Here, the t in the subscript refers to the 2-sided t -test, the p for the p -value used, and χ to the MCC correction. Our canonical example is $BD_{t,0.05,BH}(D)$, with $p = 0.05$, and $\chi = BH$ corresponding to the Benjamini/Hochberg correction. We use $BD(D)$ for a generic algorithm.
biomarker	“a characteristic (feature) that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention” ¹
$D, D^{(n/2)}$	D refers to a labeled dataset, with n subjects, each described by r features, and an outcome. $D^{(n/2)}$ refers to a dataset formed containing half of the elements of the original dataset D – this is balanced, containing $n_+/2$ +-labeled subjects and $n_-/2$ -labeled subject
feature	sometimes called the “independent variables” or “covariates” – here, we assume they are all real valued
gene expression	continuous value that indicates how much RNA of a specific gene has been expressed
GWAS	genome-wide association study
$J(A, B)$	Jaccard score of pair of sets A and B ; see Equation 3
$n, [n_+, n_-]$	By convention, n is the number of subjects in the dataset, which includes n_+ subjects that are labeled +, and n_- subjects labeled -.
$oRS(D, BD(\cdot), k)$	(an estimate of) an overbound of $RS(D, BD(\cdot))$ where $k = \#repetitions$; see Equation 6
outcome	sometimes called the “label”, “response” or “dependent variable” – here, just considering binary (think “case vs control”)
$\vec{p}(\cdot) \quad \vec{p}^{[n_+, n_-]}(\cdot)$	$\vec{p}(\cdot) = [p_{j,c}(\cdot)]_{j,c}$ is the $2 \times r$ matrix of the distributions for the 2 outcomes and r features; $\vec{p}^{[n_+, n_-]}(\cdot)$ is the distribution that draws, for each feature j , n_+ subjects from $p_{j,+}$ and n_- from $p_{j,-}$
$\widehat{p}_D(\cdot)$	empirical estimate of $\vec{p}(\cdot)$, based on the context of the dataset D
$RS(D, BD(\cdot))$	“Reproducibility Score” of the biomarkers produced, by running biomarker-discovery algorithm $BD(\cdot)$ on the labeled dataset D ; see Equation 4
$RS^*(\vec{p}(\cdot), [n_+, n_-], BD(\cdot))$	reproducibility score for the biomarkers produced by the $BD(\cdot)$ biomarker discovery tool on two datasets of size $[n_+, n_-]$ drawn from $\vec{p}(\cdot)$; see Equation 2
SNP	single nucleotide polymorphism – discrete, with values {AA, Ab, bb} indicating homozygous dominant, heterozygous, and homozygous minor, respectively
subject	a “record” in the dataset (e.g., a row in the table shown in Figure 1). Also called “element” or “instance”
$uRS(D, BD(\cdot), k)$	(an estimate of) an underbound of $RS(D, BD(\cdot))$, where $k = \#repetitions$; see Equation 10

A.2 How biomarker discovery differs from supervised machine learning

We earlier noted a few differences between association studies versus predictive studies. Here, we list two others:

(a) In general, a predictive model provides some information about *an individual* (corresponding to a **row** of the matrix at the top of Figure 1) – e.g., whether she has some specific disease. By contrast, an association study identifies *features* (each corresponding to a **column**), with the prediction that each considered a biomarker will each exhibit some *population difference with respect to a dataset of many individuals*.

(b) It is relatively easy to evaluate the quality of a learned predictive model, by running that predictor on a held-out set of subjects. 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 – e.g., if they all set $p = 1$, then each discoverer will return all features; this will have low variance, but presumably high bias.)

A.3 Why do many papers not provide biological validation?

This may be because such biological validation is not yet implementable, or the technology is not yet available. Alternatively, the biological validation may be possible but be a major project that those authors hope to explore in future works.

A.4 Our model (Equation 1) deals with just a single step

Some modern GWAS studies 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, etc⁷³. Here, our analysis is relevant to any one of these phases; see Figure 1. Also, some studies regress out covariates before finding biomarkers; here, we assume that this has happened and our analysis uses those already-regressed-out values.

A.5 Some features may only be important in combination

Sometimes a feature may be completely independent of the outcome, by itself, but become relevant, in combination with another (independent) feature. As an example, consider babies *in utero*, whose descriptions each include the Rh blood type of its mother, $MRh \in \{+, -\}$, and also of its father $FRh \in \{+, -\}$, among other features, where the outcome is the health H of that baby. Note that H may be completely uncorrelated with MRh , and also completely uncorrelated with FRh – which means neither MRh nor FRh could be a biomarker. However, suppose finding these blood factors are different $MRh \neq FRh$, increases the baby’s risk. Assuming balanced sampling (with an equal number of $MRh=+$ and $MRh=-$ subjects, and similarly for FRh), this means an effective predictive model would need to include both features, even though neither is a biomarker. Moreover, 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

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 a set 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; the analysis described in the paper still apply. Note, however, that here we assume this super-feature construction is known initially, and in particular, this paper 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 the relevance of one feature is implicitly conditioned on other features simultaneously – *e.g.*, multiple regression models.

B Exploring Other Settings

For consistency, all of the experiments in the main text used the same $BD_{t,0.05,BH}$ biomarker discovery algorithm. However, there are many other approaches that have been used in other association studies. Here, we continue to consider only the t-test as the main statistical significance test. First, Appendix B.1 explores the obvious bootstrap method, and demonstrates that its estimate is consistently worse than our over-bound method. The other sub-appendices consider only the algorithms explicitly described in the main text. Appendix B.2 explores different options for the p-value threshold and the p-value adjustment method, to see how changing these affect the reproducibility results. This paper introduced two different approximations for the *Reproducibility Score* – uRS and oRS. Appendix B.3 explores how these approximations change as we adjust the size of the dataset n , and the number of iterations of running the algorithms, k . Finally, Appendix B.4 motivates then presents the PO score, an alternative to the Jaccard score.

B.1 Bootstrap Sampling Method

We initially explored the standard bootstrap sampling method bRS as an overbound measure. Given a dataset D , this bRS method produces new datasets by drawing $n = |D|$ subjects from D , with replacement: In particular, for $i = 1..k$, $bRS(D, BD(\cdot), k)$ draws a pairs of such bootstrap datasets $D_{i,1}$ and $D_{i,2}$ from D , then computes the Jaccard score of their respective $BD(\cdot)$ -biomarkers, and returns the average. Figures 7 and 8 show these results for continuous and SNP datasets, respectively. These experiments show, for all 25 datasets, and for both $D^{(n)}$ and $D^{(n/2)}$ datasets (when using all n , and also when using $\frac{n}{2}$ subjects), that this method is a more extreme over-bound:

$$bRS(D, BD(\cdot), k) \geq oRS(D, BD(\cdot), k) \geq RS(D, BD(\cdot)) \quad (14)$$

Figure B.1. (a) 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 MCC correction ($BD_{t,0.05,BH}$) and the y -value which represents the score without MCC correction ($BD_{t,0.05,-}$). Each point above the diagonal line means the MCC correction led to inferior performance for the associated dataset. (b) Reproducibility scores $\widehat{RS}$ for all 25 datasets when using $D^{(n/2)}$, for different p-value adjustment methods – i.e., $BD_{t,0.05,\chi}$ for 5 different FDR adjustment methods χ , including BH and “no”.

which means it is a less-useful measure.

B.2 p-value adjustment methods and p-value threshold

Figure B.1(a) shows the effect of Benjamini and Hochberg (BH) MCC correction on the reproducibility scores, by comparing the average Jaccard score for the biomarkers found for a pair of complementary $D^{(n/2)}$ datasets when using the standard MCC correction ($BD_{t,0.05,BH}$), versus without MCC correction ($BD_{t,0.05,-}$) across all datasets. We see that this MCC correction 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 MCC, in addition to Benjamini+Hochberg (BH)⁴⁹, including: Benjamini and Yekutieli (BY)⁷⁴, Bonferroni⁷⁵, Hochberg⁷⁶ and Holm⁷⁷. Figure B.1(b) shows the results of these 5 MCC methods, as well as the “no-FDR” approach, over all 25 datasets – here showing RS with respect to the half-datasets $D^{(n/2)}$; see Equation 8. We see again that “no-MCC” 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., $BD_{t,\tau,BH}$, 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 B.2 shows that the reproducibility score for $D^{(n/2)}$ appears monotonic with τ – within this $\tau \in (0, 0.1)$ range, larger τ produces higher RS.

B.3 Changing k , the number of data-subset pairs drawn

Each of our approximation algorithms uses k , the number of data-subset pairs drawn. 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 therefore ran these algorithms for our largest dataset, Metabric, but varied this k . Figure B.3 shows that we obtained very similar results, whether we used $k = 10$, up to $k = 50$, for both uRS and oRS, for both Metabric and Metabric $^{(n/2)}$. We also computed the mRS values for all 25 datasets D when running the algorithm for $k = 10$ versus $k = 50$, and found the difference between the two, $mRS(D^{(n/2)}, BD(\cdot, k=50)) - mRS(D^{(n/2)}, BD(\cdot, k=10))$, is very close to 0 for most cases; see Figure B.4.

B.4 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 $J(A, B) = J(B, A)$, which 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.*⁸ meta-study: for each

Figure B.2. Reproducibility scores $\widehat{RS}$ for various datasets when using $D^{(n/2)}$ for different p-value thresholds – $BD_{t,\tau,BH}$, for various $\tau \in (0, 0.1)$. (left) considers 5 datasets, for a range of 20 different values of τ ; (right) plots the Jaccard scores for all 25 datasets, for 5 different values of τ .

Figure B.3. Reproducibility scores for different numbers of iterations, for the Metabric^(n/2) datasets.

(ordered) pair of biomarker sets $[B_i, B_j]$,

$$PO(B_i, B_j) = \frac{|B_i \cap B_j|}{|B_i|} \times 100\% . \quad (15)$$

Notice this is an asymmetric variant of the Jaccard score (Equation 3). (As that paper also reported the number of biomarkers found for each dataset, we could therefore recover the associated Jaccard score.) This measure can be reported with a set of biomarkers, to estimate the number of these biomarkers that should recur, if performing another similar study.

RG: [What else to say here?]

C Arguments Supporting the Heuristics

The overbound and underbound algorithms, uRS and oRS, are based on some intuitions, appearing as Heuristics 7 and 9. Our empirical results, Table 1 and 2, support these claims over 25 real-world datasets – as uRS is consistently below $\widehat{RS}$, which is below oRS (as visualized by the relative heights of uRS vs $\widehat{RS}$ vs oRS in Figures 4(a,b)).

This appendix further motivates these heuristics. Subappendices C.1 and C.2 first provide arguments that motivate these heuristics (both focus on simple use of t-test $BD_{t,\alpha,-}$), then Appendix C.3 provides further empirical evidence that these claims hold in practice. Appendix C.4 provides additional theoretical justifications for the overlaps claims underlying Heuristic 7.

Figure B.4. uRS and oRS values for all 25 datasets when running $k = 10$ versus $k = 50$ iterations. Note the black line, hovering around 0, is the difference between the uRS (left) values when using $k = 10$ iterations and when using $k = 50$ iterations; and similarly for the oRS values (right).

C.1 Motivation for Heuristic 7

Here, we motivate Heuristic 7 by explaining why we expect the Jaccard score of the biomarkers found from two related datasets, to increase as the subjects forming the dataset have larger overlap. We initially consider a single feature, call it g , and explore a relevant statistic from these two datasets. Here, we assume that every subject of g associated with a positively (resp., negatively) labeled subject is drawn from a distribution with mean μ_+ and variance σ_+^2 (resp., μ_- and σ_-^2). (Of course, for non-biomarkers, $\mu_+ = \mu_0$ and $\sigma_+^2 = \sigma_-^2$.) Each dataset has n^+ positive subjects and n^- negative subjects, and these are drawn i.i.d., except that each dataset includes r^+ common positive subjects and r^- common negative subjects.

For notation, let $D_1 = E_1^+ \cup E_1^-$ and $D_2 = E_2^+ \cup E_2^-$ be the two datasets where E_1^+ and E_2^+ (resp., E_1^- and E_2^-) are the values of feature g associated with the positively (resp., negatively) labeled subjects:

$$\begin{aligned} E_1^+ &= \{a_1, \dots, a_{r^+}, b_1, b_2, \dots, b_{k^+}\} \\ E_2^+ &= \{a_1, \dots, a_{r^+}, c_1, c_2, \dots, c_{k^+}\} \\ E_1^- &= \{d_1, \dots, d_{r^-}, e_1, e_2, \dots, e_{k^-}\} \\ E_2^- &= \{d_1, \dots, d_{r^-}, f_1, f_2, \dots, f_{k^-}\} \end{aligned}$$

where each a_i, b_i, c_i is drawn, i.i.d., from a distribution with variance σ_+^2 , and each d_i, e_i, f_i is drawn, i.i.d., from a distribution with variance σ_-^2 .

As desired, E_1^+ and E_2^+ share $r^+ = n^+ - k^+$ elements in common, and E_1^- and E_2^- share $r^- = n^- - k^-$ elements in common.

Now define the means:

$$\begin{aligned} \bar{X}_1^+ &= \frac{1}{n^+} \left[\sum_{i=1}^{r^+} a_i + \sum_{j=1}^{k^+} b_j \right] & \bar{X}_1^- &= \frac{1}{n^-} \left[\sum_{i=1}^{r^-} d_i + \sum_{j=1}^{k^-} e_j \right] \\ \bar{X}_2^+ &= \frac{1}{n^+} \left[\sum_{i=1}^{r^+} a_i + \sum_{j=1}^{k^+} c_j \right] & \bar{X}_2^- &= \frac{1}{n^-} \left[\sum_{i=1}^{r^-} d_i + \sum_{j=1}^{k^-} f_j \right] \end{aligned}$$

then define the joint total variance values as

$$\begin{aligned} \bar{S}_1 &= \sum_{i=1}^{r^+} (a_i - \bar{X}_1^+)^2 + \sum_{j=1}^{k^+} (b_j - \bar{X}_1^+)^2 + \sum_{i=1}^{r^-} (d_i - \bar{X}_1^-)^2 + \sum_{j=1}^{k^-} (e_j - \bar{X}_1^-)^2 \\ \bar{S}_2 &= \sum_{i=1}^{r^+} (a_i - \bar{X}_2^+)^2 + \sum_{j=1}^{k^+} (c_j - \bar{X}_2^+)^2 + \sum_{i=1}^{r^-} (d_i - \bar{X}_2^-)^2 + \sum_{j=1}^{k^-} (f_j - \bar{X}_2^-)^2 \end{aligned}$$

Note that $\bar{S}_1$ and $\bar{S}_2$ should be extremely similar, as each is the sum of n^+ terms, each with expected value of σ_+^2 , and of n^- terms each with expected value σ_-^2 . We will assume they are effectively the same value; call it $\bar{S}_{1,2} = \frac{\bar{S}_1 + \bar{S}_2}{2}$.

Now observe the t-statistics of each dataset:

$$t(E_1) = \frac{\bar{X}_1^+ - \bar{X}_1^-}{\kappa \sqrt{\bar{S}_1}} \quad t(E_2) = \frac{\bar{X}_2^+ - \bar{X}_2^-}{\kappa \sqrt{\bar{S}_2}} \quad (16)$$

where $\kappa = \sqrt{\frac{1}{n_+ + n_- - 2} \left[\frac{1}{n_+} + \frac{1}{n_-} \right]}$, and define

$$Q = t(E_1) - t(E_2) = \frac{1}{\kappa} \left[\frac{\bar{X}_1^+ - \bar{X}_1^-}{\sqrt{\bar{S}_1}} - \frac{\bar{X}_2^+ - \bar{X}_2^-}{\sqrt{\bar{S}_2}} \right] \approx \frac{1}{\kappa} \left[\frac{(\bar{X}_1^+ - \bar{X}_1^-) - (\bar{X}_2^+ - \bar{X}_2^-)}{\sqrt{\bar{S}_{1,2}}} \right]$$

as the different between the t-statistics of these datasets. Note that the expected value $E[Q] = 0$. We want to show that its variance decreases as r^+ or r^- increases.

We will view $\bar{S}_{1,2}$ as a constant, and focus on the numerator

$$\begin{aligned} R &= (\bar{X}_1^+ - \bar{X}_1^-) - (\bar{X}_2^+ - \bar{X}_2^-) = (\bar{X}_1^+ - \bar{X}_2^+) - (\bar{X}_2^- - \bar{X}_1^-) \\ &= \frac{1}{n^+} \sum_{j=1}^{k^+} (b_j - c_j) + \frac{1}{n^-} \sum_{j=1}^{k^-} (f_j - e_j) \end{aligned}$$

We can now observe that

$$\begin{aligned} \text{Var}[R | k^+, k^-] &= \sum_{j=1}^{k^+} \text{Var}\left[\frac{1}{n^+} b_j\right] + \text{Var}\left[\frac{1}{n^+} c_j\right] + \sum_{j=1}^{k^-} \text{Var}\left[\frac{1}{n^-} f_j\right] + \text{Var}\left[\frac{1}{n^-} e_j\right] \\ &= \sum_{j=1}^{k^+} \left(\frac{1}{(n^+)^2} \sigma_+^2 + \frac{1}{(n^+)^2} \sigma_+^2 \right) + \sum_{j=1}^{k^-} \left(\frac{1}{(n^-)^2} \sigma_-^2 + \frac{1}{(n^-)^2} \sigma_-^2 \right) \end{aligned} \quad (17)$$

$$= \frac{2k^+}{(n^+)^2} \sigma_+^2 + \frac{2k^-}{(n^-)^2} \sigma_-^2 \quad (18)$$

As a sanity check: if $k^+ = k^- = 0$, then E_1 and E_2 are identical, meaning that they will have the same t -statistics, which means Q and hence R will be 0, as confirmed by this formula. Clearly this quantity **decreases** as k^+ and k^- increases, which means, as $r^+ = n^+ - k^+$ and $r^- = n^- - k^-$ **increase**, we expect E_1 and E_2 to agree more with respect to the biomarker g – either both agree it will be a biomarker, or both agree it will not.

Let $\text{BD}(E_1[r])$, (resp., $\text{BD}(E_2[r])$) be the set of biomarkers found (by our $\text{BD}(\cdot)$) for E_1 (resp E_2), when there are $r = [r^+, r^-]$ common elements, and let its complement $n\text{BD}(E_1[r]) = F - \text{BD}(E_1[r])$ and $n\text{BD}(E_2[r]) = F - \text{BD}(E_2[r])$ be the set of non-biomarkers, using F as the set of all features. The argument above suggests that $|\text{BD}(E_1[r]) \cap \text{BD}(E_2[r])|$ monotonically increases with r , as does $|n\text{BD}(E_1[r]) \cap n\text{BD}(E_2[r])|$.

Now consider the Jaccard score, and notice

$$J(\text{BD}(E_1[r]), \text{BD}(E_2[r])) = \frac{|\text{BD}(E_1[r]) \cap \text{BD}(E_2[r])|}{|F| - |n\text{BD}(E_1[r]) \cap n\text{BD}(E_2[r])|} \quad (19)$$

where $|\chi|$ is the total number of elements in the set χ . As r increases, we expect the numerator to increase, and the denominator to decrease, both of which means the ratio will increase.

(In general, we expect smaller values of $\text{Var}[R]$ to increase the chance that E_1 and E_2 will agree on the status of g . There is one exception here: say E_2 thinks that g is a biomarker as $t(E_2)$ is very negative, but E_1 thinks it is not a biomarker, as $t(E_1)$ is near 0. Here, imagine a variant of E_1 – call it E_1' – had a larger value of $t(E_1')$; large enough that it this value meant that g was considered a biomarker. Here E_1 and E_2 disagreed about g with their $Q = t(E_1) - t(E_2)$, but E_1' and E_2 agreed about g , despite having a larger value of $Q' = t(E_1') - t(E_2)$. This is very unlikely, as the numerator (and hence the sign) of $t(E_1)$ is basically an empirical estimate of $\mu_1 - \mu_2$, which means finding a large positive value for $t(E_1)$ is unlikely unless $\mu_1 \gg \mu_2$ – but this condition means it is very unlikely that $t(E_2)$ (which is another estimate of $\mu_1 - \mu_2$) will be very negative. That is, if τ is the threshold for declaring a feature to be a biomarker: $P(t(E_2) < -\tau | \mu_1 \gg \mu_2)$ is tiny. Similarly,

$$P(t(E_1) > \tau, t(E_2) < -\tau | \mu_1 \gg \mu_2) \ll \epsilon \quad (20)$$

$$P(t(E_1) > \tau, t(E_2) < -\tau | \mu_1 \approx \mu_2) \ll \epsilon \quad (21)$$

$$P(t(E_1) > \tau, t(E_2) < -\tau | \mu_1 \ll \mu_2) \ll \epsilon \quad (22)$$

where “ $\ll \epsilon$ ” means a very small probability. Equation 20 holds as $t(E_2) < -\tau$ is unlikely in this case; Equation 21 as both $t(E_2) < -\tau$ and $t(E_1) > \tau$ are each unlikely in this case; and Equation 22 as $t(E_1) > \tau$ is unlikely in this case.)

Final note: This derivation is for the general statement of Heuristic 7. Our specific oRS algorithm imposes an additional constraint – in essence, that E_1 will have two copies of each b_j and each e_j , and that E_2 will have two copies of each c_j and each f_j . The claim still holds – but with a variance for R that is twice as large as Equation 18: there are now only half as many b_j s (and c_j , e_j , f_j), but each is implicitly multiplied by $\frac{2}{n^+}$ or $\frac{2}{n^-}$, and variance goes with the square of that constant.

C.2 Motivation for Heuristic 9

Reusing the notation from above, assume D_1 and D_2 are each samples of size n , drawn from n^+ positive subjects and n^- negative subjects, independently for all of the features F . We also use BM to denote the set of all true biomarkers and NBM to denote the set of all true non-biomarkers.

For each gene g , our use of simple independent t-test means there are probabilities $p, q(n)$ such that

$$P(g \in \text{BD}_{t,p,no}(D) | g \in \text{NBM}) = p \quad (23)$$

$$P(g \in \text{BD}_{t,p,no}(D) | g \in \text{BM}) = q(n) \quad (24)$$

(Recall that $n = n^+ + n^-$ is the total number of subjects; we will insist on a constant ratio of $n^+ : n^-$ as we increase the total number of subjects.) Equation 23 is the definition of false-positive, which corresponds to the p-value of the t-test; Equation 24 is the “power” of the statistical test; we only need to know that $q(n)$ is monotonically increasing with n .

Given $E_1[n]$ and $E_2[n]$ are each based on n draws from the $\vec{p}$ distribution, we define the expected value

$$EJ(E_1[n], E_2[n]) = \mathbb{E} \left(\frac{|\text{BD}(E_1[n]) \cap \text{BD}(E_2[n])|}{|\text{BD}(E_1[n]) \cup \text{BD}(E_2[n])|} \right) \quad (25)$$

Heuristic 9 is claiming that this function is monotonically increasing as function of n .

To motivate this claim, we instead approximate this expectation of a ratio EJ (Equation 25) by the ratio of the expectations, $\tilde{J}$ (Equation 26, below), which Lemma 1 proves is monotonically increasing in n .

Lemma 1 Let E_1, E_2 each be drawn from $\vec{p}(\cdot)$ of size n – drawing $n_+ \in \mathbb{Z}^+$ instances from the joint distribution $[p_{1,1}(\cdot), \dots, p_{r,1}(\cdot)]$ associated with the outcome of $+$ -instances, and $n_- \in \mathbb{Z}^+$ instances from the joint distribution $[p_{1,2}(\cdot), \dots, p_{r,2}(\cdot)]$ associated with the outcome of $-$ -instances. Assuming that biomarkers are based on $\text{BD}_{t,p,no}(D)$, whose true positive rate is $q = q(n)$, and let

$$\tilde{J}(n; E_1, E_2) = \frac{\mathbb{E}(|\text{BD}(E_1) \cap \text{BD}(E_2)|)}{\mathbb{E}(|\text{BD}(E_1) \cup \text{BD}(E_2)|)} \quad (26)$$

Now assume that $q = q(n)$ is increasing in n , and that $q > p/2$. Then $\tilde{J}(n; E_1, E_2)$ increases as n increases

Proof: We first consider a single feature g , and consider the likelihoods that the two sample E_1 and E_2 will agree:

1. What is the chance that E_1 and E_2 both agree that g is a biomarker?

(1a) If g is a biomarker:

$$\begin{aligned} P(g \in \text{BD}(E_1), g \in \text{BD}(E_2) | g \in \text{BM}) &= P(g \in \text{BD}(E_1) | g \in \text{BM}) \times P(g \in \text{BD}(E_2) | g \in \text{BM}) \\ &= q(n) \times q(n) \end{aligned}$$

as the draws leading to E_1 and E_2 are independent.

(1b) If g is not a biomarker:

$$\begin{aligned} P(g \in \text{BD}(E_1), g \in \text{BD}(E_2) | g \notin \text{BM}) &= P(g \in \text{BD}(E_1) | g \notin \text{BM}) \times P(g \in \text{BD}(E_2) | g \notin \text{BM}) \\ &= p \times p \end{aligned}$$

2. What is the chance that E_1 and E_2 both agree that g is not a biomarker? Here again use $n\text{BD}(D) = F - \text{BD}(D)$ as the set of non-biomarkers found by BD – using F to represent the set of all features. Here we again deal first with the “ g is a biomarker”, then “ g is not a biomarker”, cases:

$$\begin{aligned} P(g \notin \text{BD}(E_1), g \notin \text{BD}(E_2) | g \in \text{BM}) &= P(g \notin \text{BD}(E_1) | g \in \text{BM}) \times P(g \notin \text{BD}(E_2) | g \in \text{BM}) \\ &= (1 - q(n)) \times (1 - q(n)) \\ P(g \in n\text{BD}(E_1), g \in n\text{BD}(E_2) | g \notin \text{BM}) &= P(g \in n\text{BD}(E_1) | g \notin \text{BM}) \times P(g \in n\text{BD}(E_2) | g \notin \text{BM}) \\ &= (1 - p) \times (1 - p) \end{aligned}$$

Hence, the expected size of the intersection of the biomarkers found by both samples (resp., NOT found by both) is

$$\begin{aligned}
 & \mathbb{E}[\text{BD}(E_1) \cap \text{BD}(E_2)] \\
 &= \sum_{g \in \text{BM}} P(g \in \text{BD}(E_1), g \in \text{BD}(E_2) | g \in \text{BM}) + \sum_{g \notin \text{BM}} P(g \in \text{BD}(E_1), g \in \text{BD}(E_2) | g \notin \text{BM}) \\
 &= |\text{BM}| \times p^2 + (|F| - |\text{BM}|) \times q(n)^2 \\
 & \mathbb{E}[n\text{BD}(E_1) \cap n\text{BD}(E_2)] \\
 &= \sum_{g \in \text{BM}} P(g \in n\text{BD}(E_1), g \in n\text{BD}(E_2) | g \in \text{BM}) + \sum_{g \notin \text{BM}} P(g \in n\text{BD}(E_1), g \in n\text{BD}(E_2) | g \notin \text{BM}) \\
 &= |\text{BM}| \times (1-p)^2 + (|F| - |\text{BM}|) \times (1-q(n))^2
 \end{aligned}$$

We now use the observation, from Equation 19, that $\text{BD}(E_1) \cup \text{BD}(E_2) = F - (n\text{BD}(E_1) \cap n\text{BD}(E_2))$, which means Equation 26 reduces to

$$\begin{aligned}
 \tilde{J}(n; E_1, E_2) &= \frac{|\text{BM}| \times p^2 + (|F| - |\text{BM}|) \times q(n)^2}{|F| - (|\text{BM}| \times (1-p)^2 + (|F| - |\text{BM}|) \times (1-q(n))^2)} \\
 &= \frac{t \times p^2 + s \times q(n)^2}{1 - [t \times (1-p)^2 + s \times (1-q(n))^2]}
 \end{aligned}$$

where $t = |\text{BM}|/|F|$ is the fraction of biomarkers (which is typically very small), and $s = 1 - t$ as the fraction of non-biomarkers. Also write $q = q(n)$.

Now note that $\frac{\partial \tilde{J}}{\partial n} = \frac{\partial \tilde{J}}{\partial q} \frac{\partial q}{\partial n}$, and recall that $\frac{\partial q}{\partial n} \geq 0$. Hence, to show that $\tilde{J}$ increases with n , it suffices to show that $\frac{\partial \tilde{J}}{\partial n} \geq 0$. First,

$$\frac{\partial \tilde{J}}{\partial q} = \frac{1 - [t(1-p)^2 - s(1-q)^2] \times 2sq - [tp^2 + sq^2] \times (-s) \times 2 \times (1-q) \times (-1)}{[1 - t(1-p)^2 - s(1-q)^2]^2}$$

As the denominator $[1 - t \times (1-p)^2 - s \times (1-q)^2]^2$ is positive, so we can ignore it. (Below we use $\propto$ to mean “same wrt sign”):

$$\begin{aligned}
 \frac{\partial \tilde{J}}{\partial q} &\propto [1 - t(1-p)^2 - s(1-q)^2] \times 2sq - [tp^2 + sq^2] \times 2s \times (1-q) \\
 &\propto [1 - t(1-p)^2 - s(1-q)^2] \times q + [tp^2 + sq^2] \times (q-1) \tag{27}
 \end{aligned}$$

$$\begin{aligned}
 &= (q - qt(1+p^2-2p) - sq(1+q^2-2q)) + [qp^2 + qs^2] - [tp^2 + sq^2] \\
 &= q - qt - qtp^2 + 2pqt - sq - sq^3 + 2sq^2 + qp^2 + sq^3 - tp^2 - sq^2 \\
 &= q - qt + 2pqt - sq + sq^2 - tp^2 \\
 &= q - qt + 2pqt - (1-t)q + (1-t)q^2 - tp^2 \\
 &= q - qt + 2pqt - q + tq + q^2 - tq^2 - tp^2 \\
 &= q^2 - t(q^2 + p^2 - 2pq) \\
 &= q^2 - t(q-p)^2 \\
 &\geq q^2 - (q-p)^2 = p(2q-p) \tag{28}
 \end{aligned}$$

$$\propto q - \frac{p}{2} \tag{29}$$

$$\geq 0 \tag{30}$$

where the “ $\propto$ ” in Line 27 as $2s > 0$; the “ $\geq$ ” in Line 28 follows from $t > 0$; the “ $\propto$ ” in Line 29 from knowing that p , and hence $p/2$, is > 0 , and the “ $\geq$ ” in Line 30 from our assumption that $q > \frac{p}{2}$. ■ (Lemma 1)

C.3 Empirical Evidence for Heuristics 7 and 9

Our empirical results, over 25 datasets, demonstrate that algorithms based on Heuristics 7 and 9 work effectively; see Tables 1 and 2. This subappendix provides further empirical evidence by running a number of simulations, based on realistic distributions of data. In particular, we form distributions based on each of 4 real-world datasets: $\Gamma = \{ \text{GDS968}, \text{GDS2771}, \text{GSE11121}, \text{GDS4185} \}$. (We selected these datasets to span a wide range of situations – in terms of number of subjects, number of features and the Reproducibility Scores; see Figure 5.) For each dataset, we compute the empirical mean and variance of each feature j

Figure C.5. Average underbound uRS, overbound oRS and true Jaccard histograms shown for 100 pairs of simulated datasets (for the true Jaccard) and 100 simulated datasets (for underbound and overbound), with distributions for each feature/outcome based on the empirical means and variances of features of 4 datasets { GDS968, GDS2771, GSE11121, GDS4185 }. Note that the x axis range is different for the different plots.

and for each outcome c , then define $p_{j,c}(\cdot)$ to be a univariate Gaussian with this mean and variance. We can then form new datasets by drawing $n_+(\rho)$ subjects of each $c = +$ outcome and $n_-(\rho)$ subjects of each $c = -$ outcome, matching the size of the original $\rho \in \Gamma$ dataset. (We will use $n(\rho)$ to be the total number of elements in that dataset.)

We then ran 3 experiments on each of the 4 distributions of data. The first experiment explored the claim that “Underbound $\leq$ True Jaccard $\leq$ Overbound”. Here, we simulated 100 pairs of datasets, for each of the 4 real-world datasets, then computed the average true Jaccard score between the biomarker sets found for each pair. We also computed the average uRS and oRS score for each of these simulated datasets. Figure C.5 shows the histograms of values for each of these 4 different distributions. As expected, we see no overlap between the histograms of the underbound, true and overbound, except for GDS968. Here, we found that the true Jaccard was higher than the overbound in just 8 of the 100 cases; and the largest difference here was 0.013 (and the average difference, over these 8, was only 0.004). Figure C.6 shows the median and interquartile range for each of the 3 algorithms {uRS, true_Jaccard, oRS} for each of the 4 distributions in Γ .

The second experiment explores Heuristic 7. For each distribution (corresponding to a dataset in $\rho \in \Gamma$), for each proportion $r \in \{0.1, 0.25, 0.5\}$, we generate 100 pairs of datasets $\{[D1_{\rho,r}^i, D2_{\rho,r}^i]\}_{i=1..100}$. For each $i \in \{1..100\}$, we first draw $n(\rho) \cdot r$ subjects – these will be in both datasets $D1_{\rho,r}^i$ and $D2_{\rho,r}^i$. We then complete each pair by drawing the remaining $n(\rho) \cdot (1 - r)$ subjects from the underlying distribution for $D1_{\rho,r}^i$, and then another $n(\rho) \cdot (1 - r)$ for $D2_{\rho,r}^i$. Hence, each of the ρ -datasets will have $n(\rho)$ subjects, and each $[D1_{\rho,r}^i, D2_{\rho,r}^i]$ will share r proportion of the subjects in common.

Figure C.7 shows the Jaccard scores of each pair of datasets. We see that, as Heuristic 7 predicts, as the number of subjects in common between a pair of datasets increases, so does the average Jaccard score of the discovered-biomarker sets.

The third experiment explores Heuristic 9. In this experiment, we generate 100 pairs of datasets with $n = |D|$ subjects (same as the associated dataset D), 100 pairs with $n \times 2$ and 100 pairs with $n/2$ subjects, and compute the average true Jaccard score for the discovered biomarkers. Figure C.8 shows that, as Heuristic 9 predicts, as the number of subjects increases, the average Jaccard score increases as well.

Figure 4(a) earlier presented similar results, over 5 different datasets, but there taking (disjoint) subsets of the original dataset (rather than generating new subjects from the estimated distribution), and by presenting mean/variance values, rather than median/interquartile values, over 10 database-sizes (rather than 3). See also the black $\bar{R}S$ line in Figure 4(b), for yet another dataset.

C.4 Overlap between Two Multi-Sets

This subappendix explores some relevant theoretical properties of the oRS algorithm: Recall this algorithm partitions the “doubled version” DD of the original dataset D , into 2 size- n sub(multi)sets, $D1$ and $D2$. Below we show that, in expectation, one-half of the elements of D will appear in both both $D1$ and $D2$ – i.e., $\mathbb{E}[|D1 \cap D2|] = n/2$. We then provide a natural notion

Figure C.6. Average underbound uRS, overbound oRS and true Jaccard values for 100 pairs of simulated datasets (for the true Jaccard) and 100 simulated datasets (for underbound and overbound), with distributions based on the empirical means and variances of feature/outcome pairs of 4 datasets – using box-and-whiskers plots containing the median and IQR.

of overlap between two multi-sets, and prove that the expected overlap between 2 random draws is 0, between the $D1$ and $D2$ defined above is $1/2$, and between two bootstrap samples is 1.

Expected Overlap between Multi-Sets produced by oRS

Let $DD = \{a_1, b_1, a_2, b_2, \dots, a_n, b_n\}$ denote a set of $2n$ distinct elements. Note $|D1 \cap D2|$ is the number of indices i such that exactly one of a_i and b_i is in $D1$. We provide a formal proof below. First, to provide the intuition: If the elements drawn in $D1$ were done independently, then clearly there is a 50% chance that any element of DD will appear in $D1$ and a 50% chance that another element will *not* appear in $D1$. If these events were independent, then the chance that $a_i \in D1$ and $b_i \notin D1$ would be $P(a_i \in D1, b_i \notin D1) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$. Similarly, $P(a_i \notin D1, b_i \in D1) = \frac{1}{4}$. Hence, the chance that exactly one of $\{a_i, b_i\}$ is in A is $P(a_i \in D1 \oplus b_i \in D1) = P(a_i \in D1, b_i \notin D1) + P(a_i \notin D1, b_i \in D1) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$.

To be precise, we need to deal with the observation that the draws from DD are not independent. Now let A be a size- n subset drawn uniformly from DD (corresponding to the $D1$ mentioned above). Let $\mathcal{P}_n$ denote the set of all size- n subsets of DD , equipped with the uniform distribution. For any $A \in \mathcal{P}_n$, let $O_n(A)$ denote the number of indices $1 \leq i \leq n$ such that exactly one of a_i or b_i is in A , so $O_n : \mathcal{P}_n \rightarrow \mathbb{R}$ is a random variable. We view a fixed $A \in \mathcal{P}_n$ as the outcome of drawing a random size- n subset of DD . Note $|O_n|/n$ is the fraction of DD that is in both subsets.

Lemma 2 Let O_n be defined as above. We have

$$\lim_{n \rightarrow \infty} \frac{\mathbb{E}[O_n]}{n} = \frac{1}{2}$$

monotonically from above.

Proof Note that $\#\mathcal{P}_n = \binom{2n}{n}$, where $\#$ denotes the cardinality of a set.

Suppose we are given distinct elements $Z = \{z_1, \dots, z_k\}$ where $k \leq n$. We compute the probability that $\{z_1, \dots, z_k\} \subset A$ where $A \in \mathcal{P}_n$ is a randomly chosen subset. Note that there is a bijection between size n subsets of DD containing $\{z_1, \dots, z_k\}$ and size $n-k$ subsets of $DD \setminus Z$ where given a subset E of $DD \setminus Z$ of size $n-k$, we get the corresponding subset $E \cup Z$ of DD . Thus the number of elements in $\mathcal{P}_n$ containing Z is $\binom{2n-k}{n-k}$, where

$$\begin{aligned} \frac{\binom{2n-k}{n-k}}{\binom{2n}{n}} &= \frac{\frac{(2n-k)!}{(n-k)!n!}}{\frac{(2n)!}{n!n!}} = \frac{\frac{(2n-k)!n(n-1)\dots(n-k+1)}{n!n!}}{\frac{(2n)!}{n!n!}} \\ &= \frac{(2n-k)!n(n-1)\dots(n-k+1)}{(2n)!} = \frac{n(n-1)\dots(n-k+1)}{2n(2n-1)\dots(2n-k+1)}. \end{aligned}$$

Figure C.7. Box and whiskers plot showing the average true Jaccard values for 100 pairs of simulated datasets for 4 different distributions, in 3 settings, where each pair has 10%, 25% or 50% of the subjects in common.

Moreover, the probability that Z is disjoint from a randomly chosen element of $\mathcal{P}_n$ is given by the same probability since for $A \in \mathcal{P}_n$, we have

$$Z \subseteq A \iff Z \cap (DD \setminus A) = \{\}$$

and the map $A \mapsto (DD \setminus A)$ is a bijection from $\mathcal{P}_n$ to $\mathcal{P}_n$ since $\#A = n$.

In particular, for a fixed index $1 \leq i \leq n$, the probability that $\{a_i, b_i\}$ is disjoint from a random subset of DD of size n is

$$\frac{1}{4} \cdot \left(\frac{n-1}{n-\frac{1}{2}} \right). \quad (31)$$

Now, let S_j be the indicator random variable for the event that either a_j or b_j (or both) are selected, i.e. for a set $E \in \mathcal{P}_n$,

$$S_j(E) = \begin{cases} 1 & : \{a_j, b_j\} \cap E \neq \{\} \\ 0 & : \{a_j, b_j\} \cap E = \{\} \end{cases}$$

In particular, the number of distinct elements in E , denoted $D_n(E)$, is just given by $D_n(E) = \sum_{j=1}^n S_j(E)$. Thus by linearity of expectation and Equation 31, we have

$$\begin{aligned} \mathbb{E}[D_n] &= \sum_{j=1}^n \mathbb{E}[S_j] = \sum_{j=1}^n \mathbb{P}(\{a_j, b_j\} \cap E \neq \emptyset) = \sum_{j=1}^n (1 - \mathbb{P}(\{a_j, b_j\} \cap E = \emptyset)) \\ &= \sum_{j=1}^n \left(1 - \frac{1}{4} \cdot \left(\frac{n-1}{n-1/2} \right) \right) \\ &= n \left(1 - \frac{1}{4} \cdot \left(\frac{n-1}{n-1/2} \right) \right) \end{aligned}$$

Let's now compute the expected overlap of the sets A and $DD \setminus A$ for $A \in \mathcal{P}_n$. If there are n_1 indices in A that occur exactly once, and n_2 indices that occur exactly twice, then

$$\begin{aligned} n &= \#A = n_1 + 2n_2 \\ D_n(A) &= n_1 + n_2. \end{aligned}$$

Figure C.8. Box and whiskers plot showing the average true Jaccard values for 100 pairs of simulated datasets for 4 different distributions, and for 3 different number of subjects (x-axis) relative to the original real datasets.

Since each index has two corresponding elements in DD , an index occurs in both A and $DD \setminus A$ if and only if it occurs exactly once in A . In particular, if $O_n(A)$ denotes the size of the overlap of A and $DD \setminus A$, we have $O_n(A) = 2D_n(A) - n$ and

$$\begin{aligned} o_n &:= \mathbb{E}[O_n] = 2\mathbb{E}[D_n] - n = 2n \left(1 - \frac{1}{4} \cdot \left(\frac{n-1}{n-1/2} \right) \right) - n \\ &= n \left(1 - \frac{1}{2} \cdot \left(\frac{n-1}{n-1/2} \right) \right) = n \left(\frac{n}{2n-1} \right). \end{aligned}$$

Clearly $\lim_{n \rightarrow \infty} \frac{o_n}{n} = \frac{1}{2}$, as claimed.

Defining Overlap between Multi-Sets, in General Heuristic 7 stated that the Jaccard score of biomarkers found based on two datasets, depends on the overlap between those datasets. The standard definition of overlap applies to the sets that we considered so far. To deal with bootstrap sampling, we need to extend that definition. We also show that, with this extension, we see that our oRS model has a smaller overlap than bootstrap sampling, explaining our empirical results that oRS is better than bRS.

To illustrate, let $c(x, S) =$ the number of times that x appears in the (multi-)set S . For example, if

$$S_1 = [2 \ 2 \ 8 \ 1 \ 8 \ 2 \ 6 \ 7 \ 6 \ 3]$$

then

$$c(1, S_1) = 1 \quad c(2, S_1) = 3 \quad c(3, S_1) = 1 \quad c(4, S_1) = 0$$

Assume $|S_1| = |S_2|$, and define:

$$\text{OVERLAP}(S_1, S_2) = \frac{1}{|S_1|} [\#(2 \in S_2) + \#(2 \in S_2) + \#(8 \in S_2) + \dots + \#(3 \in S_2)] = \frac{1}{|S_1|} \sum_x c(x, S_1) \times c(x, S_2)$$

Lemma 3 (a) If I_1 and I_1 are two size- n samples drawn independently from some continuous distribution over the reals (say from a mixture of Gaussians), then $\mathbb{E}[\text{OVERLAP}(I_1, I_2)] = 0$

(b) If O_1 and O_1 are two size- n samples produced by the oRS algorithm, from a size- n sample D , then $\mathbb{E}[\text{OVERLAP}(O_1, O_2)] = \frac{1}{2}$.

(c) If B_1 and B_1 are two size- n boot-strap samples from a size- n sample D , then $\mathbb{E}[\text{OVERLAP}(B_1, B_2)] = 1$.

Proof: (a) follows by realizing that the probability of any real value appearing in two size- n draws is effectively 0.

(b) follows from the observation that we can partition $X = X_{12} \cup X_1 \cup X_2$, where X_{12} contains elements that appear exactly 1

time in each of O_1 and O_2 , X_1 contains elements that appear 2 times in O_1 and 0 times in O_2 , and X_2 contains elements that appear 2 times in O_2 and 0 times in O_1 . Hence,

$$\begin{aligned}
 \mathbb{E}[\text{OVERLAP}(S_1, S_2)] &= \frac{1}{n} \sum_{x \in X} c(x, O_1) \times c(x, O_2) \\
 &= \frac{1}{n} \left[\sum_{x \in X_{12}} c(x, O_1) \times c(x, O_2) + \sum_{x \in X_1} c(x, O_1) \times c(x, O_2) + \sum_{x \in X_2} c(x, O_1) \times c(x, O_2) \right] \\
 &= \frac{1}{n} \left[\sum_{x \in X_{12}} 1 \times 1 + \sum_{x \in X_1} 2 \times 0 + \sum_{x \in X_2} 0 \times 2 \right] \\
 &= \frac{1}{n} [\mathbb{E}[|X_{12}|] + 0 + 0] \\
 &= \frac{\mathbb{E}[|X_{12}|]}{|X|} = \frac{1}{2}
 \end{aligned}$$

where the last line used Lemma 2, which proves that we expect X_{12} to contain half the elements of X .

(c) To prove (c),

$$\begin{aligned}
 \mathbb{E}[\text{OVERLAP}(B_1, B_2)] &= \frac{1}{n} \sum_{x \in X} \mathbb{E}[c(x, B_1) \times c(x, B_2)] \\
 &= \frac{1}{n} \sum_{x \in X} \mathbb{E}[c(x, B_1)] \times \mathbb{E}[c(x, B_2)] \tag{32}
 \end{aligned}$$

$$\begin{aligned}
 &= \frac{1}{n} \sum_{x \in X} 1 \times 1 \tag{33} \\
 &= \frac{1}{n} \times |X| = 1
 \end{aligned}$$

where Line 32 follows as B_1 and B_2 are independent bootstrap samples from X . To prove Line 33, note that $\sum_{x \in X} c(x, B) = n$ as some value appears in each position of $B \Rightarrow \mathbb{E}[\sum_{x \in X} c(x, B)] = n$.

Also $\mathbb{E}[c(x_a, B)] = \mathbb{E}[c(x_b, B)] \quad \forall x_a, x_b \in X$ by symmetry

$$\Rightarrow n = \mathbb{E}[\sum_{x \in X} c(x, B)] = n \times \mathbb{E}[c(x, B)] \Rightarrow \mathbb{E}[c(x, B)] = \frac{n}{n} = 1.$$

■