STRATEGIES FOR IMPROVING EPISTASIS DETECTION AND REPLICATION

Elizabeth Rachel Piette

A DISSERTATION

in

Genomics and Computational Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2018

Supervisor of Dissertation

____

Jason H. Moore

Professor of Biostatistics, Epidemiology, and Informatics

Graduate Group Chairperson

____

Li-San Wang

Associate Professor of Pathology and Laboratory Medicine

Dissertation Committee

John H. Holmes (Chair), Professor of Medical Informatics

Laura Almasy, Professor of Genetics

Hongzhe Li, Professor of Biostatistics

Scott M. Williams (External), Professor of Population and Quantitative Health Sciences (Case Western Reserve University)

STRATEGIES FOR IMPROVING EPISTASIS DETECTION AND REPLICATION

This is dedicated to my dearest husband Brian.

Your love makes life beautiful.

ACKNOWLEDGMENT

I extend my heartfelt appreciation to the many who have supported me in my academic endeavors and with their love and friendship.

To Jason, thank you for your kindness, sense of humor, and thoughtful perspectives as a scientist. I will always be grateful for your encouragement and aspire to do justice to your example.

To my thesis committee, John, Laura, Hongzhe, and Scott, thank you for your helpful advice and critiques. I truly appreciate your guidance and am honored to learn from your expertise.

To the students, postdocs, faculty, and staff of Moore Lab, IBI, and GCB, thank you for being there for me on a daily basis and helping me feel welcome at Penn. Special thanks to Maureen, Sallie, and Hannah for their invaluable helpfulness and organization.

To my mom, dad, brother, and the beloved friends I call family, thank you for your compassion, reassurance, and love. You bring me joy and strength.

To my husband Brian, I cannot overstate my love and appreciation. I am so fortunate to have a true partner in life.

This work was supported by National Institutes of Health grants LM009012, LM010098, LM12601, AI116794, DK112217, and ES013508.

ABSTRACT

STRATEGIES FOR IMPROVING EPISTASIS DETECTION AND REPLICATION

Elizabeth Rachel Piette

Jason H. Moore

Genome-wide association studies (GWAS) have been extensively critiqued for their perceived inability to adequately elucidate the genetic underpinnings of complex human phenotypes. Of particular concern is "missing heritability," or the difference between the total estimated heritability of a phenotype and that explained by GWAS-identified loci. There are numerous proposed explanations for this missing heritability, but a frequently ignored and potentially vastly informative alternative explanation is the contribution of epistatic interactions underlying complex phenotypes.

Given our understanding of how biomolecules interact in networks and pathways, it is not unreasonable to conclude that the effect of variation at individual genetic loci may non-additively depend on and should be analyzed in the context of their interacting partners. It has been recognized for over a century that deviation from expected Mendelian proportions can be explained by the interaction of multiple loci, and the epistatic underpinnings of phenotypes in model organisms have been extensively experimentally quantified. Therefore, the dearth of inspiring single locus GWAS hits for complex human phenotypes (and the inconsistent replication of these between populations) should not be surprising, as one might expect the joint effect of multiple perturbations to interacting partners within a functional biological module to be more impactful than individual main effects.

Current methods for analyzing GWAS data are not well-equipped to detect epistasis or replicate significant interactions. The multiple testing burden associated with testing each pairwise interaction quickly becomes nearly insurmountable with increasing numbers of loci. Statistical and *machine learning approaches that have worked well for other types of high-dimensional data are appealing and may be useful for detecting epistasis, but potentially require adaptations to suit interaction analyses. Biological knowledge may also be leveraged to guide the search for epistasis candidates, but requires context-appropriate application (as, for example, two loci with significant main effects may not have a significant interaction, and vice versa).*

Rather than renouncing GWAS and the wealth of associated data that has been accumulated as a failure, I propose the development of new techniques and incorporation of diverse data sources to analyze GWAS data in an epistasis-centric framework.

TABLE OF CONTENTS

LIST OF TABLES

IMPROVING THE REPRODUCIBILITY OF GENETIC ASSOCIATION RESULTS USING GENOTYPE RESAMPLING METHODS

IMPROVING MACHINE LEARNING REPRODUCIBILITY IN GENETIC ASSOCIATION STUDIES WITH PROPORTIONAL INSTANCE CROSS VALIDATION (PICV)

IDENTIFICATION OF EPISTATIC INTERACTIONS BETWEEN THE HUMAN RNA DEMETHYLASES FTO AND ALKBH5 WITH GENE SET ENRICHMENT ANALYSIS INFORMED BY DIFFERENTIAL METHYLATION

LIST OF ILLUSTRATIONS

ENHANCING THE REPRODUCIBILITY OF MACHINE LEARNING ANALYSES OF GENOMIC DATA

IMPROVING THE REPRODUCIBILITY OF GENETIC ASSOCIATION RESULTS USING GENOTYPE RESAMPLING METHODS

IMPROVING MACHINE LEARNING REPRODUCIBILITY IN GENETIC ASSOCIATION STUDIES WITH PROPORTIONAL INSTANCE CROSS VALIDATION (PICV)

IDENTIFICATION OF EPISTATIC INTERACTIONS BETWEEN THE HUMAN RNA DEMETHYLASES FTO AND ALKBH5 WITH GENE SET ENRICHMENT ANALYSIS INFORMED BY DIFFERENTIAL METHYLATION

ENHANCING THE REPRODUCIBILITY OF MACHINE LEARNING ANALYSES OF GENOMIC DATA

Introduction

As the falling cost of DNA sequencing continues to outpace Moore's Law, and new parallel and distributed computing paradigms and technologies arise to manage the rapid accumulation of big data, biomedical scientists are faced with new practical and analytical challenges including data storage, merging heterogeneous data types from diverse sources, handling differentially missing data, and more [87, 138]. The increasingly high-throughput nature of complex biomedical studies pairs well with analysis via machine learning methods, which in general aim to automatically improve performance of a task through experience [57, 71]. Machine learning has gained incredible traction in the biomedical sciences and adjacent fields, evidenced by numerous fruitful applications to diverse issues in genetics and genomics such as identifying binding sites from sequences, functionally annotating hierarchical gene ontologies, building gene expression networks for regulatory context, and more [78]. However, as machine learning methods continue to increase in popularity and accessibility, it is critical to emphasize the importance of thoughtfully considering how choices at each stage can impact a given analysis in the context of the unique data challenges associated with big biomedical data. This review discusses reproducibility and replicability in general before providing overviews of current issues in and methods for enhancing reproducibility in the contexts of genome-wide association studies (GWAS) and machine learning, both individually and in conjunction. Analyzing GWAS data beyond the traditional interrogation of single-locus main effects to consider interactions or incorporate additional complementary data sources may facilitate the discovery of new insights from pre-existing data. Finally, this review concludes with a discussion of future directions for the field and suggestions for challenging current paradigms in both genomics and machine learning to better integrate the two to promote high-quality, reproducible science.

Reproducibility vs. replicability

Variations on the terms “reproducibility” and “replicability” are used, sometimes interchangeably, to broadly refer to the ability to achieve the same experimental results or arrive upon the same conclusions given repetition of the same or similar experiment or analysis. This reproducibility is generally accepted to be the defining property of robust scientific research and provides the basis upon which future research depends. Despite the collective emphasis on the importance of reproducibility/replicability in conducting research, which has been particularly acute in the biomedical sciences in recent years given the explosion of data and the need to perform analyses across bench lab conditions and computing environments, distinct scientific traditions have proposed and embraced varying definitions over time of what these terms refer to and the relative importance of the concepts they describe [53, 94, 107-9, 119]. An early definition of reproducibility in the era of computational analyses orients us to an example of a major desired quality of reproducible analysis, the ability to redraw a figure from the data and software provided with a paper:

> “A revolution in education and technology transfer follows from the marriage of word processing and software command scripts. In this marriage an author attaches to every figure caption a pushbutton or a name tag usable to recalculate the figure from all its data, parameters, and programs. This provides a concrete definition of reproducibility in computationally oriented research.” [38]

Over time, the definition of reproducibility has expanded and become more nuanced. The following semantic distinctions can help contextualize these terms. However in common usage these distinctions may be ignored, either term may refer to the same overarching concept, their definitions may be flipped, or “reproducibility” may generally be preferred in computational analyses and “replicability” in biological and other traditional experimental sciences [44, 95].

Reproducibility: Research may be considered reproducible if the data collected by the original researcher and the analyses applied to that data (such as the software code used for data cleaning, statistical analyses, and figure generation) can be re-run by another individual to produce the same numerical results, figures, etc. as in the original analysis.

Replicability: Research may be considered replicable if independently-performed data collection and analyses lead to the same overall conclusion as the original research.

Therefore, replicability may be considered to be more robust than reproducibility in lending support to a scientific finding, as it refers to independent confirmation from differing experimental conditions. However, improving both replicability and reproducibility is necessary and desirable for distinct reasons that feed into each other, especially in the current computational research environment. For example, reproducing a certain experiment may be infeasible due to lack of adequate computational resources; however, adhering to the principles of reproducible research by sharing the code used in this analysis may allow an external researcher to analyze a different data set with the same software to replicate the original finding.

Regardless of distinctions in definitions, in recent years the scientific communities across a number of fields have acknowledged and struggled to address the reproducibility/replicability crisis. Large-scale efforts to replicate prior experiments, analyses of experiments that have had follow-up studies, and surveys of attempts to reproduce previously-conducted research have found high rates of failure to replicate even one's own experiments [4, 99, 110]. Explanations for the reproducibility crisis are distributed across many steps and agents in the scientific process with suggestions for improvements ranging from addressing the ways in which grants are reviewed to how results are disseminated, but generally agreed upon top contributing factors include "failure to adhere to good scientific practice and the desperation to publish or perish" [12]. The latter concern requires institutional changes outside the scope of this paper, although advances in scientific data dissemination such as with journals that publish data descriptions and negative results are an improvement. The era of high-throughput experiments has produced unprecedented volumes of data which may be analyzed in widely varying computational environments, necessitating new definitions of "good scientific practice" for this research landscape. In order to improve the reproducibility and replicability of machine learning analyses of genomic data, it will be necessary to acknowledge contributing factors of failure to reproduce/replicate in both of these fields and develop novel, integrative solutions.

Replicability of genome-wide association studies (GWAS)

GWAS aim to identify genetic variants associated with a phenotype, such as single nucleotide polymorphisms that differ in frequency between cases and controls. Replication of a genotype-phenotype association in another cohort is considered necessary to substantiate GWAS findings, but there are numerous reasons why findings may fail to replicate despite a true association [33, 54]. The significance threshold for GWAS is Bonferonni-corrected for the estimated number of independent tests of association being performed based on linkage disequilibrium, generally 5×10^{-8} for individuals of European descent and lower for other populations with greater genetic diversity [70, 104]. Kraft et al. provide a helpful description of a Bayesian framework for considering evidence of replicability, most importantly highlighting that "the probability that an observed association truly exists in the sampled population depends not only on the observed p-value for association, but also the power to detect the association (a function of minor allele frequency, effect size and sample size), the prior probability that the tested variant is associated with the trait under study, and the anticipated effect size" [74]. Therefore, at the widely-accepted GWAS significance threshold, differences in minor allele frequency, effect size, and sample size between the populations used in the discovery and replication cohorts may all contribute to failure to replicate, but these factors are often not used to adjust the significance threshold accordingly. Replicating GWAS findings of gene-gene interactions is even more difficult than for single variants due to the increased multiple testing burden associated with naively testing each combination of loci, but analyses that move beyond single variants are likely to become more important in light of the growing disappointment in the amount of phenotypic heritability explained by individual loci and the acknowledgement of the distribution of risk across the genome [90, 96]. In addition to these statistical concerns, differences in genetic architecture and environmental exposures between populations, differences in unmeasurable confounders between studies, or technical sources of failure to replicate such as differences between genotyping platforms or quality control protocols may all diminish the replicability of GWAS findings.

Methods for improving GWAS replicability

GWAS replication may be improved at the design stage when planning initial data collection and analysis, or increasingly commonly in conducting secondary analyses that may combine pre-existing GWAS data with other data sources. In the design stage, choosing a replication cohort of a sufficient size that is of the same ancestry as the discovery cohort is generally recommended to avoid the effects of population stratification, although effects that are replicated across different populations may be considered more robust [91, 111]. Care should also be taken to ensure consistent and accurate phenotyping between cohorts, as lack of phenotype harmonizing can result in perceived poor replication due to making comparisons of associations with two overlapping but distinct phenotypes; this is likewise a concern in performing meta-analysis of a GWAS phenotype [13].

Methods for improving GWAS replication may be most fruitful if they aim to replicate network- or pathway-based findings rather than single variant main effects [47, 67-9, 92, 105, 147]. Given the complexity of the genome and of many common diseases/phenotypes, analysis of the cumulative effects of numerous perturbations to a pathway may be more informative and replicable than the effects of individual variants. These may also benefit from integration of additional data types such as methylation or expression data, or incorporation of expert knowledge, to provide prior knowledge or to use for dimensionality reduction to minimize the number of hypotheses being tested [39, 61, 130]. Collecting and analyzing multiple sources of data in conjunction is becoming increasingly feasible, and seems a logical next step for attempting to explain more of the genetic variation of a trait than GWAS alone. Performing secondary analyses that repurpose data from multiple sources, which has been both maligned and praised as “research parasitism”, is likely to become increasingly popular given the modern emphasis on data sharing and decreasing financial barriers to accessing adequate computational resources [9, 85, 118]. Improving the methods by which diverse data types are consolidated, the ways in which machine learning methods and parameters are chosen to suit the data and shape the analytic pipeline, and good software and computational environment sharing practices will shape the future of genomic analyses, and thoughtful guidelines surrounding each choice will be necessary to prevent the propagation of errors.

Reproducibility of machine learning analyses

As lack of replicability across many traditional scientific fields has reached the point of crisis, improving the reproducibility of computational analyses has concurrently emerged as a solution for minimizing or eliminating many of the preventable inconsistencies that hinder replicable research [43, 125]. Version control repositories and code hosting platforms such as GitHub enable sharing exact code and data sets used for analyses, reducing the need for re-implementation of algorithms to solve problems that have already been well-realized and improving the ability to suggest and disseminate new versions [84]. Using tools such as Docker that consolidate entire computational environments, or sharing a cloud instance between collaborators at multiple research sites rather than performing multiple individual analyses with different high performance clusters, removes concerns regarding software versioning and inconsistent analytic pipelines or quality control steps between sites [32, 45, 64, 93]. These are particularly relevant to GWAS analyses because seemingly minor differences in data pre-processing or model training may result in different top hits.

Methods for improving machine learning analysis reproducibility

Reproducibility of computational analyses is largely thought of in terms of implementing the same exact analysis on the same data in the same computational environment, but if improving replicability of GWAS findings is the goal, it may be as important to consider how a machine learning analysis of one data set may need to be altered to replicate a finding in another cohort, which requires understanding how the choices made in a given analysis can impact the result. For example, a discovery data set may have observations missing at random, and a replication data set may have observations missing not at random in a modelable way, which could warrant applying different imputation techniques. In this sense, reproducibility may be seen as a way to help facilitate future replication. The following section documents how choices the researcher may make at each stage of a typical machine learning analysis may impact the ultimate analytic conclusion. Figure 1 provides an overview of a generic supervised machine learning workflow. Although accompanied by examples of applications to GWAS analysis or the biomedical sciences more broadly, these concerns are overarching and may be translatable across many research domains wherever large, complex data is analyzed.

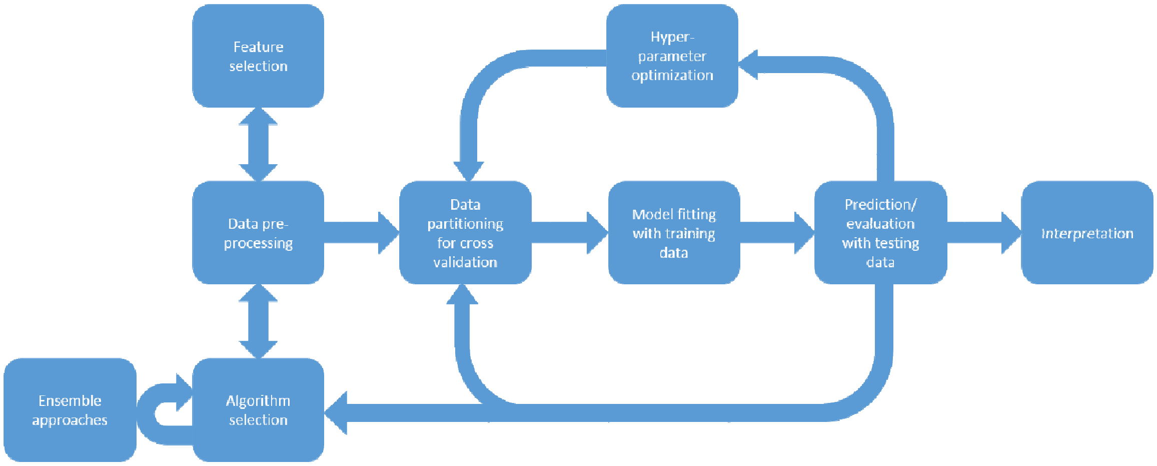

Feature selection
Hyper-parameter optimization
Data pre-processing
Data partitioning for cross validation
Model fitting with training data
Prediction/ evaluation with testing data
Interpretation
Ensemble approaches
Algorithm selection

Figure 1. A generic supervised machine learning analysis. Machine learning workflows are often iterative with interrelated stages, so comprehensive or rigid definitions are unsuitable. For example, the researcher may already have a particular algorithm in mind and perform imputation to ensure the data is in a suitable format free of missingness prior to feature selection. Alternatively, the researcher may perform feature selection prior to data pre-processing if unimputed variables with low missingness are considered higher quality or more potentially informative.

Choosing between algorithms

Algorithm choice may be driven by a combination of the overarching goals of the analysis, number and structure of predictors, presence of interactions or non-linearity, time and computational constraints, and perhaps at the most basic level the availability of labeled data. Machine learning approaches are typically classified as either supervised or unsupervised depending upon whether the learner uses labeled training data, although there also exist semi-supervised approaches in which a subset of the data is labeled and the rest remains unlabeled and reinforcement learning approaches in which [146]. In a typical supervised analysis, an algorithm is applied to a portion of a data set designated for use in the “training” stage of a classification or regression task, after which the model derived from this training is applied to held out data in the “testing” stage and performance is evaluated. Ideally the model will be generalizable enough that it reasonably captures underlying trends in the data and will perform well for data that it has not encountered during training, avoiding the “overfitting” that results from overly complex models or irrelevant predictors that pick up on the noise inherent in the training data [50]. Some classes of algorithms have become largely acknowledged as being particularly well-suited to certain applications based on complimentary characteristics of the algorithm and data, such as support vector machines for functional classification of genes from gene expression data [25]. Unsupervised analyses attempt to uncover structure in data lacking labels, for example k-means clustering of gene expression data to identify novel tumor subtypes with clinical significance [128]. A number of helpful sources exist that provide guidance regarding algorithm selection considering the analytic goal, computational burden, number of features, data sparseness, etc. [48, 121].

Ensemble approaches

Considering the “no free lunch” theorem that essentially posits no single algorithm will be optimal for all applications, ensemble approaches in machine learning that aggregate the decisions of multiple predictors are appealing [42, 139]. Early forays into model combination in the statistical and especially economics fields laid the basis for the development of Bayesian model averaging, in which the posterior distribution of each potential model is weighted by its posterior model probability [6, 8, 63, 77]. Although this is an appealing way to circumvent the uncertainty associated with model specification, it may be computationally expensive or infeasible to implement as the number of models increases, and estimating the posterior probability can also be a challenge. See Yeung et al. for an implementation of Bayesian model averaging in the context of gene selection and classification of microarray data [143]. Bagging, short for bootstrap aggregating, is a popular ensemble approach in which new training sets are generated via draws with replacement from the original training set and models are fitted to each of these then subsequently aggregated via averaging or voting [22]. Dudoit and Fridlyand describe a combined bagged clustering procedure for identifying tumor subtypes from gene expression profiles [46]. Boosting is another popular ensemble approach in which many weak classifiers are iteratively learned and weighted relative to their performance to create a single strong one, for example via their linear combination [21, 120]. Niu et al apply AdaBoost, one of the most popular boosting algorithms, to a protein structural class prediction problem [101].

Hyperparameter optimization

The performance of machine learning methods that require hyperparameters, which are parameters of the algorithm itself assigned before any training takes place (such as number of trees in a random forest or layers in a neural network), may be largely impacted by their choice [17, 127]. Popular hyperparameter optimization algorithms include grid search and Bayesian optimization. A grid search, or parameter sweep, is an exhaustive search of the combinations of hyperparameter values within user-defined bounds. The hyperparameter combination that best optimizes a user-specified loss function, for example prediction accuracy on held-out testing data, is chosen. Grid search may be computationally expensive due to its exhaustiveness, which random search circumvents via instead sampling a subset of the hyperparameter combination space, but may be readily parallelized [17]. Grid search may be particularly relevant to GWAS simulation applications that aim to characterize and compare performances with machine learning methods across a range of scenarios in which factors such as minor allele frequency, heritability, prevalence, etc. are varied. Bayesian optimization assigns a (generally Gaussian) prior to the loss function, uses an acquisition function to determine where to sample from in the loss function, then updates the posterior to determine the next sampling location [124, 23]. The researcher may alternatively define more meaningful prior distributions that incorporate additional meta-data about the experiments or data sets. Since Bayesian optimization does not require searching the entire hyperparameter space, it also reduces the computational resources required.

Data pre-processing

Multiple iterations of pre-processing may be necessary prior to conducting a machine learning analysis in order to render the data into an appropriate format for the algorithm of interest, reduce dimensionality, combine multiple disparate sources of data, etc. Algorithms often require data sets to be free of missingness in order to run, in which case the researcher must consider the type of missingness and decide upon a strategy for addressing missing observations. Little and Rubin provide a definitive guide to handling missing data [81]. Even if the algorithm implementation in question includes an embedded approach to produce a complete data set, or alternatively the researcher has decided to choose an approach that allows for missingness and avoids imputation, it is important to consider the pattern and mechanism of the missingness and the biases specific to the experimental conditions that produced said data. Pre-processing in GWAS requires multiple steps of quality control to remove poor quality samples, SNPs, and batch effects, and different threshold choices at each step and even the order in which the steps are performed may all impact the resulting top hits and may be difficult to coordinate even between sites working on the same data set [129]. A Jupyter notebook that allows users to execute and supervise an automated cloud-based quality control pipeline can both allow for quality control harmonization between sites, and for the alteration and sharing of pipelines to better suit individual analytic needs. Data pre-processing to incorporate multiple heterogeneous data sources may be an even greater challenge, requiring pre-processing not only of each data source individually, but in consideration of whether and how data will be merged and analyzed [114].

Feature selection

Feature selection is often a critical component of analyses of data sets containing large numbers of variables, and can help improve predictions by utilizing the most informative features, reduce the time and computational resources needed to run an algorithm, enhance interpretability and visualization, and avoid the "curse of dimensionality" [14]. Many commonly-used methods are based on either filters that rank variables based on intrinsic data qualities such as correlation with other variables, wrappers that assess groups of variables as they optimize accuracy in a predictive model, or combined and embedded methods [55, 65, 117]. Feature selection may be impacted by pre-processing so it may be beneficial to consider how the two may play in to each other, for example by weighting variable rankings according to a data quality score reflective of missingness. Feature selection is incredibly pertinent to GWAS analyses in order to circumvent the high multiple testing burden. SNPinfo is an example of a SNP selection tool that incorporates functional predictions, linkage disequilibrium, and GWAS results to select relatively small numbers of SNPs for GWAS [140]. Cantor et al. provide a review of how meta-analysis, epistasis testing, and pathway analysis can be used to prioritize GWAS results [28].

Cross validation

Implementing a supervised machine learning analysis generally constitutes training on a subset of the data and evaluating performance on a held-out test set, which allows for unbiased evaluation of a model's performance and helps avoid overfitting [112]. The simple hold-out method is a popular way to choose training and testing partitions by randomly dividing the data into two non-overlapping sets. Also popular are the number of variations on *k*-fold cross validation, in which the data is divided into *k* subsets, rounds of training are performed on each subset and tested on all remaining data, and performance is estimated by averaging across all rounds . In addition to choices regarding how to partition the data and how many folds and/or repetitions should be used, the cross validation method chosen should be appropriate for the data and application particularly regarding desires to minimize bias or variance [133]. For example the repeated *k*-fold cross validation method is biased by overlap between training and testing sets. Additionally, bias is introduced by performing feature selection or other classifier training or parameter selection outside of the cross validation procedure; these must be performed separately within each round. Cross validation can be affected by class imbalance with regard to the dependent or independent variables, and re-weighting and re-sampling methods have been developed to address the inconsistencies that may arise because of this [5, 24]. In GWAS analyses, consistent cross validation results from any number of supervised algorithms may be used to select subsets of SNPs or interaction models [11, 60].

Interpretation

The choice of algorithm and its associated assumptions can greatly impact interpretability of results from machine learning analyses. The enduring popularity of linear models despite the potential accuracy improvement that may be gained from applying more complicated “black box” models may be partially ascribed to their ease of interpretation. Therefore, the researcher should consider whether the aim of an analysis is solely to maximize accuracy or some other performance measure, or if meaningful knowledge concerning the model features is desired. The gap in our current ability to apply some machine learning methods very accurately yet fail to be able to adequately interpret them may be reduced by improvements in the intersection of dimensionality reduction and visualization techniques [135].

Conclusion and future directions

Improving the reproducibility of machine learning analyses of genomic data will require thorough, rigorous interrogation of how decisions at each stage of a machine learning analysis impacts results across a range of data types with varying biases to ensure that analyses are paired with the most appropriate choices of algorithm, hyperparameters, meaningful features, etc. Large benchmarking efforts that explore which machine learning methods are best suited to data sets of varying complexity, size, numbers of features, and feature types are already under way [103]. Extending this paradigm to benchmark as many publicly available genomics data sets as possible can improve the basis upon which artificial intelligence methods can suggest data analysis pairings [102]. Improved standardization of data collection tools to feed directly into analysis may help further streamline the process overall and help reduce careless errors related to data entry or transfer. The future of genomic data analysis is likely to be increasingly integrative, with GWAS as one of many data sources that must be processed and linked to be used in machine learning analyses.

IMPROVING THE REPRODUCIBILITY OF GENETIC ASSOCIATION RESULTS USING GENOTYPE RESAMPLING METHODS

Introduction

Replication is the gold standard for substantiating the validity of results across the spectrum of biological sciences, and is a cornerstone of rigorous hypothesis-driven research. In this era of big data and complex, computationally-intensive research, true replication may be impossible or infeasible, and reproducibility of analyses is a proximate concern [107]. Both replication and reproducibility are beset with challenges associated with a diversity of issues ranging from data access and storage, to availability of requisite computational resources, to thoughtfully implemented, high-quality code, all in the context of a constantly shifting field with high software, hardware, and ideological turnover. While advances such as portable, versioned workflows for computational environments and proposed statistical frameworks for defining replication and reproducibility themselves are addressing some of these issues, certain roadblocks to replication and reproducibility have yet to be resolved and may continue to remain impractical and inaccessible to the average researcher, such as for the analysis of data sets involving millions of parameters, multiple processers, and finite time [19, 75, 106, 116].

In the context of genome-wide association studies, failure to replicate previously-observed findings in a second population may be attributable to a combination of statistical and biological factors. Investigating the genetic underpinnings of complex diseases presents a special challenge given the evidence for multi-locus or network-based models of disease and the increased multiple-testing burden associated with fitting interaction models over single-locus models [90]. This is in addition to considerations of the heterogeneity of disease etiology, underlying genetic architecture, and other confounding factors that vary across populations. In this study, we explore epistatic interactions as a case study of a phenomenon that may be inherently difficult to replicate, and attempt to recapitulate the power to detect epistatic interactions between single nucleotide polymorphisms (SNPs) in two populations with differing minor allele frequencies (MAFs).

Epistasis, briefly defined as interactions between genetic loci that non-additively contribute to phenotype, is suspected to be both ubiquitously implicated in susceptibility to non-Mendelian disease and difficult to detect and replicate [96]. Resampling populations so that they appear more similar for the genotypes of interest may allow us to compare them in a more meaningful way. In this study, we propose a method for improving detection of epistatic SNP-SNP interactions between genome-wide association study (GWAS) data sets with differing minor allele frequencies for the SNPs of interest via resampling by genotype such that genotypes that are underrepresented in the replication population relative to the discovery population are oversampled, and genotypes that are overrepresented are undersampled. We substantiate the efficacy of this method via simulations. Application of this method may help inform scenarios in which findings of interest with potential functional significance from a discovery population sample fail to reach statistical significance in a replication population sample.

Data sets and methods

The following subsections describe our methods workflow (refer to Fig. 1 for accompanying graphical abstract). Briefly, we begin with data set simulation for a selection of models with varying penetrance functions, minor allele frequencies for two SNPs, heritabilities, and prevalences. Then, we use the discovery penetrance tables to generate replication data with the same underlying penetrances but differing minor allele frequencies. Next, we analyze the SNP-SNP interactions for all discovery scenarios by calculating the p-value for the likelihood ratio test comparing the logistic regression models with and without the interaction between the two SNPs, and estimate power to detect the interaction over 1000 simulations. Replication data sets are resampled to match the genotype proportions of their relative discovery data sets, and interaction analysis and power estimation is performed again post-resampling. We also test negative simulations to address the possibility of erroneously significant interactions – sample data sets with significant p-values for the interaction, despite being drawn from an underlying population without a significant interaction.

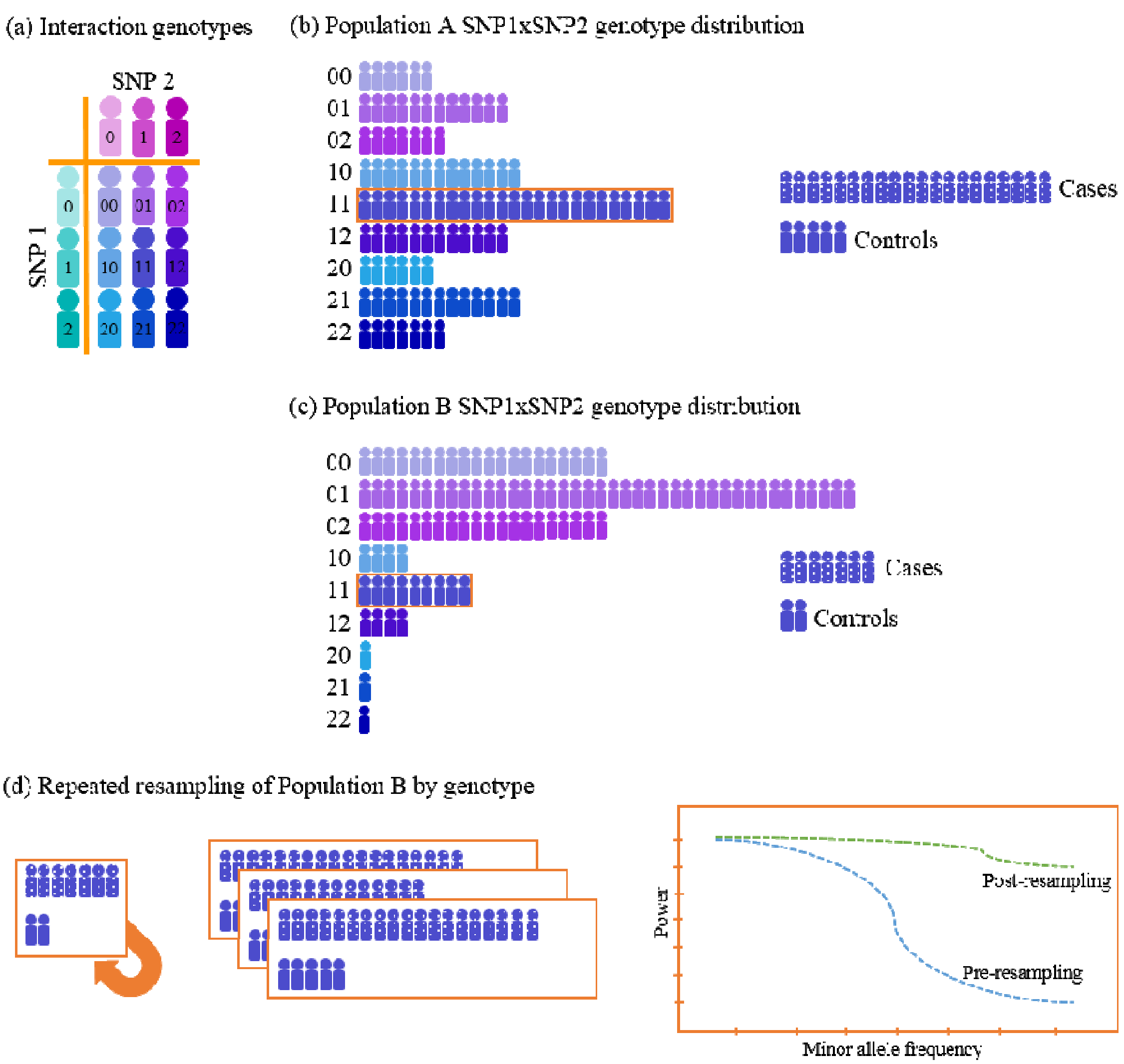

(a) Interaction genotypes
SNP 2
SNP 1
0
1
2
00
01
02
10
11
12
20
21
22
(b) Population A SNP1xSNP2 genotype distribution
Cases
Controls
(c) Population B SNP1xSNP2 genotype distribution
Cases
Controls
(d) Repeated resampling of Population B by genotype
Power
Minor allele frequency
Post-resampling
Pre-resampling

Figure 3. Graphical abstract (a) A SNP-SNP interaction results in 9 genotypes (b) In Population A, the SNP1 minor allele frequency is 0.5, and the SNP2 minor allele frequency is also 0.5. A GWAS of Population A reveals a significant association between the ‘11’ genotype and disease status. (c) Replication of this interaction is sought in Population B, and another GWAS is performed. However, in Population B, the SNP1 minor allele frequency is 0.1, and the SNP2 minor allele frequency is 0.5, so the relative distribution of genotypes is different. A GWAS of Population B does not reveal a significant association between the ‘11’ genotype and disease status, despite the same penetrance for genotype ‘11’ in Population B and in Population A, due to the low minor allele frequency of SNP1/low prevalence of genotype ‘11’ in Population B. (d) Resampling by genotype allows us to observe what our Population B sample would look like if the minor allele frequencies for SNP1 and SNP2 were the same as in Population A. Performing resampling numerous times allows for an empirical estimation of power to detect a significant interaction.

Discovery data set simulation

Penetrance functions and data sets for eight discovery scenarios were generated using GAMETES, an algorithm and software package that facilitates generation of complex epistatic models and data sets based upon these models [131]. Our test parameters (Table 1) included minor allele frequencies of 0.5 and 0.5 or 0.1 and 0.1 for two SNPs, heritabilities of 0.05 or 0.005, and prevalences of 0.5 or 0.1. Case-control data sets of size 2,000 and 4,000 were tested, and all simulation scenarios were replicated 1,000 times.

Replication data set simulation

Replication data sets were generated using the discovery penetrance tables to create data sets with the same underlying penetrances but differing minor allele frequencies. First, each simulated individual is assigned a value of 0, 1, or 2 for SNP 1 genotypes, with probabilities corresponding to genotype frequencies in Hardy-Weinberg equilibrium. This is repeated for SNP 2. Then, each simulated individual is assigned their case-control status based on their assigned values for SNP1 and SNP2 and the corresponding penetrance for that genotype from the discovery penetrance function. All discovery scenarios had corresponding replication scenarios with all two-SNP minor allele frequency combinations of {0.5, 0.4, 0.3, 0.2, 0.1}. We also include a finer resolution version replicating the discovery scenario with minor allele frequencies of 0.5 and 0.5, heritability of 0.005, and prevalence of 0.5 (Model 1 from Table 1) with replication SNP1 minor allele frequency fixed at 0.5, and SNP2 minor allele frequency from 0.5 to 0.01 by 0.01. 1,000 data sets of sizes 2,000 and 4,000 were generated for each replication scenario.

Table 4. Discovery data set simulation parameters: minor allele frequencies, heritabililties, prevalences, and penetrance tables used to generate discovery data

	Model 1			**Model 2**		
SNP1 MAF:	0.5			0.5		
SNP2 MAF:	0.5			0.5		
Heritability:	0.005			0.05		
Prevalence:	0.5			0.5		
Penetrance:	0.475	0.469	0.586	0.495	0.646	0.214
	0.524	0.516	0.443	0.449	0.451	0.649
	0.476	0.498	0.528	0.608	0.451	0.498
	Model 3			**Model 4**		
SNP1 MAF:	0.5			0.5		
SNP2 MAF:	0.5			0.5		
Heritability:	0.005			0.05		
Prevalence:	0.1			0.1		
Penetrance:	0.145	0.092	0.071	0.130	0.105	0.061
	0.065	0.110	0.115	0.081	0.145	0.029
	0.125	0.088	0.098	0.108	0.006	0.281
	Model 5			**Model 6**		
SNP1 MAF:	0.1			0.1		
SNP2 MAF:	0.1			0.1		
Heritability:	0.005			0.05		
Prevalence:	0.5			0.5		
Penetrance:	0.507	0.475	0.407	0.524	0.395	0.458
	0.467	0.624	0.906	0.387	0.999	0.689
	0.548	0.274	0.692	0.604	0.033	0.462
	Model 7			**Model 8**		
SNP1 MAF:	0.1			0.1		
SNP2 MAF:	0.1			0.1		
Heritability:	0.005			0.05		
Prevalence:	0.1			0.1		
Penetrance:	0.097	0.113	0.102	0.115	0.036	0.029
	0.117	0.024	0.093	0.031	0.394	0.417
	0.035	0.391	0.092	0.118	0.017	0.128

Interaction analysis

For all replicates of all discovery scenarios, we calculated p-values for the likelihood ratio test comparing the logistic regression models with and without the interaction term for the two SNPs of interest, where the two models being compared are:

$$P(case) = \frac{1}{1+e^{-(\beta_0+\beta_1^{SNP1}+\beta_2^{SNP2})}} \qquad (1)$$

And

$$P(case) = \frac{1}{1+e^{-(\beta_0+\beta_1 SNP1+\beta_2 SNP2+\beta_3 SNP1 \cdot SNP2)}} \qquad (2)$$

Where P(case) is a binary indicator of disease status, SNP1 and SNP2 are categorical variables with values of 0, 1, or 2 corresponding to homozygous dominant, heterozygous, or homozygous recessive genotypes, and SNP1*SNP2={00, 01, 02, 10, 11, 12, 20, 21, 22} is the Cartesian product of SNP1 and SNP2.

Power estimation

For the purpose of our power calculations we define power not in the traditional statistical sense, but rather as an empirical measure of the number of successes (where success is defined as a p-value of less than 0.05 for the likelihood ratio test comparing the two models above) out of the total number of simulated replication data sets for each scenario.

Replication data set resampling

All replication data sets were resampled to match the genotype proportions of their corresponding discovery data sets by taking a random sample with replacement of the desired number of observations for each genotype, as follows. First, calculate desired genotype proportions from the crossproduct of discovery SNP1 proportions and discovery SNP2 proportions. Then, multiply desired genotype proportions by data set size to obtain the desired number of observations per genotype (if the discovery and replication data sets are the same size, there are simply equal numbers of individuals per SNP-SNP genotype combination, otherwise, they are proportionate). Next, ensure that there is at least one case and one control per genotype, and if not, add single pseudo-observations to ensure non-zero case and control sampling probabilities. Finally, for each genotype, take a random sample with replacement to the desired number of observations. The resampled replication data set is the composite of these samples by genotype. The following pseudocode outlines the resampling method.

```
INITIALIZE data frame to store resampled data set

FOR each SNP-SNP genotype
      SUBSET all observations of the genotype from the replication data set
      IF there are no case observations in this subset THEN
            APPEND a single case pseudo observation
      IF there are no control observations in this subset THEN
            APPEND a single control pseudo observation
      SAMPLE with replacement to size proportionate to discovery genotype
      APPEND sample to resampled data set
```

Negative simulation methods

Negative simulation data sets were generated such that the SNP-SNP interaction is significant in the underlying discovery population but not the replication population. We generated 1000 discovery data sets of size 4000 with minor allele frequencies of 0.5 and 0.5 for the interacting SNPs with penetrances of 0.5 for 8 of the 9 SNP-SNP genotypes and a penetrance of 0.9 for the SNP-SNP genotype where both SNPs have two doses of the minor allele. We next generated 15,000 replication data sets, 1000 each of the 15 two-SNP minor allele frequency combinations of {0.5, 0.4, 0.3, 0.2, 0.1}. All replication data sets were generated with a penetrance of 0.5 for all SNP-SNP genotypes, that is, none of the interacting SNP-SNP genotype combinations are significant. All power calculations for the discovery and replication data sets are estimated as described above in the “Power Estimation” section, and resampling of the replication data sets that were false positives was performed as described in “Replication Data set Resampling”.

Results

Positive simulation results

We found that performing resampling of the replication data sets generally resulted in better or comparable power to detect the interaction between SNP1 and SNP2 compared to the power to detect the interaction using the unadjusted replication data sets. Figure 2 illustrates a scenario in which the minor allele frequencies of both SNP1 and SNP2 are 0.5 in the discovery data set, and in the replication data sets the SNP1 MAF is held constant at 0.5 and only SNP2 is varied from 0.5 to 0.01 by 0.01 increments. There is increasing divergence of the unadjusted replication power to detect the interaction as the minor allele frequency of SNP2 decreases. Resampling results in better power to detect the interaction, with the worst performance observed for scenarios with the greatest difference between discovery and replication SNP2 minor allele frequency.

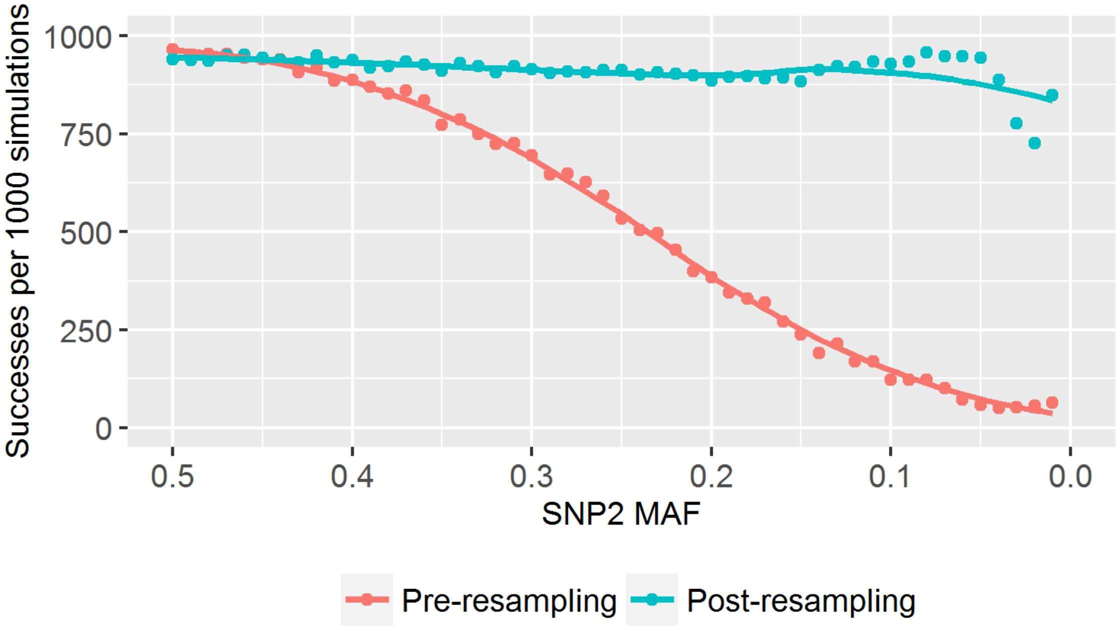


Figure 4. Detection of a SNP-SNP interaction in unadjusted versus resampled replication data sets. Model 1: discovery SNP1 MAF = 0.5, SNP2 MAF = 0.5. Replication SNP1 MAF = 0.05, SNP2 MAF from 0.5 to 0.01 by 0.01. Heritability = 0.005 and prevalence = 0.5 for all discovery and replication data sets.

Likewise, Figure 3 illustrates a comparable trend when both replication SNP1 and SNP2 minor allele frequencies are varied, with increasingly poor power to detect the interaction for unadjusted replication data sets with SNP1 and SNP2 minor allele frequencies that are more distant from those of the discovery data set. Once again, we estimate replication powers to detect the interaction that are much improved after performing resampling. Table 2 tabulates the pre- and post-resampling powers for the interaction for the remainder of our test models.

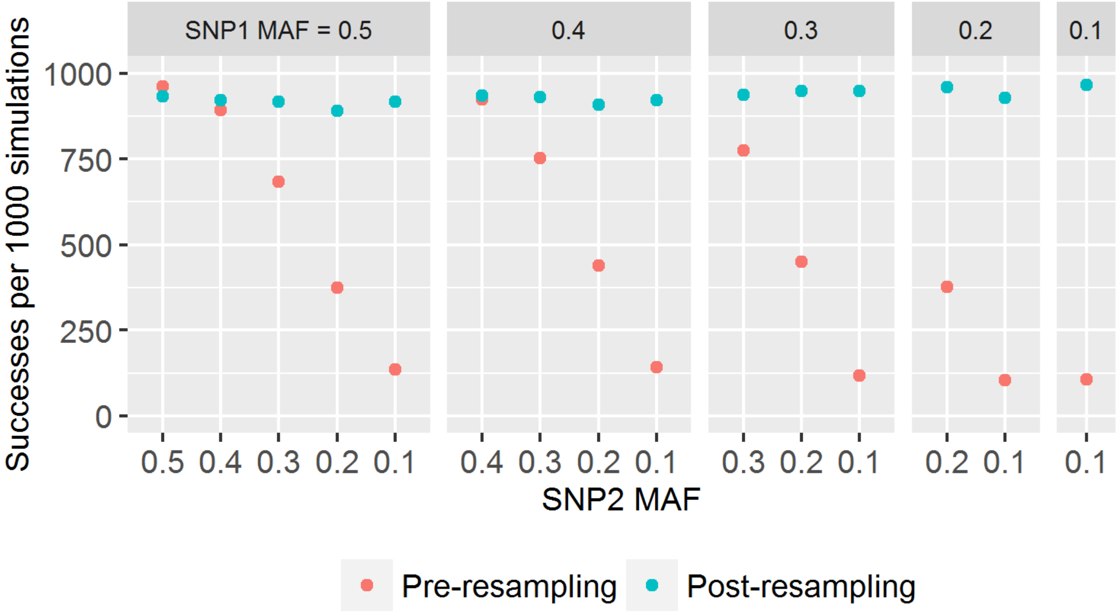


Figure 5. Detection of a SNP-SNP interaction in unadjusted versus resampled replication data sets. Model 1: discovery SNP1 MAF = 0.5, SNP2 MAF = 0.5. Replication SNP1 MAF and SNP2 MAF combinations from 0.5 to 0.1 by 0.1. Heritability = 0.005 and prevalence = 0.5 for all discovery and replication data sets.

Negative simulation results

We developed a negative simulation in order to both establish that our simulated data sets have a realistic false positive rate, and to investigate the factors that contribute to why data sets may fail to recapitulate the true underlying population significance of an interaction following application of our resampling method. In the vast majority of cases, if a SNP-SNP interaction is significant in a discovery population and not significant in a replication population, we will indeed observe respectively significant and non-significant p-values for the likelihood ratio test comparing the logistic regression models with and without the interaction in samples taken from these populations. However, samples do not always provide good representations of the underlying population. For a p-value cutoff of 0.05, we expect a 5% false positive rate, and indeed, 747 of the 15,000 simulated replication data sets (4.98%) yielded false positives (e.g. the data set yielded a significant p-value for the likelihood ratio test, even though the interaction was truly non-significant). We also expect the majority of these false positives to be unsuitable candidates for application of our resampling method because erroneous significance is driven by genotypes with so few observations that the sample penetrance is not representative of that of the underlying population, which is what we observe - 694 of the 747 false positives (92.9%) remain false positives after resampling.

Table 5. Pre- and post-resampling power (successes per 1000 simulated data sets) summary for all SNP-SNP minor allele frequency combinations for all models. See Table 1 for data set parameters by model number.

		Model # (Pre-resampling power, Post-resampling power)							
SNP1 MAF	**SNP2 MAF**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
0.5	**0.5**	962	1000	969	1000	1000	1000	938	1000
		932	1000	944	1000	1000	1000	1000	1000
	0.4	893	1000	989	1000	1000	1000	934	1000
		921	1000	967	1000	1000	1000	1000	1000
	0.3	684	1000	984	1000	1000	1000	951	1000
		916	1000	966	1000	1000	999	1000	1000
	0.2	375	1000	988	1000	1000	1000	946	1000
		890	1000	977	1000	1000	1000	1000	1000
	0.1	136	1000	857	999	1000	1000	944	1000
		917	1000	958	1000	1000	1000	1000	1000
0.4	**0.4**	923	1000	993	1000	1000	1000	941	1000
		935	1000	971	1000	1000	1000	1000	1000
	0.3	753	1000	994	1000	1000	1000	950	1000
		929	1000	965	1000	1000	999	1000	1000
	0.2	439	1000	976	1000	1000	1000	953	1000
		908	1000	984	1000	1000	1000	1000	1000
	0.1	141	998	850	999	1000	1000	952	1000
		922	1000	973	1000	1000	1000	999	1000
0.3	**0.3**	774	1000	991	1000	1000	1000	942	1000
		937	1000	975	1000	1000	1000	1000	1000
	0.2	451	1000	978	1000	1000	1000	967	1000
		947	1000	971	1000	1000	1000	1000	1000
	0.1	117	999	807	967	1000	1000	947	1000
		948	1000	991	1000	1000	1000	999	1000
0.2	**0.2**	377	1000	958	971	1000	1000	941	999
		959	1000	988	1000	1000	1000	1000	1000
	0.1	105	989	729	746	998	1000	950	1000
		928	1000	997	1000	1000	1000	999	1000
0.1	**0.1**	106	903	520	315	964	1000	934	999
		965	1000	998	1000	1000	1000	920	1000

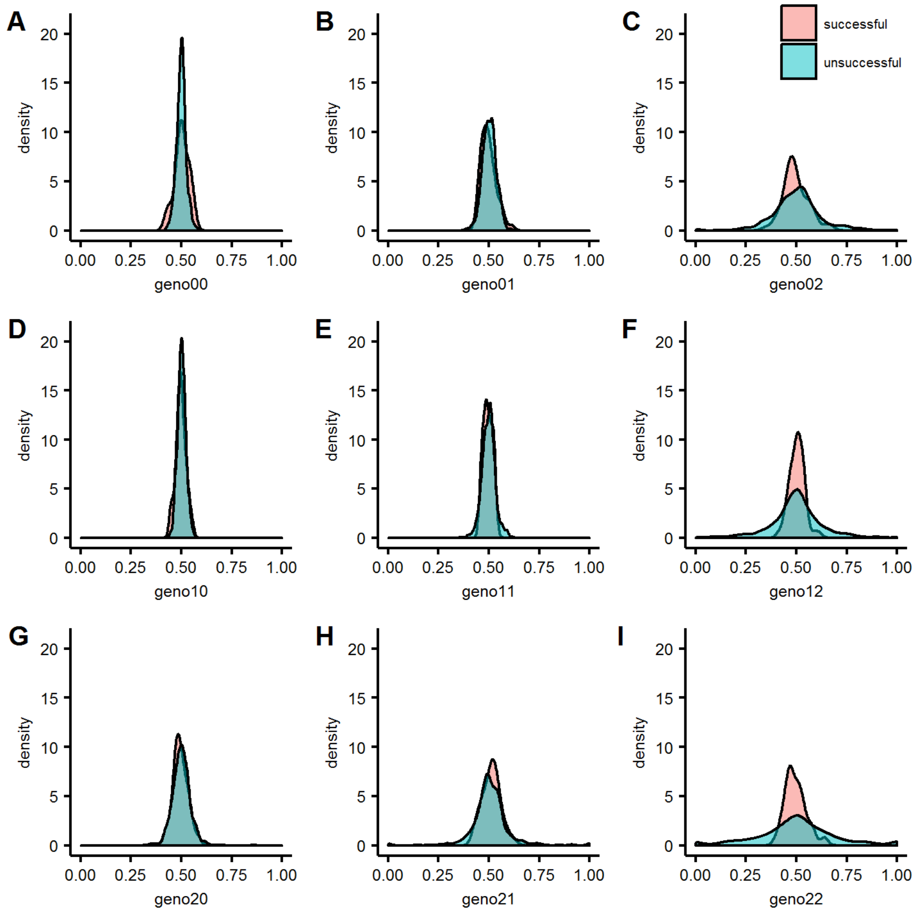


Figure 6. Distributions of penetrances by SNP-SNP genotype. Note that 48 of the data sets that did not successfully resample had zero observations for the 22 genotype (Part I), so the density reflects only those data sets for which we can calculate penetrances.

Comparing the penetrances and number of observations by SNP-SNP genotype provides insight into the factors that render our resampling method inappropriate. In Figure 4 one may observe that, while the mean penetrances for each genotype are comparable and centered around the expected value of 0.5 for those that do and do not successfully recapitulate the initial significance of the interaction following resampling, the variances differ significantly, particularly for the rarer genotypes. Table 3 provides a summary of the penetrance distributions and the number of observations by genotype. Recall that the minor allele frequencies of the two SNPs in the replication data sets are both 0.1, so the SNP-SNP genotype where both SNPs have two doses of the minor allele is quite rare, and in some cases, replication data sets completely lacked observations for this genotype. Looking at this genotype in particular (Figure 4, part I), the “unsuccessful” density is quite broad and flat, which stands to reason - since there are generally so few observations of this genotype, a population sample can be quite unrepresentative of the underlying population. Small perturbations in the observed number of cases per genotype can greatly skew the perceived significance of an interaction when the number of observations per genotype is quite low. In order to make reasonable inferences on the significance of an interaction, we recommend ensuring an adequate number of observations for the lowest frequency genotypes. This resampling method is only applicable if the underlying penetrances by genotype are reasonable approximations of the true population penetrances; our method simply provides a power boost for situations where we don't have enough observations of certain genotypes relative to others, but the observations we do have must be representative draws from the underlying population.

Table 3. Penetrance distribution summary statistics and number of observations, by SNP-SNP genotype, for data sets that successfully versus unsuccessfully resample

	Successfully resampled (n=53)		**Unsuccessfully resampled** (n=694)	
SNP-SNP genotype	**Mean penetrance** (SD)	**Median observations** (min, max)	**Mean penetrance** (SD)	**Median observations** (min, max)
00	0.502 (0.036)	489 (232, 1244)	0.499 (0.023)	904 (223, 2679)
01	0.498 (0.038)	489 (179, 817)	0.502 (0.034)	478 (154, 917)
02	0.496 (0.063)	162 (9, 269)	0.500 (0.122)	78 (5, 284)
10	0.497 (0.025)	733 (460, 1610)	0.502 (0.020)	996 (454, 1684)
11	0.495 (0.023)	920 (348, 1022)	0.498 (0.033)	618 (115, 1047)
12	0.503 (0.037)	290 (20, 534)	0.498 (0.123)	77 (4, 547)
20	0.495 (0.036)	330 (180, 815)	0.500 (0.046)	309 (21, 850)
21	0.512 (0.045)	42 (106, 537)	0.505 (0.095)	189 (2, 548)
22	0.496 (0.053)	108 (9, 263)	0.501 (0.194)	25 (0, 281)

Discussion

This study critiques the validity of replication as the gold standard for substantiating GWAS hits, and proposes the exploration of alternative approaches that consider how differences in factors such as minor allele frequency can modulate the observed significance of SNP-SNP interactions. This may expand the usefulness of data that is already collected, which can in turn better direct our resources to future studies that will be most fruitful.

For the purposes of our positive simulations, we tested an exploratory range of heritabilities, prevalences, and sample sizes, but did not explore a wide range of differences in these parameters between populations. Small differences in minor allele frequency between discovery and replication data sets can greatly reduce the power to detect main effects, and the ability to detect this may differ by heritability, so it seems plausible that the ability to detect interactions may follow similar and possibly even more pronounced trends [54]. Future investigation into the potential joint effects of these parameters may yield further insight into the factors that affect our ability to detect and assess interactions in diverse populations, and subsequently direct study design to better control for these differences. Similarly, more negative simulations should be performed that systematically cover a range of scenarios to illuminate the conditions under which we can reliably regain power to detect interactions in replication studies; our negative simulations do not establish guidelines for across the entire space of observable population parameters.

Additional future analyses may also aim to demonstrate that shifts in interaction significance following resampling can alter which ones are selected for model inclusion, thereby modulating our ability to predict case-control status. Investigating shifts in variable inclusion following resampling is likely to yield interesting biological insight. Indeed, the future of extracting meaningful findings from GWAS is likely to be driven by investigating SNPs that are initially identified either based on expert knowledge or via bioinformatics methods incorporating prior assumptions, including hierarchical models that consider groups of SNPs and their functional relationships and interactions, in order to bypass the extreme prejudice of multiple testing burden

[27, 97, 141]. Furthermore, the genome-wide significance level is unlikely to be an appropriate one-size-fits-all cutoff for multiple reasons, including evidence for the successful replication of borderline statistically significant genotype-phenotype associations [104]. It also stands to reason that diverse populations with different genetic architectures may require variable significance cutoffs that better reflect their patterns of linkage disequilibrium; hopefully, the increasing quantity of genotyped diverse populations will result in more pertinent high-quality reference genomes that will better enable identifying and replicating genetic associations between populations, enabling more accurate comparison of populations in the context of structural differences between diverse genomes [36, 115].

As new computational methods for GWAS present solutions to the various challenges associated with the accumulation of vast amounts of ever denser genetic data, and is mutually reinforced with increasing integration with clinical and epidemiological data, it is important to keep sight of the end goal of practical application of this knowledge to the betterment of both population health and personalized medicine. As such, the ultimate takeaway from this study should be that the purpose of resampling is to identify potential candidates for further biological validation, with the intention of using these findings to reduce structural inequalities in health and medicine.

IMPROVING MACHINE LEARNING REPRODUCIBILITY IN GENETIC ASSOCIATION STUDIES WITH PROPORTIONAL INSTANCE CROSS VALIDATION (PICV)

Background

Genome-wide association studies (GWAS) have been frequently critiqued for failing to explain the “missing heritability” of complex disease in terms of single-locus main effects [89, 137]. In addition to interrogating the contributions of rare variants, non-coding regions, structural variation, etc., a logical reactionary paradigm to embrace involves revisiting heritability estimates to consider the effect of interactions and developing approaches that acknowledge that loci do not exist in isolation but rather act in complex networks of interacting partners in the dynamic, three-dimensional genome and in tissue-specific and environmental context [39, 40, 79, 148]. Utilizing pre-existing GWAS data to test a curated set of potentially biologically-relevant interactions, such as those identified as being plausible via expert knowledge, integrating data from gene set enrichment analyses, chromatin capture experiments, co-expression data sets, etc. provides a way to overcome the multiple testing burden of naively testing every possible interaction and motivates future bench science experimentation [26, 114]. Accordingly, machine learning methods are appealing for the analysis of this big, complex data, and have been applied to diverse problems and data types across the biological sciences [76, 78]. However, machine learning should not be viewed as a panacea that can be readily applied to all genomics problems. Beyond concerns regarding model choice and interpretability, there are numerous reasons why valid biological interactions may fail to appear statistically significant and vice versa [54, 97-8]. Therefore, typical machine learning tools, techniques, and standards from other fields may need tweaking to be appropriate for use in genomics considering the unique biases in generating genomic data sets, the structure of the genome, the validity of model assumptions, etc.

Improving the reproducibility of machine learning analyses of genomic data will require methodological and analytic advances from not only both the computational and wet laboratory sides, but also their consideration in conjunction with each other as a greater whole. Sharing data publicly for secondary analyses, writing open-source code in executable notebook format, and using container and cloud services all contribute to a culture of reproducibility that enhances the capacity for integrative and innovative computational analyses [10, 72, 85, 93]. Likewise, thoughtfully interrogating methodological, environmental, and other determinants of inconsistencies in bench experimentation results lends robustness to findings, and this greater understanding of sources of variation can in itself lead to worthwhile new hypotheses [62]. Ideally, technological supports such as mobile applications for data collection will increasingly allow for recording more complete and consistent data in a format that can be seamlessly analyzed with software tools developed or modified to consider the unique intricacies of the data at hand [110].

Epistasis, or the non-additive interaction between genotypes to produce phenotype, is difficult to detect statistically but is of biological interest in light of a multifactorial view of disease [29, 90, 96]. This study is motivated by poor cross-validation performance observed for epistasis data sets with an interaction between two single nucleotide polymorphisms (SNPs). Cross validation is a widely-used standard for evaluating the performance of a machine learning analysis in which the data is split into training and testing partitions, a model is fit using the training set, and its performance is evaluated on predicting the classes of the held out test set observations [3]. Typically the overall data set is split such that the resultant training and testing partitions are random, independent draws from the same probability distribution, although there are also methods that consider the data structure, generally in terms of maintaining outcome class proportions between the training and testing data sets [49, 59, 132]. In this study, we propose a new cross validation method, proportional instance cross validation (PICV), that preserves the relative distribution of an independent variable (in our example application, SNP-SNP interaction genotypes) when dividing the overall data set into train and test partitions. We demonstrate significantly improved sensitivity and positive predictive value across all tested scenarios with application of PICV relative to a traditional cross validation implementation. We additionally apply PICV to primary open-angle glaucoma GWAS data to investigate an interaction previously reported to be significant in two independent data sets. Although this interaction is not observed to be significant in our analysis, PICV produced more consistent estimates than a traditional cross validation implementation. This approach is not only appropriate for epistasis data but may be readily applied to comparable imbalanced variable problems.

Methods

Data set generation

All data sets were generated using GAMETES, a tool that produces epistatic models between SNPs and creates data sets based off these models [131]. Penetrance functions were generated for SNP-SNP interaction scenarios for all 15 combinations of minor allele frequencies (MAFs) of {0.5, 0.4, 0.3, 0.2, and 0.1}, with SNP heritability kept constant at 0.005 and population prevalences of 0.5, 0.1, and 0.02 (Table 1, Supplemental tables 1-2). Although a prevalence of 0.5 may seem high for a given disease, numerous risk factors for chronic and complex diseases in the United States population that may be phenotypes of interest are as or more prevalent, including being overweight or obese, lack of physical activity, excessive sodium consumption, lack of fruit and vegetable consumption, etc [30]. The simulated data with prevalence of 0.1 is intended to reflect the US prevalence of common complex diseases such as diabetes or cardiovascular disease [100]. The simulated data sets of 0.02 prevalence approximately reflect the US prevalence of primary open-angle glaucoma, which is investigated in the real data application [51]. Balanced case-control ratio data sets of size 2,000 and 10,000 were generated for the 0.5 prevalence scenario and of size 10,000 for the 0.1 and 0.02 prevalence scenarios.

	Scenario 1			Scenario 2			Scenario 3		
SNP1 MAF:	0.1			0.2			0.2		
SNP2 MAF:	0.1			0.1			0.2		
Penetrance:	0.493	0.531	0.522	0.507	0.480	0.556	0.514	0.481	0.425
	0.526	0.387	0.410	0.471	0.590	0.249	0.467	0.544	0.674
	0.611	0.008	0.358	0.485	0.532	0.482	0.539	0.447	0.304
	Scenario 4			**Scenario 5**			**Scenario 6**		
SNP1 MAF:	0.3			0.3			0.3		
SNP2 MAF:	0.1			0.2			0.3		
Penetrance:	0.513	0.494	0.456	0.488	0.525	0.450	0.481	0.533	0.446
	0.438	0.530	0.696	0.527	0.455	0.562	0.525	0.468	0.513
	0.520	0.475	0.506	0.478	0.458	0.814	0.483	0.470	0.734
	Scenario 7			**Scenario 8**			**Scenario 9**		
SNP1 MAF:	0.4			0.4			0.4		
SNP2 MAF:	0.1			0.2			0.3		
Penetrance:	0.484	0.501	0.535	0.490	0.523	0.455	0.502	0.523	0.425
	0.570	0.494	0.359	0.512	0.468	0.568	0.499	0.472	0.588
	0.545	0.551	0.245	0.565	0.395	0.668	0.495	0.503	0.501
	Scenario 10			**Scenario 11**			**Scenario 12**		
SNP1 MAF:	0.4			0.5			0.5		
SNP2 MAF:	0.4			0.1			0.2		
Penetrance:	0.476	0.535	0.449	0.306	0.333	0.341	0.476	0.521	0.482
	0.506	0.473	0.568	0.428	0.314	0.256	0.521	0.472	0.536
	0.536	0.503	0.410	0.322	0.198	0.595	0.715	0.392	0.502
	Scenario 13			**Scenario 14**			**Scenario 15**		
SNP1 MAF:	0.5			0.5			0.5		
SNP2 MAF:	0.3			0.4			0.5		
Penetrance:	0.500	0.520	0.459	0.422	0.515	0.547	0.440	0.560	0.440
	0.477	0.480	0.563	0.548	0.491	0.470	0.522	0.484	0.509
	0.608	0.482	0.429	0.531	0.492	0.485	0.515	0.472	0.542

Table 6. Data set simulation parameters. Minor allele frequencies and penetrance tables used to generate balanced case-control ratio data sets of size 2,000 and 10,000. Heritability = 0.005 and prevalence = 0.5 constant across all simulations.

Implementation and evaluation of traditional cross validation

For each of the 15 scenarios for each investigated prevalence and sample size combination, we perform 1000 replicates of a standard cross validation procedure in which two-thirds of observations are randomly allocated to be used for training and the remaining third is used for testing. The training data is then used to fit the following logistic regression models with and without the SNP-SNP interaction:

$$P(case) = \frac{1}{1+e^{-(\beta_0+\beta_1 SNP1+\beta_2 SNP2)}} \qquad (1)$$

$$P(case) = \frac{1}{1+e^{-(\beta_0+\beta_1 SNP1+\beta_2 SNP2+\beta_3 SNP1*SNP2)}} \qquad (2)$$

Where P(case) is a binary indicator of case-control status, SNP1 and SNP2 are categorical variables in which 0 corresponds to the homozygous dominant genotype, 1 to the heterozygous, and 2 to the homozygous recessive, and SNP1*SNP2 corresponds to the Cartesian product of the two {00, 01, 02, 10, 11, 12, 20, 21, 22}.

These models fit to the training data are then used to predict case-control status for the held-out testing data, using a cutoff of 0.5 for case versus control prediction assignment from the fitted values. These predictions are then used to calculate the sensitivity, specificity, positive predictive value, and negative predictive value for the testing data.

Implementation and evaluation of proportional instance cross validation (PICV)

For the proportional instance cross validation procedure, rather than randomly allocating each observation to be included in the training or testing set, observations are allocated in a genotype-specific fashion (Figure 1). Two-thirds of the observations of each SNP-SNP genotype are randomly allocated to be used for training and the remaining third is used for testing. Therefore the same total proportion of individuals used for training versus testing is maintained as in the traditional cross validation procedure, and additionally, the relative proportions of each genotype are preserved between the overall data set and the training and testing partitions. Model fitting with the training data, testing data predictions, and performance measure calculations are conducted as for the traditional cross validation.

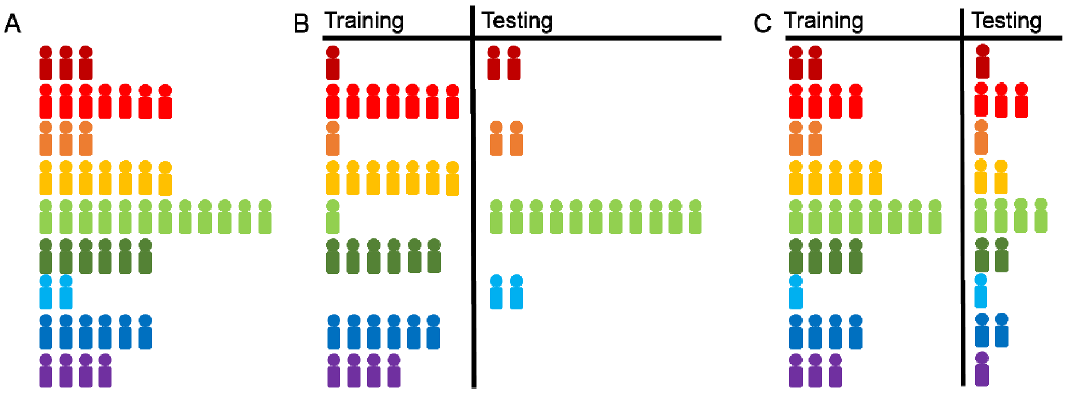


Figure 1. Comparing traditional cross validation and proportional instance cross validation (PICV). A) The overall distribution of 9 SNP-SNP interaction genotypes in a population of individuals. B) Traditional cross validation in which 2/3 of observations are randomly allocated to the training set and the remaining 1/3 are allocated to the testing set can result in draws with imbalanced genotype proportions. C) PICV randomly allocates 2/3 of observations of each genotype to the training set and the remaining 1/3 to the testing set, ensuring that the relative proportions of genotypes are maintained.

Comparison of traditional cross validation and proportional instance cross validation (PICV)

For both traditional cross validation and PICV, we calculate the absolute value of the difference between training and testing for each of four performance measures (sensitivity, specificity, positive predictive value, and negative predictive value) over 1,000 trials for each of the 15 scenarios. We calculate p-values for the two-sample Kolmogorov-Smirnov test with the null hypothesis that there is no difference between the traditional cross validation implementation and PICV distributions of the difference between training and testing for each performance measure, with the one-sided alternative that the PICV distribution is smaller, with a significance threshold of $\alpha = 0.05$.

Results

Implementing PICV for our simulated epistasis examples (that is, performing cross validation data set splitting such that observations are allocated to maintain the same relative proportions of each SNP-SNP genotype in the training and testing sets as in the data set overall) significantly improved the consistency between training and testing sensitivities and positive predictive values. Figure 2 illustrates comparisons of training/testing consistencies for PICV versus a traditional cross validation procedure in which observations are allocated to the training and testing sets without regard to genotype (see Supplemental Figures 1-60 for all minor allele frequency, prevalence, and cohort size combinations). P-values listed are for the two-sample Kolmogorov-Smirnov test of the distributions of the absolute values of the differences between the training and testing performance measure (e.g. sensitivity) over 1,000 trials per scenario for these two cross validation approaches, with a one-sided alternative hypothesis that the split-by-genotype distribution is smaller. Table 2 summarizes these performance measures across all 15 SNP-SNP genotype MAF combination scenarios for the 0.5 prevalence simulations of size 2,000 (see Supplemental Table 3 for prevalence = 0.5 and n = 10,000, Supplemental Table 4 for prevalence = 0.1, Supplemental Table 5 for prevalence = 0.02). Sensitivity and positive predictive value were significantly more consistent between test and train for PICV than for traditional cross validation across all 15 scenarios tested for both n=2,000 and n=10,000. Although the specificity and negative predictive value comparisons mostly did not meet statistical significance, smaller medians and maximum values for the differences in these performance measures between training and testing were observed for the PICV approach for the majority of scenarios (Table 3).

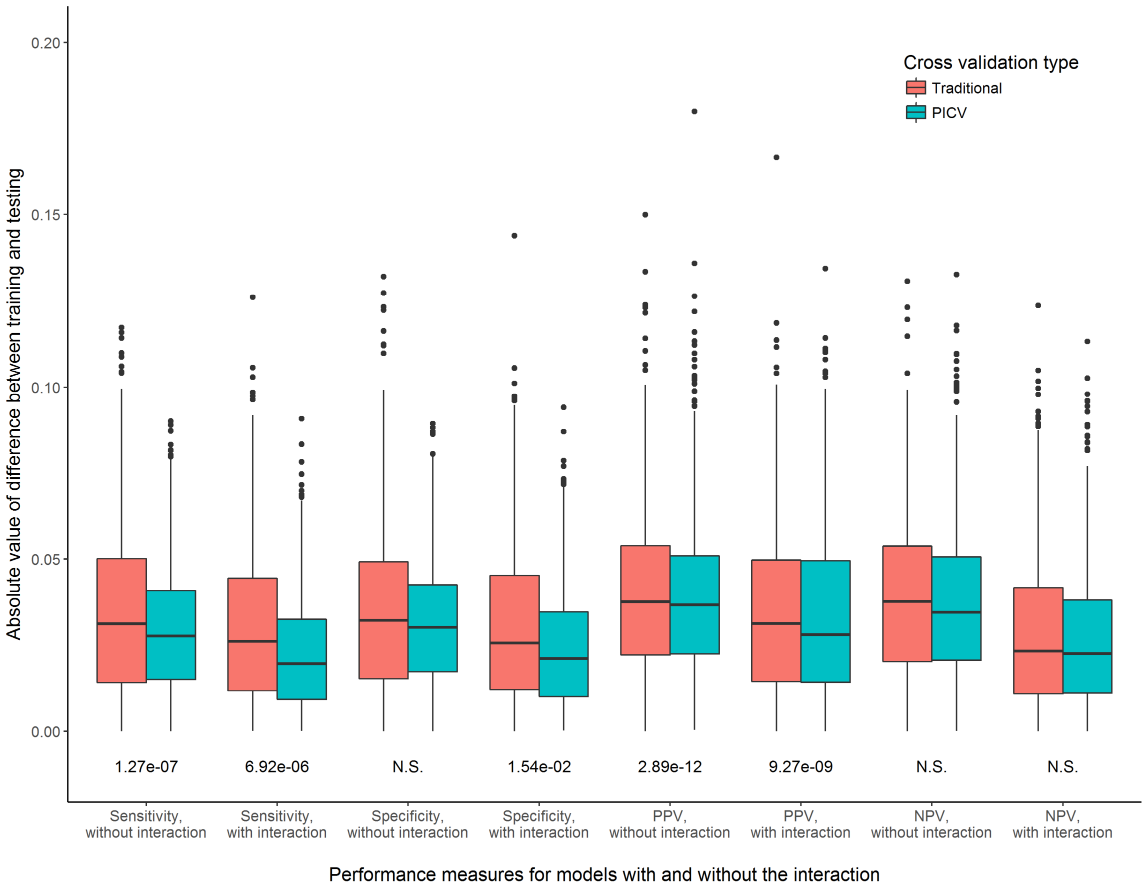


Figure 2. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario in which both SNPs have a MAF of 0.5, prevalence = 0.5, n=2,000.

Table 2. Summary of performance measures across minor allele frequency combinations, prevalence = 0.5, n = 2,000.

Measure, Model Scenario	Sens, without int	Sens, with int	Spec, without int	Spec, with int	PPV, without int	PPV, with int	NPV, without int	NPV, with int
SNP1 MAF: 0.1 SNP2 MAF: 0.1	3.06e-17	9.89e-08	N.S.	N.S.	1.90e-18	7.67e-08	N.S.	N.S.
SNP1 MAF: 0.2 SNP2 MAF: 0.1	7.04e-20	4.54e-05	3.88e-02	N.S.	3.68e-11	5.56e-06	4.35e-02	1.89e-02
SNP1 MAF: 0.2 SNP2 MAF: 0.2	1.69e-10	1.69e-10	N.S.	6.87e-03	4.06e-09	4.06e-09	N.S.	N.S.
SNP1 MAF: 0.3 SNP2 MAF: 0.1	1.59e-08	2.47e-05	4.35e-02	N.S.	9.27e-09	2.47e-05	3.46e-02	N.S.
SNP1 MAF: 0.3 SNP2 MAF: 0.2	6.14e-04	5.02e-11	N.S.	N.S.	3.07e-16	1.22e-14	N.S.	N.S.
SNP1 MAF: 0.3 SNP2 MAF: 0.3	5.16e-04	4.33e-04	N.S.	N.S.	1.75e-04	1.75e-04	N.S.	N.S.
SNP1 MAF: 0.4 SNP2 MAF: 0.1	9.94e-05	7.67e-08	N.S.	N.S.	3.52e-08	5.53e-10	N.S.	N.S.
SNP1 MAF: 0.4 SNP2 MAF: 0.2	6.65e-17	1.45e-04	N.S.	N.S.	5.36e-09	2.42e-02	N.S.	N.S.
SNP1 MAF: 0.4 SNP2 MAF: 0.3	2.71e-08	4.54e-05	N.S.	N.S.	8.97e-07	4.46e-06	N.S.	N.S.
SNP1 MAF: 0.4 SNP2 MAF: 0.4	1.63e-05	1.41e-03	N.S.	N.S.	2.66e-03	8.62e-04	N.S.	N.S.
SNP1 MAF: 0.5 SNP2 MAF: 0.1	8.97e-07	7.06e-09	N.S.	N.S.	2.27e-06	1.27e-07	4.85e-03	N.S.
SNP1 MAF: 0.5 SNP2 MAF: 0.2	9.42e-18	6.75e-05	1.28e-02	N.S.	4.00e-12	8.60e-06	N.S.	N.S.
SNP1 MAF: 0.5 SNP2 MAF: 0.3	4.38e-07	2.47e-05	N.S.	N.S.	7.67e-08	4.12e-10	N.S.	1.46e-02
SNP1 MAF: 0.5 SNP2 MAF: 0.4	2.69e-07	5.54e-05	N.S.	N.S.	7.06e-09	8.62e-04	N.S.	N.S.
SNP1 MAF: 0.5 SNP2 MAF: 0.5	1.27e-07	6.92e-06	N.S.	1.54e-02	2.89e-12	9.27e-09	N.S.	N.S.

Table 3. Number of scenarios for which PICV yielded smaller median, maximum differences between training and testing, n = 10,000

	PICV median less than traditional CV median (out of 15)			**PICV maximum less than traditional CV maximum (out of 15)**		
	Prevalence			Prevalence		
Measure, Model	0.02	0.1	0.5	0.02	0.1	0.5
Specificity, without interaction	15	15	12	15	15	15
Specificity, with interaction	15	15	15	15	15	15
NPV, without interaction	14	9	9	11	8	8
NPV, with interaction	8	9	10	8	9	9

Primary open-angle glaucoma interaction analysis

Prior interaction analyses of primary open-angle glaucoma identified several pairs of replicating interactions using the eMERGE and NEIGHBOR data [136]. We attempted to replicate the most significant interaction (between ALX4 and RBFOX1) in the GLAUGEN data set (dbGaP Study Accession: phs000308.v1.p1, available at https://www.ncbi.nlm.nih.gov/gap), which is harmonized with NEIGHBOR. The GLAUGEN model is adjusted for age, sex, site, and the first 6 principal components to reflect the eMERGE and NEIGHBOR models (the eMERGE and NEIGHBOR models additionally adjusted for platform, but all GLAUGEN samples were assessed on the same platform). Our analysis did not find a significant interaction between the two variants (Table 4). However, application of PICV to this data did yield training and testing p-values (0.376 and 0.323, respectively) more consistent with the overall LRT p-value (0.327) than a traditional cross validation procedure (0.442 and 0.470, respectively).

Table 4. Comparison of interaction significance across data sets

Data set	ALX4 variant	RBFOX1 variant	LRT p-value
eMERGE	rs10838251	rs653127	7.29E-06
NEIGHBOR	rs7126447	rs11077011	1.62E-06
GLAUGEN	rs7126447	rs11077011	0.327

Discussion

Implementing a cross validation splitting procedure that maintains the relative proportions of each SNP-SNP genotype when dividing the overall data set significantly improved the sensitivity and positive predictive value consistencies between the training and testing partitions in each of the experimental scenarios tested. Although specificity and negative predictive value improvement did not meet statistical significance in most cases, application of the PICV approach did yield smaller median and maximum absolute differences between training and testing in the majority of scenarios. The interaction analysis did not replicate the prior finding between ALX4 and RBFOX1, however PICV still produced more consistent estimates than a traditional cross validation procedure for this data. Verma et al note that RBFOX1 has been previously shown to be associated with myopia, and that eMERGE primary open-angle glaucoma cases had not been screened for myopia; GLAUGEN excluded individuals with more than 8 diopters of myopia. This inconsistent finding highlights the importance of considering epidemiological confounders and co-morbidities of complex phenotypes in genetic analyses.

Class imbalance is a well-recognized issue in machine learning analyses, particularly for the analysis of high-dimensional data sets as in genomics and other biomedical applications [80]. If the main objective of a machine learning analysis is maximizing accuracy, and the minority class is very small, simply predicting the majority class for each observation may yield high overall accuracy, as in the spam filtering problem [56]. Clearly, adoption of a balanced accuracy measure or a cost-sensitivity analysis that weighs the relative importance of avoiding false positives versus false negatives is critical for such problems, and numerous methods have been developed to address this issue including novel fitness functions, sampling-based approaches, and ensemble methods, including for epistasis modeling [52, 83, 86, 134]. The present study, though thematically similar to the class imbalance problem, instead addresses imbalance in observations of classes of an independent variable, e.g. the SNP-SNP interaction genotype. This is also adjacent to the covariate and data set shift problems, in which the training and testing distributions differ (for example due to model training using clean data from consistent laboratory conditions to produce models that then fail to hold for experimentally gathered data with unanticipated environmental differences), but for internal cross validation [1, 123, 126]. Solutions to problems of both of these genres include re-weighting and –sampling techniques, whereas the present study circumvents the need for either via splitting the data to ensure balanced proportions by genotype between training and testing sets. The example application of imbalanced SNP-SNP genotypes considers a categorical variable, but the underlying idea of preserving the distribution of instances between training and testing with regard to an independent variable could be extended to continuous variables or combinations of variables via binning, propensity scores, etc.

Conclusions

Although the contribution of epistatic interactions may help explain the “missing heritability” of complex disease, statistical detection of epistasis remains challenging and can require adjustment of general machine learning protocols. With decreasing minor allele frequencies, the number of observations for rare SNP-SNP interaction genotypes becomes quite small in a GWAS of typical size, and a standard cross validation procedure may result in training/testing data set splits that poorly represent the data as a whole. This diminishes the ability to identify interactions of potential interest for experimental follow-up, and underscores the need to perform interaction analyses in an interaction-specific framework. A potentially overlooked element of performing reproducible analyses includes the imperative to develop and modify methods considering how intrinsic characteristics of the data and its structure may contribute to statistical failure to replicate despite biological (or other scientific) validity. Genomics and the biomedical sciences in general benefit from their increasingly multidisciplinary nature by incorporating methodology and theory from adjacent computational fields, but thoughtful contextualization of the data in view of the underlying biology is necessary to reap the potential benefits of applied machine learning methods and to successfully reproduce them.

IDENTIFICATION OF EPISTATIC INTERACTIONS BETWEEN THE HUMAN RNA DEMETHYLASES FTO AND ALKBH5 WITH GENE SET ENRICHMENT ANALYSIS INFORMED BY DIFFERENTIAL METHYLATION

Introduction

The Genetic Analysis Workshop (GAW) is a forum for investigators to develop and critique new analytical methods for complex traits on a shared data set. The GAW20 data provide simulated replications based on the Genetics of Lipid-lowering Drugs and Diet Network (GOLDN) clinical trial, had participants been subject to treatment with a fictitious drug with a pharmaco-epigenetic effect on triglyceride response [66]. These data present a unique analysis opportunity as all phenotypes, subject covariates, genotypes, pre-treatment methylation levels, etc. are real data from the trial, but are accompanied by simulated post-treatment methylation and triglyceride levels. GAW participants choose to analyze the data with or without knowing the simulation methods; the simulated data analysis was performed prior to attending GAW, without knowledge of the simulation methods. Analysis of the real data was performed following GAW attendance, as it was revealed that the data simulation methods did not consider interactions, and therefore analysis of interactions in the simulated data was not appropriate.

Despite evidence for multi-locus underpinnings of phenotype-genotype association, the multiple-testing burden associated with fitting interaction models is stringent due to the high dimensionality of genomic data [90]. Strategies for better detecting these interactions can aim to avoid exhaustively testing each potential interaction via data reduction methods, integrating expert knowledge, and/or consolidating multiple sources of evidence to narrow the search space [26, 97, 113].

In this study, we hypothesize that CpG sites that are differentially methylated with respect to treatment are associated with the pharmaco-epigenetic mechanism of the fictitious drug. Considering the evidence for multi-locus models of complex disease etiology, we hypothesize that the drug response is better evaluated by gene set enrichment analyses than single locus models. By integrating results from Gene Ontology (GO), drug-disease association, and microRNA (miRNA) target analyses we find evidence implicating the relevance of adenosine and miRNAs with known epigenetic regulation and roles in lipid metabolism. From this, we infer the potential importance of the N^6-methyladenosine modification in the pharmaco-epigenetic response on triglycerides, and consider how miRNA adenosine methylation rather than CpG methylation may impact the phenotype. Lacking direct data for miRNA adenosine methylation, we perform a targeted epistasis search between loci on the two RNA demethylases FTO and ALKBH5, and find evidence for statistical epistasis between one variant within each respective gene. Repeating the analysis with the real data revealed four significant interactions between variants across these genes. Overall, we present an example workflow (Figure 1) in which integration of multiple sources of information can help uncover biological meaning in the absence of significant main effects.

CpG site filtering	Paired t-tests for pre- and post-treatment methylation levels	α = 0.05 463,995 hypotheses Bonferroni cutoff 1.08 x 10^{-7} 212,018 pass
Modeling relationship between phenotype and CpG site methylation	Linear models: log ratio of post-treatment to pre-treatment TG level ~ log ratio of post-treatment to pre-treatment methylation	α = 0.05 212,018 hypotheses Bonferroni cutoff 2.36 x 10^{-6} NONE pass 4433 gene list constructed from all CpG sites with p-values ≤ 0.05
Gene set enrichment analyses	WebGestalt enrichment analyses	Gene ontology Drug association MicroRNA targets Top drug: Adenosine (8 x 10^{-21}) Top miRNA: miR-124a (1.70 x 10^{-42})
Targeted epistasis search	Likelihood ratio test comparing models with and without the interaction Only interactions of SNPs between the two genes FTO and ALKBH5, not within	α = 0.05 340 hypotheses (simulated) 255 hypotheses (real) Bonferroni cutoff 0.00015 (simulated) Bonferroni cutoff 0.000196 (real) One significant interaction (simulated) Four significant interactions (real)

Figure 1. Workflow overview and results summary

Data set and methods

Data set

The GOLDN data and companion simulations for GAW20 are previously described [66]. Relevant to this analysis, subject data includes fasting lipid profiles prior to and post-treatment, methylation at > 450,000 CpG sites prior to and post-treatment, GWAS of > 700,000 autosomal SNPs, and covariates including age, center, metabolic syndrome-related traits, and smoking status. This analysis is of the pre-defined single representative replicate (n=680) of the post-treatment methylation and triglyceride levels.

Phenotype definition

We define the phenotype of interest as the log ratio of the average post-treatment triglyceride level to the average pre-treatment triglyceride level. Due to the high correlations between triglyceride levels at pre-treatment time points 1 and 2 and post-treatment time points 3 and 4 (0.90 and 0.91, respectively), and presence of a value for at least one of time point 1 or 2 and 3 or 4 for each individual, we singly impute missing values via linear regression.

CpG site filtering

Significantly differentially methylated CpG sites are identified via paired t-tests for pre- and post-treatment methylation levels ($\alpha = 0.05$ for 463,995 hypotheses yields Bonferroni cutoff of 1.08×10^{-7}).

Modeling the relationship between phenotype and CpG site methylation

Linear models are fit to test the relationships between the phenotype and methylation status of the significant CpG sites identified above, characterized as a single predictor: the log ratio of post-treatment to pre-treatment methylation ($\alpha = 0.05$ for 212,018 hypotheses yields Bonferroni cutoff of 2.36×10^{-6}).

Gene set enrichment analyses

All CpG sites that pass the initial t-test filter and have a p-value ≤ 0.05 for the phenotype ~ methylation predictor model are used to curate a list of corresponding genes with evidence for both differential methylation and association with the phenotype. This gene list is used for Gene Ontology, drug association, and microRNA target enrichment analyses using the WEB-based GEne SeT AnaLysis Toolkit (WebGestalt, http://www.webgestalt.org/) [144].

Targeted epistasis search

We investigate potential epistasis between the two RNA demethylases FTO and ALKBH5 by calculating p-values for the likelihood ratio tests comparing the linear models containing each FTO SNP – ALKBH5 SNP pair, with and without their interaction term ($\alpha = 0.05$ for 340 hypotheses yields Bonferroni cutoff of 0.00015 for the simulated data, 255 hypotheses yields a cutoff of 0.000196 for the real data).

Results

CpG site filtering

We tested 463,995 CpG sites for differential methylation prior to versus post-treatment. 212,018 CpG sites passed the Bonferroni threshold of 1.08×10^{-7}.

Modeling the relationship between phenotype and CpG site methylation

None of the 212,018 CpG sites that are significantly differentially methylated reached genome-wide significance for association with the phenotype (Figure 2).

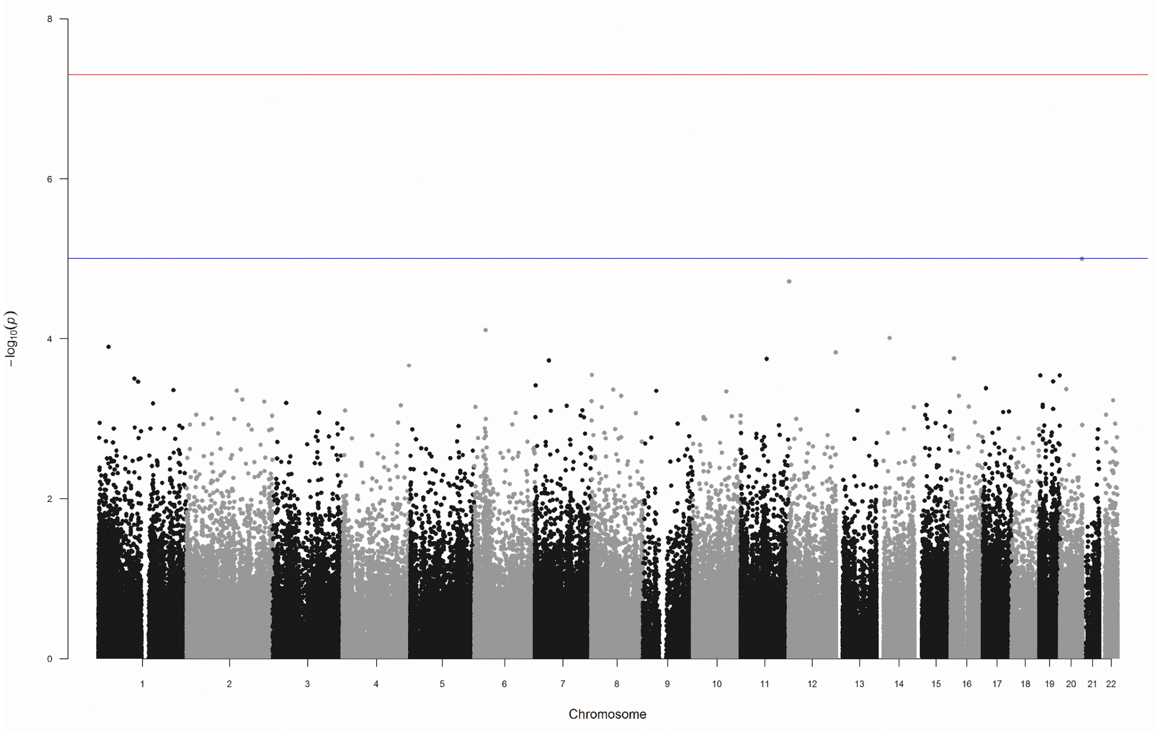


Figure 2. Manhattan plot of triglyceride phenotype ~ CpG site methylation log ratio

Gene set enrichment analyses

Our gene set is constructed from all CpG sites with p-values ≤ 0.05 for the models above for which corresponding gene annotations are available (5413 of the 212,018 differentially methylated sites). Some CpGs have more than one corresponding gene listed, and many genes have multiple CpGs, for a total gene list length of 4443. The top result from the drug association analysis is adenosine (number of reference genes in the category = 477, number of genes in the gene set and also in the category = 126, expected number in the category = 49.14, ratio of enrichment = 2.56, raw p value from hypergeometric test = 1.31 x 10^{-23}, p value adjusted by multiple test adjustment = 8 x 10^{-21}). The top result from the miRNA target analysis is miR-124a (number of reference genes in the category = 542, number of genes in the gene set and also in the category = 175, expected number in the category = 55.84, ratio of enrichment = 3.13, raw p value from hypergeometric test = 7.82 x 10^{-45}, p value adjusted by multiple test adjustment = 1.70 x 10^{-42}).

Targeted epistasis search

The GAW20 GWAS data with the simulated phenotype include complete observations for 68 SNPs on FTO and 5 SNPs on ALKBH5 for the 680 subjects. We only test for epistatic interactions between SNPs across the two genes (and do not test for interactions between SNPs on the same gene), for a total of 340 tested interactions and therefore a Bonferroni threshold of 0.00015. One pair of SNPs, rs2192872 from FTO and rs8068517 from ALKBH5 have a significant p-value for the likelihood ratio test comparing the models with and without the interaction (p = 2.01 x 10^{-5}).

The analysis of the real phenotype and GWAS data was performed in the same manner for 778 subjects for 51 SNPs on FTO and 5 SNPs on ALKBH5 (Bonferroni threshold of 0.000196). Four pairs of SNPs have significant p-values for the likelihood ratio test comparing the models with and without the interaction (Table 1). Figure 3 visually summarizes the distribution of phenotype by genotype for the two SNPs involved in the most significant identified interaction.

Table 1. Summary of significant interactions. Variant annotations are from Ensembl [142]. Base model covariate selection is based on significance at the 0.05 level and includes average pre-treatment triglyceride level, age, center, current smoker status, and sex.

SNP	Alleles	MAF	Location	Gene	Consequence Type	LRT p-value
rs1362571	G/T	0.34 (G)	16:53877858	FTO	Intron variant	2.76 x 10^{-6}
rs11655588	A/G	0.18 (G)	17:18204137	ALKBH5	Intron variant	
rs10521304	T/C	0.41 (C)	16:53874745	FTO	Intron variant	8.80 x 10^{-6}
rs11655588	A/G	0.18 (G)	17:18204137	ALKBH5	Intron variant	
rs1421090	A/G	0.29 (G)	16:53816258	FTO	Intron variant	0.000158
rs8071834	T/C	0.45 (C)	17:18196677	ALKBH5	Intron variant	
rs17820875	A/G	0.12 (G)	16:53892878	FTO	Intron variant	0.000177
rs8068517	G/A	0.24 (G)	17:18192664	ALKBH5	Intron variant	

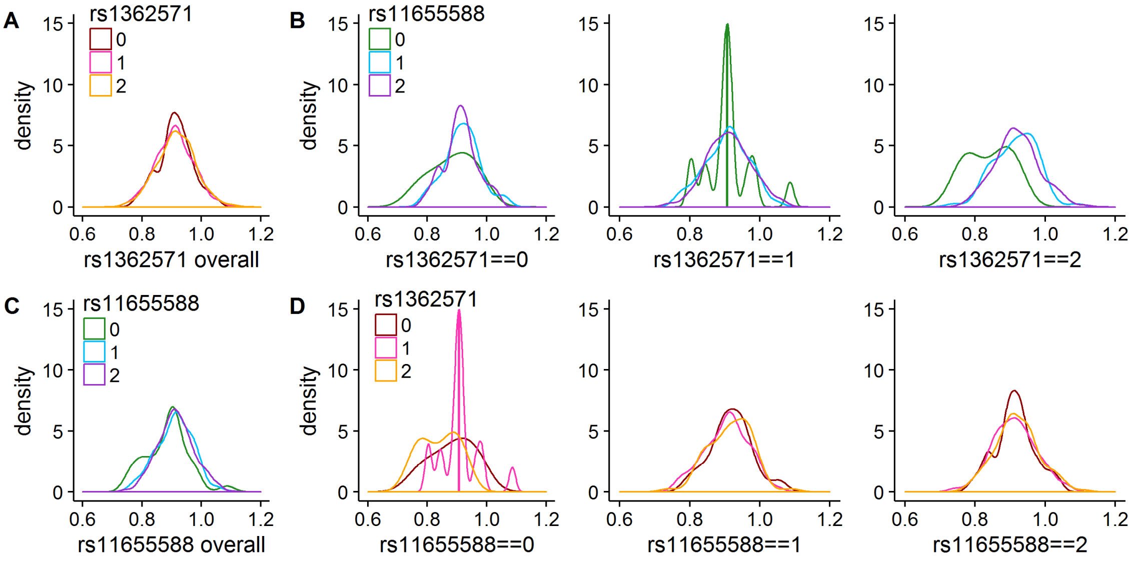


Figure 3. Phenotype distributions by genotype. A. Phenotype distributions for individuals with 0, 1, or 2 copies of the minor allele for rs1362571. B. Phenotype distributions for individuals with 0, 1, or 2 copies of the minor allele for rs11655588 amongst those with 0 copies of the minor allele for rs1362571 [left]; 1 copy [center]; 2 copies [right]. C. Phenotype distributions for individuals with 0, 1, or 2 copies of the minor allele for rs11655588. D. Phenotype distributions for individuals with 0, 1, or 2 copies of the minor allele for rs1362571 amongst those with 0 copies of the minor allele for rs11655588 [left]; 1 copy [center]; 2 copies [right].

The significant pair of SNPs identified from the simulated analysis did not reach significance in the analysis of the real data, although one of the SNPs (rs8068517) was a member of one of the significant interactions. Upon interrogating the phenotype distributions by genotype for the simulated data (Supplemental Figure 1), the real data (Supplemental Figure 2), and those individuals from the real data who were not included in the simulated data (Supplemental Figure 3), this appears to be driven by the 98 individuals who differ between the simulated and real data analyses. Although the differences in phenotype values by genotype between the real and simulated data may not appear extreme, the inclusion or exclusion of a small number of individuals may in fact significantly perturb the observed interaction. For example, there are 4 total individuals with 2 copies of the minor allele for rs2192872 and 0 copy of the minor allele for rs8068517 in the real data with a mean phenotype value of 0.900 (Supplemental Table 1). The simulated data includes only 2 of these individuals, for a mean phenotype value of 0.874, which makes this interaction appear much more protective compared to other interaction genotypes (particularly when considering the phenotype is defined as the log ratio of the average post-treatment triglyceride level to the average pre-treatment triglyceride level).

Discussion

Our gene set enrichment analyses were motivated by the goal of making inferences about the mechanism of action of the fictitious drug, assuming that differential methylation could reveal a set of genes associated with a drug that is functionally similar to the fictitious one in question. Upon attending GAW, it became apparent that interaction analysis of the simulated data was not appropriate given the nature of the simulation, and the interaction identified can therefore be considered a false positive. However, re-implementing the same analytic pipeline using the real data produced largely comparable results and identified four pairs of loci between FTO and ALKBH5 with significant interactions. The joint evidence implicating adenosine as the top result from the drug association gene set enrichment analysis and numerous miRNAs involved in metabolic traits from the miRNA target analysis, taken with the assumption that the drug has some unknown epigenetic mechanism, lead us to consider that mRNA or miRNA adenosine methylation, rather than CpG methylation, may be associated with drug response. MiRNAs in general are known important regulators of lipid metabolism [31, 50, 88]. The top miRNA hit, miR-124a, has evidence for both its role in metabolic traits [7, 37, 73, 122], and for its epigenetic regulation in the context of risk of diverse diseases [2, 15, 20, 35, 41]. N^6-Methyladenosine (m^{6}A) is a reversible, dynamic post-transcriptional modification that is regulated by miRNAs, and its demethylation has been shown to regulate adipogenesis [18, 34, 82]. Recent work has demonstrated that RNA conformational changes induced by m^{6}A determine substrate specificity for the two RNA demethylases, FTO and ALKBH5 [145]. If miRNA adenosine methylation rather than CpG methylation affects the phenotype, although the available data lacks observations on miRNA adenosine methylation, interactions between the two genes that demethylate miRNAs may be biologically relevant and can be assessed with the GOLDN SNP data. Given the evidence for physical interactions between RNA with the m^{6}A mark and these demethylases, we were motivated to check for epistasis between SNPs on these genes. Although we did find four pairs of loci with statistically significant interactions, the small sample size means that some SNP-SNP genotypes have few observations, warranting investigation of this interaction in a larger study and further molecular clarification of the distinct and mutual roles of FTO and ALKBH5. Rather than attempting to explain complex phenotypes solely in terms of single locus main effects, we posit that interaction models better represent the underlying regulatory nature of the genome, and that the joint effect of perturbations to multiple interacting partners can help better explain complex phenotypes. This analysis of epistatic interactions between loci on two genes serves as an illustrative example of how interactions can be significant in the absence of significant main effects, and highlights the need for analyses that integrate multiple sources of data to narrow the search space for plausible interactions.

CONCLUSIONS

Summary of the present work

The present work arises from the need to integrate and analyze large amounts of genomic data with a focus on challenges to the ability to detect and replicate epistatic interactions. Epistatis, or deviations from additivity in the interaction of genotypes to produce the resultant phenotype, may help explain the “missing heritability” of complex disease, but can be difficult to detect statistically due to the multiple testing burden associated with naively testing all pairwise combinations of loci. Furthermore, the potential for significant interactions in the complete absence of single locus main effects of the interacting partners and vice versa complicates prioritization of loci to test based on filtering via presence of significant main effects. Narrowing the search space to alleviate the multiple testing burden requires either the development of methods that consider qualities intrinsic to the data that may affect the power to potentially detect an interaction (such as low minor allele frequencies of the SNPs being investigated), or integration with supplemental sources of information that can provide additional context to be used for filtering or other feature selection. Beyond the difficulty associated with detecting an interaction, replication of an interaction may be hindered by population-level differences in allele frequency, heritability, prevalence, or other demographic, clinical, or environmental variables. Furthermore, differences in the computational pipeline used to analyze this data (including the ordering of workflow steps, quality control cutoffs, versions of dependent software, etc.) may determine whether an interaction is tested in the first place or thereafter ascertained to be significant. This necessitates the consideration of how standard statistical analyses, quality control and data cleaning procedures, or parameter tuning choices in machine learning analyses can be modified to improve the ability to detect and replicate epistasis.

I begin with a review on improving the reproducibility of machine learning analyses of genomic data, in which I discuss the distinctions between reproducibility and replicability, highlight potential barriers to performing analyses of genomic data, emphasize and the difficulty of detecting and replicating interactions, and consider how choices at each stage of a machine learning analysis may impact its results. This review highlights impediments to genomic big data analyses and suggests topics of open interest to contextualize the present work. The following two chapters describe methods for improving the reproducibility of interaction analyses of genomic data: implementing a resampling technique to recapitulate the power to detect interactions across populations with differences in minor allele frequencies, and implementing a cross validation technique that preserves relative genotype proportions to allow for improved detection of interactions for internal feature selection. These methods address barriers to replication of epistatic interactions between populations and reproducibility of interaction detection within a data set, respectively. The final chapter presents an example of an analysis that leverages a secondary data source with no significant association between methylation and the phenotype at any site to narrow the search space for further gene set enrichment and interaction analyses to find statistically significant interactions between the two RNA demethylases FTO and ALKBH5. This serves to emphasize that even “negative” data may be useful for contextualizing or supplementing other data sources, and by extension underscores the utility of GWAS data as one of many informative data sources that can be consolidated to produce a comprehensive, integrated personalized health risk assessment.

Overall, this work highlights and addresses some of the concerns emerging from the rapid coalescence of genomics and data science, and how experimental and analytical design choices in both domains and their fusion can greatly impact results. Understanding the downstream consequences of these choices and how they may feed in to each other demands in-depth interrogation of many small perturbations to the entire experimental process from the wet lab or clinic to the cloud. The future of high-quality, reproducible biomedical science will increasingly require the engagement of experts from diverse domains, underscoring the importance of harmonizing technical or analytical advances from other fields to suit the unique structure and biases of new data sources. The combination of technological and computational improvements that allow for the unprecedented rate of biomedical data accumulation, and the increasing emphasis on the importance of reproducible, interdisciplinary, and translational research can converge to enable an era of unprecedented scientific productivity and advancement through the cooperation of many agents. The following passages describe potential extensions of the present work and considerations for future research that arise from the present work.

Future directions

A logical extension of the present work is interrogating how the proposed resampling, cross validation, and data integration techniques may differentially influence findings across machine learning methods and how they may need modification to appropriately suit the particularities of each method. This work has investigated how factors such as minor allele frequency, prevalence, heritability, sample size, etc. impact results in the context of logistic regression models and uses the likelihood ratio test of the models with and without the interaction to quantify the significance of an interaction, as per standard analyses of GWAS data. However, we should consider other methods for detecting interactions and quantifying evidence of significance, how consistent these methods are with each other, and whether some are particularly suited for certain classes of problems. This may become increasingly important especially as analyses of GWAS data include integration or supplementation with external data sources that heighten the computational, storage-related, and interpretive demands. The choice of algorithm and its relevant parameters or hyperparameters, the ordering of steps in an analytic workflow, and the potential propagation of errors associated with integrating multiple methods and data types and sources must all be investigated to determine their relative contributions to failure to replicate across a range of population parameters.

Relatedly, the emphasis on statistical significance and p-values alone does a disservice to improving replication of both main effects and interactions. As data becomes sufficiently large, researchers are essentially guaranteed to find significant results, but the practical interpretation of these may be lacking. Re-analyzing existing GWAS data considering consistency in the direction and magnitude of effects for findings that fail to reach statistical significance and supplementing analyses with orthogonal data sources may provide additional informative clues regarding the underlying genetic architecture. A new GWAS analysis paradigm may be the automated meta-analysis of all new studies in the context of relevant evidence from prior studies. Such a system could also be used to prioritize which variants would be most informative to focus on in subsequent work, for example variants of large effect size that fail to reach statistical significance or variants at very different frequencies between populations. However, the results of meta-analyses may unfortunately obscure true relationships due to differences in bias between data sources, inconsistent phenotyping, unmeasured epidemiological confounding, or opposite directions of effect between populations. This highlights the adjacent importance of performing smaller, very well-controlled and precisely-defined studies across multiple homogeneous populations to better define and understand common versus subgroup-specific factors in the genetic architecture underlying traits to enable identification of barriers to replication. For example, it may be necessary to perform separate analyses of men and women in order to identify determinants of risk that differ by sex such as variants or interactions with variants on the X or Y. Sex-specific analysis may be especially worth consideration in light of non-genomic differences in exposures that may modify risk, such as behavioral or hormonal factors. Simulation studies that aim to characterize the ability to detect and replicate interactions across differentially confounded studies with the same true causal interaction may provide insight in interpreting real data with unknown sources of confounding.

The interactions explored in the application of PICV to primary open-angle glaucoma data and in the GAW analysis provide additional opportunities for follow-up in confounder analyses and data integration methods. Both of these interaction analyses may be improved via integration of additional sources of data such as tissue-specific gene expression data. The potential confounding by myopia identified in the PICV analysis suggests a need to better integrate genomic and traditional epidemiologic data and to design studies and develop analytic methods that can better facilitate exploration of gene-by-environmental and higher order combinations of genetic and environmental effects, perhaps with combined genomic and environmental matching and weighting techniques. The PICV analysis may for example be supplemented by data from genomic analyses of myopia, particularly with known glaucoma status and measured diopters of correction.

Towards greater reproducibility in the biomedical sciences

Although numerous fields grapple with concerns raised by the reproducibility crisis, this anxiety can be embraced as an opportunity to critique and radically reconsider how current scientific traditions and practices contribute to a culture of irreproducibility and the measures that can be taken towards creating structures and systems to enable more reproducible science. Improving reproducibility in genomics and the biomedical sciences more broadly will, besides technical advances in both the experimental and analytical branches, require improved integration of data collection and analysis, particularly in linking and jointly analyzing many heterogeneous data sources. Streamlining experimental workflows to autonomously and consistently collect and clean data in the format appropriate for analysis can reduce the potential for accidental input errors. Expanding this paradigm to facilitate the linking or integration of multiple data types in a standardized way can further promote reproducible analysis. Studies may reasonably need to synthesize data of diverse types and formats in order to consider patient demographics and personal history, family history, doctor’s notes, lab results, personal genome testing results, wearable device tracking output, etc. Analyzing this data in a reproducible framework that considers their intrinsic biases, levels of missingness, and relationships to each other suggests the need for a standardized personalized medicine hierarchy or ontology. Formalizing the hierarchical structures of the relationships between concepts could facilitate the exploration of risk attributable to perturbing exposures individually and in the context of confounders or effect measure modifiers. Such a structure could also improve the ability to compare findings between populations with different exposures and genetic architectures and provide a framework for modeling these differences or using them for weighting predictions.

Efforts to synthesize this data will likely require the expertise of individuals from numerous fields including the clinical sciences, basic sciences, computer sciences, informatics, statistics, etc. as the depth and breadth of necessary knowledge and skills is beyond an individual traditional academic domain. In addition to enhancing interdisciplinary exposures in the classroom or providing more workshop opportunities to develop working knowledge of important concepts in adjacent fields, the increasing need for scientists of diverse backgrounds to work together can be facilitated by creating more interdisciplinary programs to train scientists who can straddle these fields and serve as liaisons. Current training programs may also offer greater opportunities for trainees to interact between departments, perhaps via course offerings that may benefit diverse types of trainees (such as introductory programming and statistics courses) or research lecture series to promote current interdisciplinary expertise needs to solicit desired collaborators from other departments. This effort could be expanded to be independent of home institution via a research matchmaking web service, which would further foster relationships between scientists of varying experience and expertise around the world by connecting potential collaborators, mentors, or trainees. Seeking multiple mentors of differing expertise can provide a student with a well-rounded training experience that enhances interdisciplinary skills, while supporting trainees with a diverse range of research interests can help more established scientists stay connected to current interdisciplinary advances. Promoting overlap in domain expertise, as opposed to more rigidly defining specialization, is critical to the advancement of reproducibility in interdisciplinary academia. Scientists from more diverse training programs and with more diverse networks of collaborators can help mitigate the potential for misunderstandings that may arise from field-specific differences in vocabulary, conceptual false friends, etc. They may also be particularly poised to recognize how standards, practices, or tools from one traditional field may be superimposed upon or modified to suit analogous problems in another. Educating a new generation of creative scientists encouraged to dismiss traditional barriers and synthesize ideas from a multidisciplinary perspective will serve to enhance research across academia.

Academia may also benefit from increased collaboration with external partners such as the tech industry, or simply take inspiration from tech practices. Adopting practices that promote productivity and collaboration may improve cooperation of scientists from diverse fields, while versioning, issue tracking, automation, and backup tools may improve analytic reproducibility. De-centralizing data storage, accessibility, and analytic tools so that they are no longer siloed in individual institutions or even departments can also contribute to a culture of reproducibility by enabling the re-analysis or integrative analysis of data. Data parasitism, despite the negative connotation, serves to improve research efficiency, reduce waste, and lead to new discovery. Consolidating thematically similar sources of data and interrogating why they do or do not support the same hypotheses, re-analyzing negative data in the context of other negative data or secondary sources, assessing the most impactful data collection that can be performed to maximize the utility of existent data sources, and other secondary and integrative data analysis modalities may improve replicability of biomedical research findings via contextualization and meta-analyses.

The emergent recognition of and debate over data parasitism highlights an adjacent concern, traditional standards for publication and the increasing recognition of the utility of reporting “unsuccessful” experimental results and data descriptions. Reporting all results, even ones that are not novel or exciting, is a necessary part of improving reproducibility. This saves time, money, and effort on unnecessarily conducting analyses that other researchers have already performed, while also presenting the data for other researchers to critique and consider the impact of applying different data cleaning or analytic methods. Traditional publication of even successful results may also contribute to irreproducibility if page length and upload size limits preclude descriptions that are in-depth enough to re-run the experiment and analysis to arrive upon the same findings. A solution could be an entirely open-access journal with standardized data upload and reporting formats tailored to experiment type that performs exploratory data analysis, visualization, etc. automatically in an executable notebook. This would enable centralized data access and straightforward comparison, and all analyses performed beyond or in modification of the standard could automatically raise flags that demand an accompanying justification for each change.

The importance of epistasis and interaction-centric research

Pathway and network-based analyses of genomic data are increasingly appealing ways to represent the molecular interactions, hierarchies, and cascades that give rise to human phenotypes as we recognize that individual variants may be less important than the joint effects of many variants distributed across the genome. This implicitly acknowledges the importance of epistasis in explaining the “missing heritability” of complex disease, and also emphasizes the need to revisit heritability estimates to acknowledge the contribution of interactions between loci. It also stresses the adjacent needs to consider higher-order interactions between variants and gene-environmental interactions in risk prediction; the genome does not exist in isolation, and like interactions between variants, interactions between genes and the environment may also be non-additive and exist in the absence of main effects. Studying epistasis and adopting a more interaction-centric approach to research in general is emblematic of embracing rather than ignoring or rejecting the inherently interdependent, hierarchical, and complex nature of biology. Considering the contribution of interactions may additionally produce more reproducible research via substantiating the importance of higher-order relationships over individual variants; where single variant hits or the interactions between them may fail to replicate, quantifying the cumulative risk or protection conferred by many variants may be more stable within and across populations. Updating risk predictions for interactions between genetic and environmental determinants of disease in a probabilistic way based on observations from many studies can additionally allow for models that attempt to quantify uncertainty around individual and joint parameter combinations for a more full understanding of the structure of the relationship between variants and other risk factors.

Studying variants in terms of their greater genomic and environmental context is also necessary for ensuring equitable access to the benefits of personalized medicine in the future. In order for personalized medicine to be both maximally impactful and socially just, genomic science must address its diversity problem and focus heavily on the importance and challenges of studying non-Caucasian populations, especially considering the inequitable distribution of research resources across populations and the impact of globalization and the increasingly multiracial world. However, the current standard of substantiating GWAS findings in a replication cohort of the same ancestral background may hamper the ability to request funding or justify performing research in other populations and thereby serve to perpetuate the currently unjust state of research. Thoroughly exploring how differences in risk associated with variants or their interactions may be dependent upon population-specific genetic architecture or environmental exposures is necessary for truly quantifying the effect of variation at a given locus, and will require complementary work across many populations, particularly those with greater admixture. Estimates of variant risk derived in one population cannot be used to predict risk in another, although performing analyses across many populations can of course lend robustness to findings that replicate, and provide informative contextual clues as to why others may not. Analyses of interactions, networks, and pathways rather than individual loci may become the new paradigm for exploring genomic data in a future that embraces population differences as informative rather than inconvenient.

Conclusions

The increasing interconnectedness of the biomedical sciences, both in terms of the multidisciplinary expertise required for conducting modern research and the imperative to consider joint effects at multiple levels of complexity, can facilitate a more robust, cooperative, and translational research future. It is possible that all the necessary computational and methodological advances necessary to truly understand human genetic variation have already been made across diverse fields, and that we are on the cusp of substantial and immediate benefit to human health and personalized medicine upon context-specific translation of these tools and ideas. The rapid influx of genomic and other biomedical data should be welcomed as resources that will enable and empower researchers to explore complex systems in unprecedented depth to truly impact health and medicine. Although there remain many serious barriers to improving the reproducibility of interaction analyses and scientific reproducibility in general, we should be inspired rather than daunted by this complexity and the exciting research opportunities it provides.

APPENDIX

Supplemental Tables and Figures, Chapter 3

Supplemental table 1. Data set simulation parameters. Minor allele frequencies and penetrance tables used to generate balanced case-control ratio data sets of size 10,000. Heritability = 0.005 and prevalence = 0.1 constant across all simulations.

	Scenario 1			**Scenario 2**			**Scenario 3**		
SNP1 MAF:	0.1			0.2			0.2		
SNP2 MAF:	0.1			0.1			0.2		
	0.205	0.182	0.102	0.208	0.190	0.148	0.189	0.213	0.273
Penetrance:	0.176	0.282	0.640	0.164	0.244	0.433	0.225	0.166	0.066
	0.219	0.110	0.248	0.202	0.198	0.192	0.173	0.266	0.094
	Scenario 4			**Scenario 5**			**Scenario 6**		
SNP1 MAF:	0.3			0.3			0.3		
SNP2 MAF:	0.1			0.2			0.3		
	0.211	0.194	0.166	0.201	0.203	0.182	0.220	0.172	0.223
Penetrance:	0.153	0.227	0.334	0.203	0.201	0.181	0.191	0.220	0.157
	0.151	0.186	0.532	0.166	0.146	0.639	0.134	0.260	0.276
	Scenario 7			**Scenario 8**			**Scenario 9**		
SNP1 MAF:	0.4			0.4			0.4		
SNP2 MAF:	0.1			0.2			0.3		
	0.188	0.199	0.228	0.221	0.191	0.180	0.202	0.218	0.143
Penetrance:	0.253	0.203	0.072	0.157	0.212	0.260	0.204	0.175	0.267
	0.180	0.212	0.208	0.206	0.248	0.040	0.172	0.221	0.198
	Scenario 10			**Scenario 11**			**Scenario 12**		
SNP1 MAF:	0.4			0.5			0.5		
SNP2 MAF:	0.4			0.1			0.2		
	0.159	0.228	0.209	0.205	0.207	0.180	0.190	0.192	0.225
Penetrance:	0.211	0.186	0.217	0.170	0.177	0.277	0.216	0.224	0.136
	0.259	0.179	0.130	0.296	0.026	0.452	0.229	0.130	0.310
	Scenario 13			**Scenario 14**			**Scenario 15**		
SNP1 MAF:	0.5			0.5			0.5		
SNP2 MAF:	0.3			0.4			0.5		
	0.185	0.188	0.239	0.154	0.204	0.238	0.224	0.279	0.219
Penetrance:	0.233	0.207	0.152	0.246	0.195	0.164	0.220	0.187	0.205
	0.124	0.234	0.208	0.167	0.205	0.222	0.236	0.247	0.170

Supplemental table 2. Data set simulation parameters. Minor allele frequencies and penetrance tables used to generate balanced case-control ratio data sets of size 10,000. Heritability = 0.005 and prevalence = 0.02 constant across all simulations.

		Scenario 1			Scenario 2			Scenario 3	
SNP1 MAF:		0.1			0.2			0.2	
SNP2 MAF:		0.1			0.1			0.2	
Penetrance:	0.022	0.013	0.008	0.023	0.016	0.002	0.016	0.027	0.019
	0.013	0.052	0.035	0.007	0.036	0.010	0.028	0.006	0.006
	0.016	0.008	0.709	0.022	0.016	0.029	0.012	0.016	0.183
		Scenario 4			**Scenario 5**			**Scenario 6**	
SNP1 MAF:		0.3			0.3			0.3	
SNP2 MAF:		0.1			0.2			0.3	
Penetrance:	0.024	0.017	0.009	0.017	0.027	0.002	0.017	0.027	0.003
	0.001	0.031	0.069	0.025	0.007	0.056	0.027	0.008	0.039
	0.022	0.021	0.005	0.023	0.017	0.016	0.003	0.039	0.023
		Scenario 7			**Scenario 8**			**Scenario 9**	
SNP1 MAF:		0.4			0.4			0.4	
SNP2 MAF:		0.1			0.2			0.3	
Penetrance:	0.018	0.024	0.011	0.019	0.025	0.006	0.031	0.012	0.009
	0.028	0.009	0.058	0.025	0.008	0.047	0.010	0.028	0.023
	0.026	0.010	0.034	0.003	0.031	0.023	0.010	0.012	0.067
		Scenario 10			**Scenario 11**			**Scenario 12**	
SNP1 MAF:		0.4			0.5			0.5	
SNP2 MAF:		0.4			0.1			0.2	
Penetrance:	0.005	0.028	0.029	0.022	0.023	0.012	0.025	0.024	0.008
	0.027	0.019	0.007	0.012	0.006	0.056	0.010	0.013	0.045
	0.031	0.005	0.039	0.023	0.014	0.030	0.025	0.020	0.017
		Scenario 13			**Scenario 14**			**Scenario 15**	
SNP1 MAF:		0.5			0.5			0.5	
SNP2 MAF:		0.3			0.4			0.5	
Penetrance:	0.006	0.023	0.029	0.041	0.012	0.014	0.045	0.004	0.026
	0.036	0.019	0.006	0.010	0.026	0.019	0.013	0.026	0.016
	0.022	0.010	0.040	0.004	0.020	0.036	0.009	0.025	0.021

Supplemental Table 3. Summary of performance measures across minor allele frequency combinations, prevalence = 0.5, n = 10000.

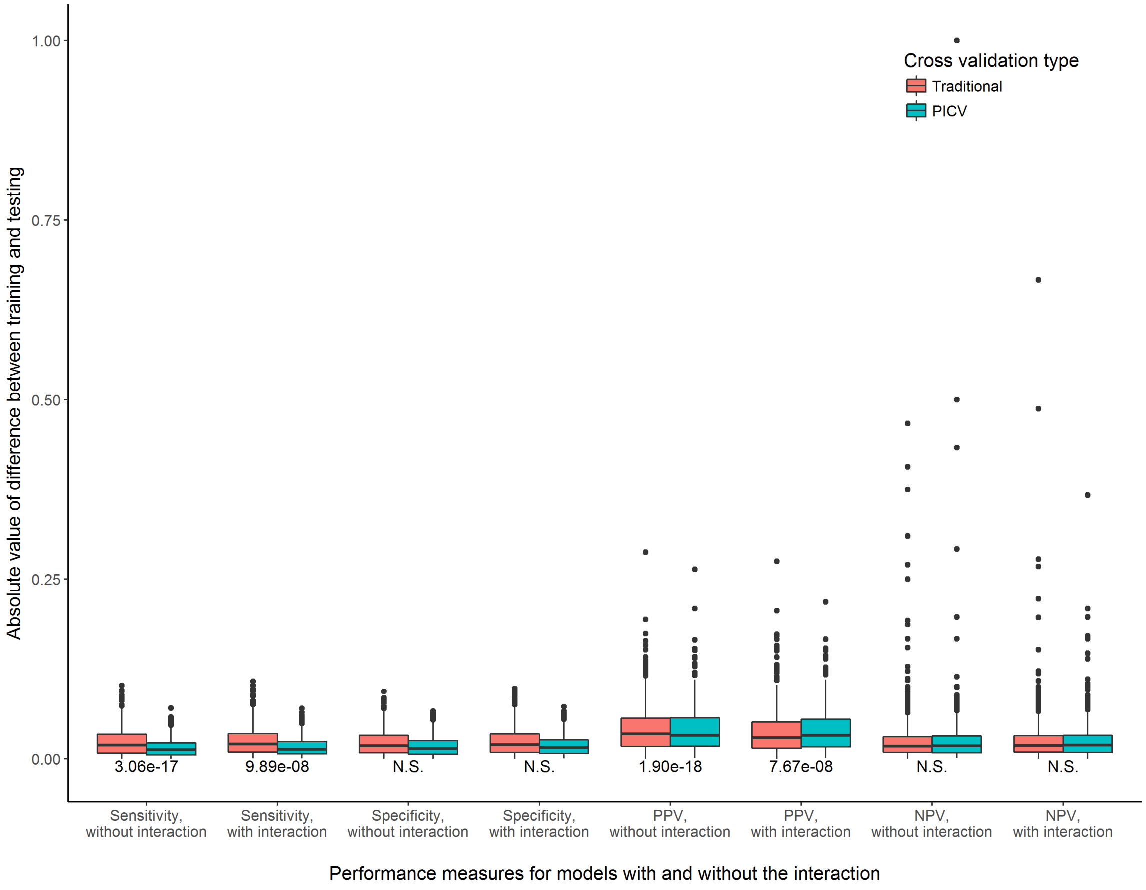
Measure, Model Scenario	Sens, without int	Sens, with int	Spec, without int	Spec, with int	PPV, without int	PPV, with int	NPV, without int	NPV, with int
SNP1 MAF: 0.1 SNP2 MAF: 0.1	8.60e-06	1.75e-04	N.S.	N.S.	4.12e-10	4.06e-09	N.S.	N.S.
SNP1 MAF: 0.2 SNP2 MAF: 0.1	6.92e-06	2.47e-05	N.S.	2.68e-02	3.03e-20	9.55e-22	4.35e-02	N.S.
SNP1 MAF: 0.2 SNP2 MAF: 0.2	9.27e-09	3.03e-05	N.S.	N.S.	5.01e-14	5.92e-15	2.73e-02	N.S.
SNP1 MAF: 0.3 SNP2 MAF: 0.1	2.27e-06	1.20e-03	N.S.	N.S.	1.74e-14	2.81e-13	N.S.	N.S.
SNP1 MAF: 0.3 SNP2 MAF: 0.2	6.65e-17	4.12e-10	N.S.	N.S.	1.48e-21	1.07e-19	N.S.	1.46e-02
SNP1 MAF: 0.3 SNP2 MAF: 0.3	5.56e-06	2.10e-07	N.S.	N.S.	1.75e-09	3.07e-16	N.S.	N.S.
SNP1 MAF: 0.4 SNP2 MAF: 0.1	8.97e-07	4.54e-05	N.S.	N.S.	7.73e-13	1.00e-13	N.S.	N.S.
SNP1 MAF: 0.4 SNP2 MAF: 0.2	2.66e-03	5.54e-05	N.S.	N.S.	2.86e-15	2.07e-17	N.S.	N.S.
SNP1 MAF: 0.4 SNP2 MAF: 0.3	4.38e-07	4.06e-09	N.S.	N.S.	1.44e-11	1.75e-09	N.S.	N.S.
SNP1 MAF: 0.4 SNP2 MAF: 0.4	3.63e-04	2.11e-04	N.S.	N.S.	1.97e-11	1.63e-05	N.S.	N.S.
SNP1 MAF: 0.5 SNP2 MAF: 0.1	5.57e-07	7.08e-07	N.S.	N.S.	1.43e-16	6.65e-17	N.S.	7.45e-03
SNP1 MAF: 0.5 SNP2 MAF: 0.2	4.06e-09	5.93e-08	N.S.	N.S.	3.95e-22	1.62e-19	N.S.	N.S.
SNP1 MAF: 0.5 SNP2 MAF: 0.3	1.27e-07	2.32e-09	N.S.	N.S.	4.00e-12	1.69e-10	N.S.	N.S.
SNP1 MAF: 0.5 SNP2 MAF: 0.4	8.97e-07	1.13e-06	N.S.	N.S.	1.43e-16	1.37e-15	3.08e-02	N.S.
SNP1 MAF: 0.5 SNP2 MAF: 0.5	4.06e-09	3.71e-05	N.S.	N.S.	5.53e-13	2.89e-12	N.S.	N.S.

Supplemental Table 4. Summary of performance measures across minor allele frequency combinations, prevalence = 0.1, n = 10000.

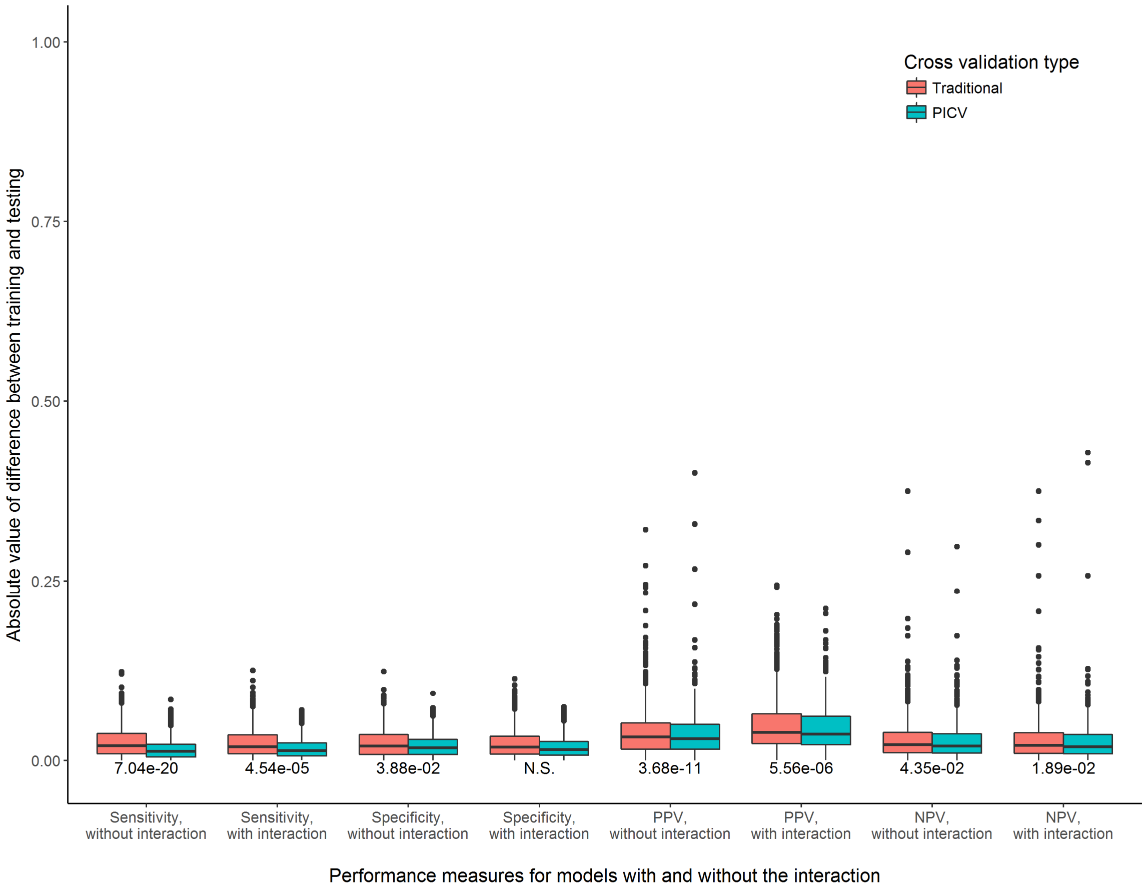
Measure, Model Scenario	Sens, without int	Sens, with int	Spec, without int	Spec, with int	PPV, without int	PPV, with int	NPV, without int	NPV, with int
SNP1 MAF: 0.1 SNP2 MAF: 0.1	6.75e-05	2.10e-07	N.S.	N.S.	1.74e-14	1.97e-11	N.S.	N.S.
SNP1 MAF: 0.2 SNP2 MAF: 0.1	2.08e-08	1.22e-08	4.86e-02	N.S.	1.97e-11	6.50e-16	N.S.	N.S.
SNP1 MAF: 0.2 SNP2 MAF: 0.2	1.07e-05	1.32e-05	N.S.	N.S.	1.97e-11	3.07e-16	N.S.	N.S.
SNP1 MAF: 0.3 SNP2 MAF: 0.1	3.07e-09	1.74e-14	N.S.	N.S.	3.07e-16	1.40e-17	N.S.	N.S.
SNP1 MAF: 0.3 SNP2 MAF: 0.2	5.57e-07	5.53e-10	N.S.	N.S.	4.47e-16	1.22e-14	N.S.	4.35e-02
SNP1 MAF: 0.3 SNP2 MAF: 0.3	1.07e-05	8.20e-05	5.21e-03	N.S.	6.15e-22	9.25e-11	N.S.	N.S.
SNP1 MAF: 0.4 SNP2 MAF: 0.1	2.71e-08	2.47e-05	N.S.	N.S.	1.22e-14	1.37e-15	N.S.	N.S.
SNP1 MAF: 0.4 SNP2 MAF: 0.2	8.60e-06	1.07e-05	N.S.	N.S.	5.92e-15	5.53e-13	N.S.	3.88e-02
SNP1 MAF: 0.4 SNP2 MAF: 0.3	9.87e-10	9.87e-10	N.S.	N.S.	5.53e-13	6.33e-18	N.S.	N.S.
SNP1 MAF: 0.4 SNP2 MAF: 0.4	3.71e-05	5.53e-10	N.S.	N.S.	5.02e-11	9.42e-18	N.S.	N.S.
SNP1 MAF: 0.5 SNP2 MAF: 0.1	9.25e-11	5.53e-10	N.S.	N.S.	7.73e-13	4.52e-17	N.S.	N.S.
SNP1 MAF: 0.5 SNP2 MAF: 0.2	5.54e-05	2.27e-06	N.S.	N.S.	1.22e-14	3.03e-20	1.46e-02	N.S.
SNP1 MAF: 0.5 SNP2 MAF: 0.3	1.32e-05	3.07e-10	N.S.	N.S.	2.27e-06	1.22e-08	N.S.	N.S.
SNP1 MAF: 0.5 SNP2 MAF: 0.4	6.75e-05	5.57e-07	N.S.	N.S.	9.25e-11	2.08e-08	N.S.	N.S.
SNP1 MAF: 0.5 SNP2 MAF: 0.5	1.22e-14	5.53e-12	N.S.	N.S.	6.50e-16	7.10e-14	N.S.	N.S.

Supplemental Table 5. Summary of performance measures across minor allele frequency combinations, prevalence = 0.02, n = 10000.

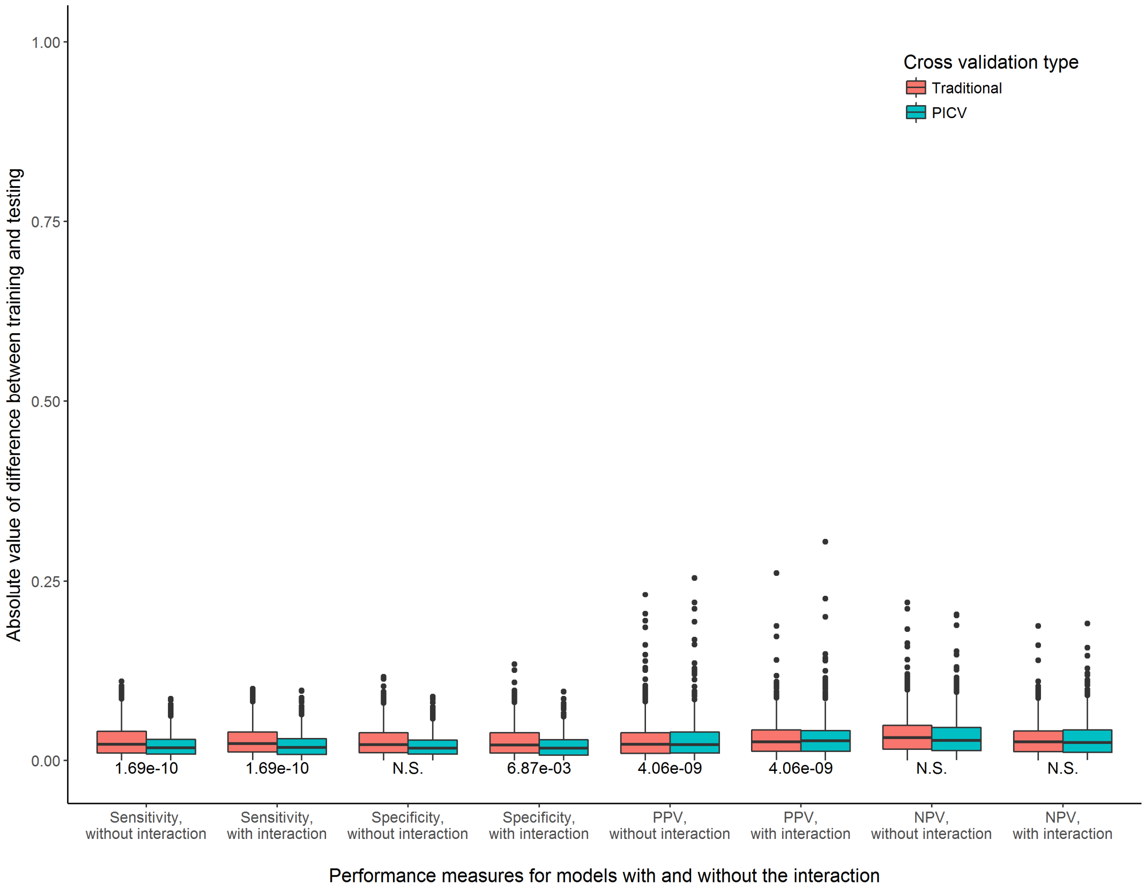
Measure, Model Scenario	Sens, without int	Sens, with int	Spec, without int	Spec, with int	PPV, without int	PPV, with int	NPV, without int	NPV, with int
SNP1 MAF: 0.1 SNP2 MAF: 0.1	2.28e-10	9.25e-11	4.20e-03	1.12e-02	1.40e-17	1.98e-20	1.46e-02	1.46e-02
SNP1 MAF: 0.2 SNP2 MAF: 0.1	1.80e-06	1.75e-09	1.03e-02	N.S.	1.25e-10	4.21e-24	N.S.	N.S.
SNP1 MAF: 0.2 SNP2 MAF: 0.2	7.10e-14	1.05e-11	N.S.	1.70e-02	9.27e-09	3.72e-19	N.S.	N.S.
SNP1 MAF: 0.3 SNP2 MAF: 0.1	3.07e-09	9.27e-09	N.S.	5.58e-03	9.25e-11	1.98e-20	N.S.	4.35e-02
SNP1 MAF: 0.3 SNP2 MAF: 0.2	1.13e-06	3.07e-09	4.09e-03	N.S.	2.32e-09	1.66e-24	N.S.	N.S.
SNP1 MAF: 0.3 SNP2 MAF: 0.3	1.32e-09	5.53e-12	3.16e-02	N.S.	1.25e-10	2.10e-16	N.S.	N.S.
SNP1 MAF: 0.4 SNP2 MAF: 0.1	1.08e-12	4.12e-10	N.S.	3.73e-02	3.52e-08	3.19e-27	N.S.	N.S.
SNP1 MAF: 0.4 SNP2 MAF: 0.2	4.12e-10	7.40e-10	1.80e-02	N.S.	8.49e-15	1.98e-15	N.S.	N.S.
SNP1 MAF: 0.4 SNP2 MAF: 0.3	7.73e-13	1.43e-16	1.07e-02	4.93e-03	7.67e-08	3.95e-22	N.S.	N.S.
SNP1 MAF: 0.4 SNP2 MAF: 0.4	5.46e-21	1.08e-12	N.S.	N.S.	2.89e-12	1.22e-14	N.S.	N.S.
SNP1 MAF: 0.5 SNP2 MAF: 0.1	5.36e-09	6.14e-04	N.S.	N.S.	6.82e-11	1.97e-11	N.S.	N.S.
SNP1 MAF: 0.5 SNP2 MAF: 0.2	4.38e-07	4.06e-09	N.S.	N.S.	2.69e-07	4.52e-17	N.S.	N.S.
SNP1 MAF: 0.5 SNP2 MAF: 0.3	7.06e-09	4.58e-08	N.S.	N.S.	7.73e-13	8.44e-19	N.S.	N.S.
SNP1 MAF: 0.5 SNP2 MAF: 0.4	1.43e-06	5.36e-09	N.S.	N.S.	1.44e-11	5.46e-21	4.35e-02	N.S.
SNP1 MAF: 0.5 SNP2 MAF: 0.5	5.02e-11	5.57e-07	N.S.	2.75e-05	3.44e-07	1.62e-19	N.S.	N.S.



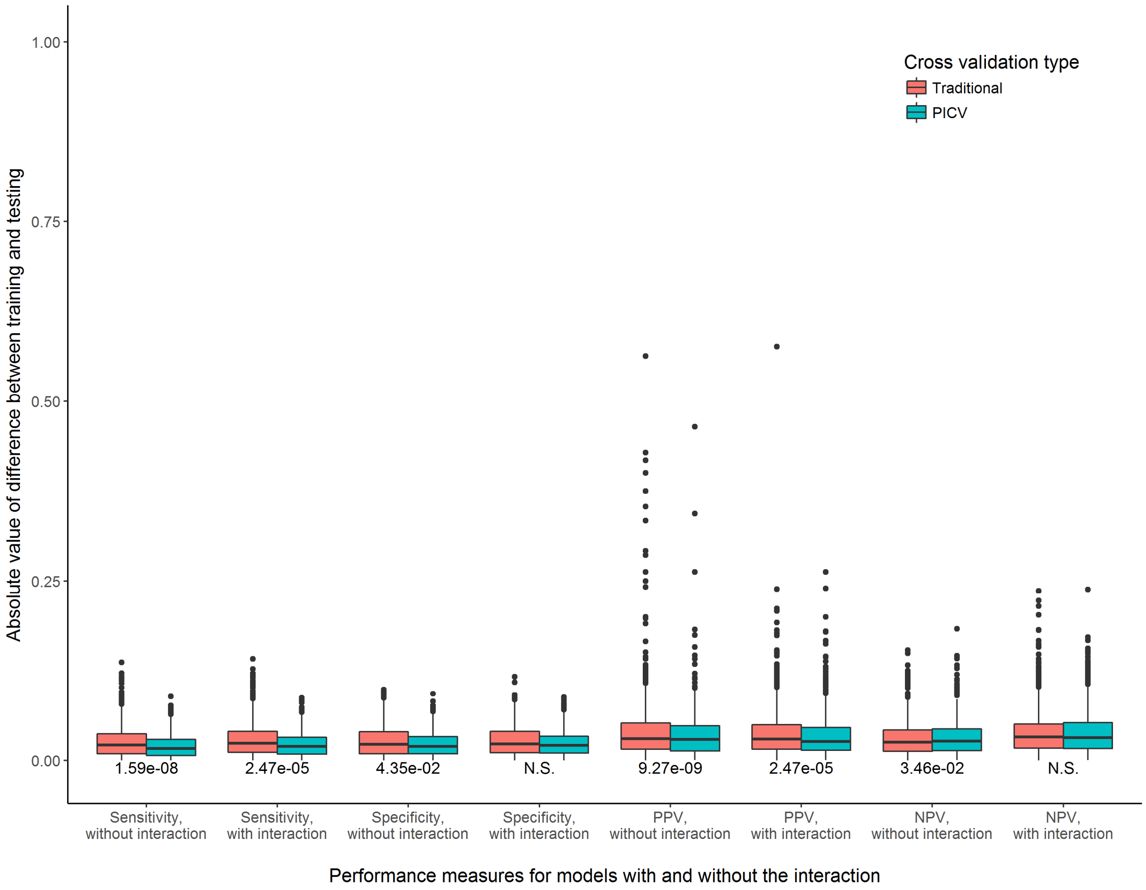
Supplemental Figure 1. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 1, prevalence = 0.5, n = 2000



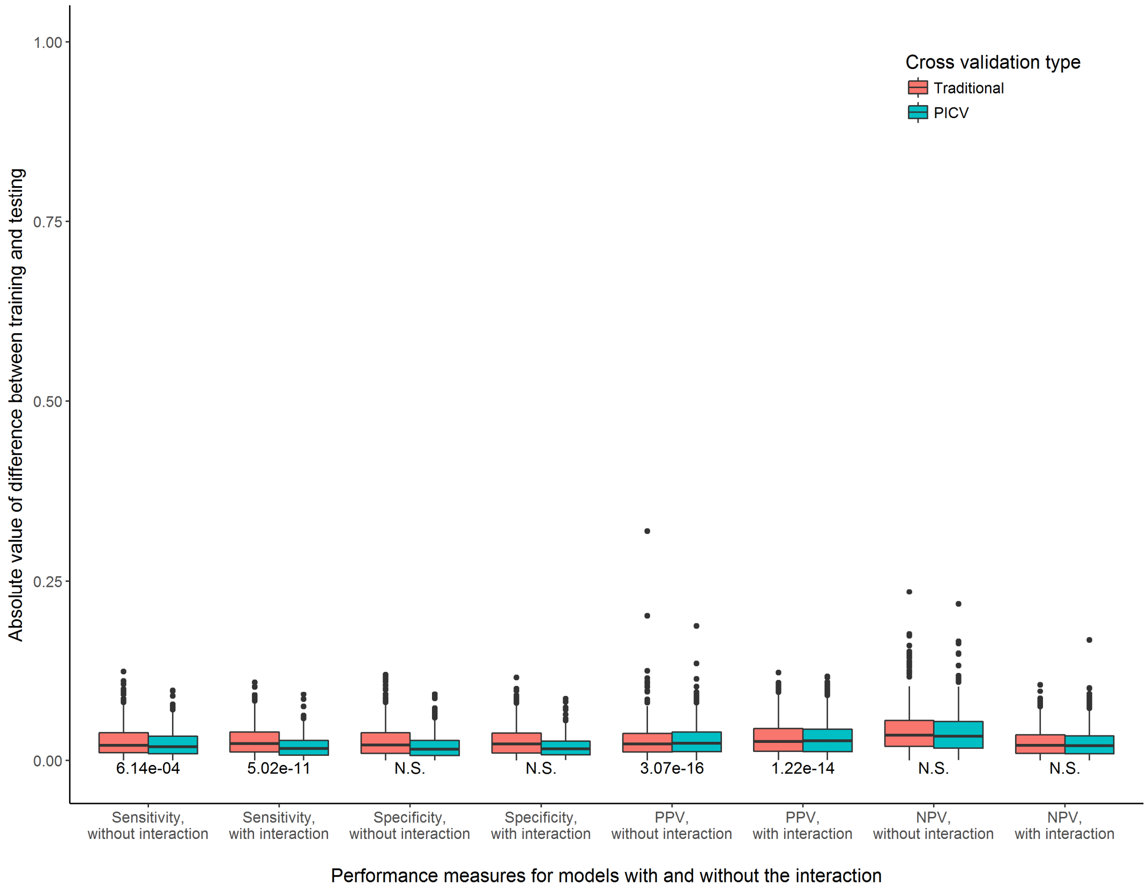
Supplemental Figure 2. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 2, prevalence = 0.5, n = 2000



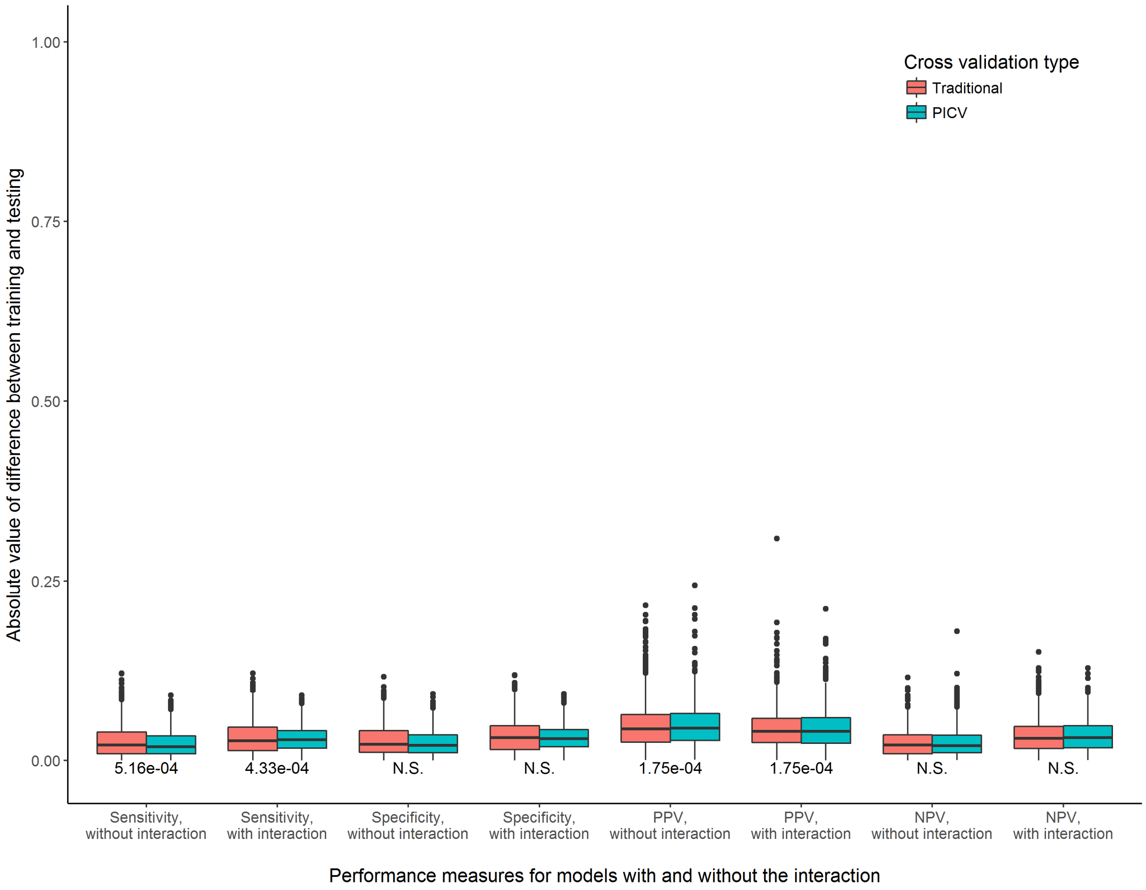
Supplemental Figure 3. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 3, prevalence = 0.5, n = 2000



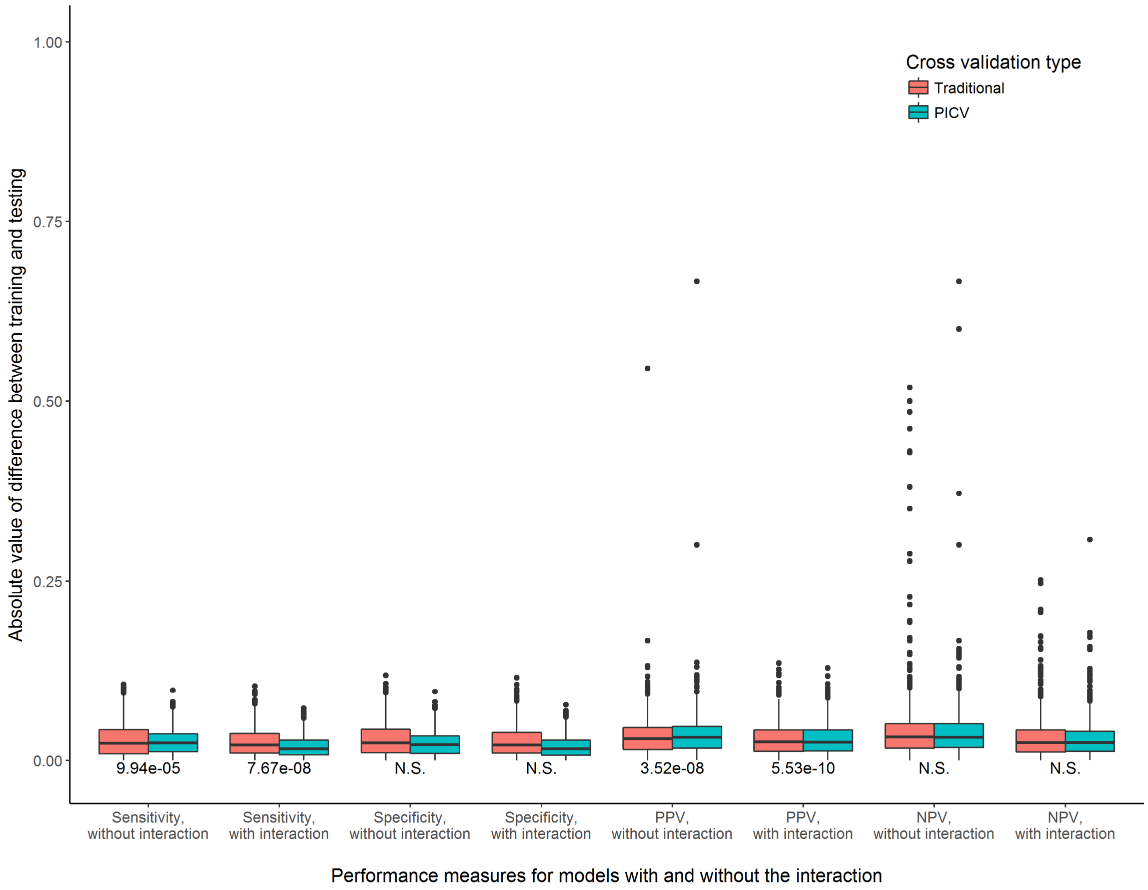
Supplemental Figure 4. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 4, prevalence = 0.5, n = 2000



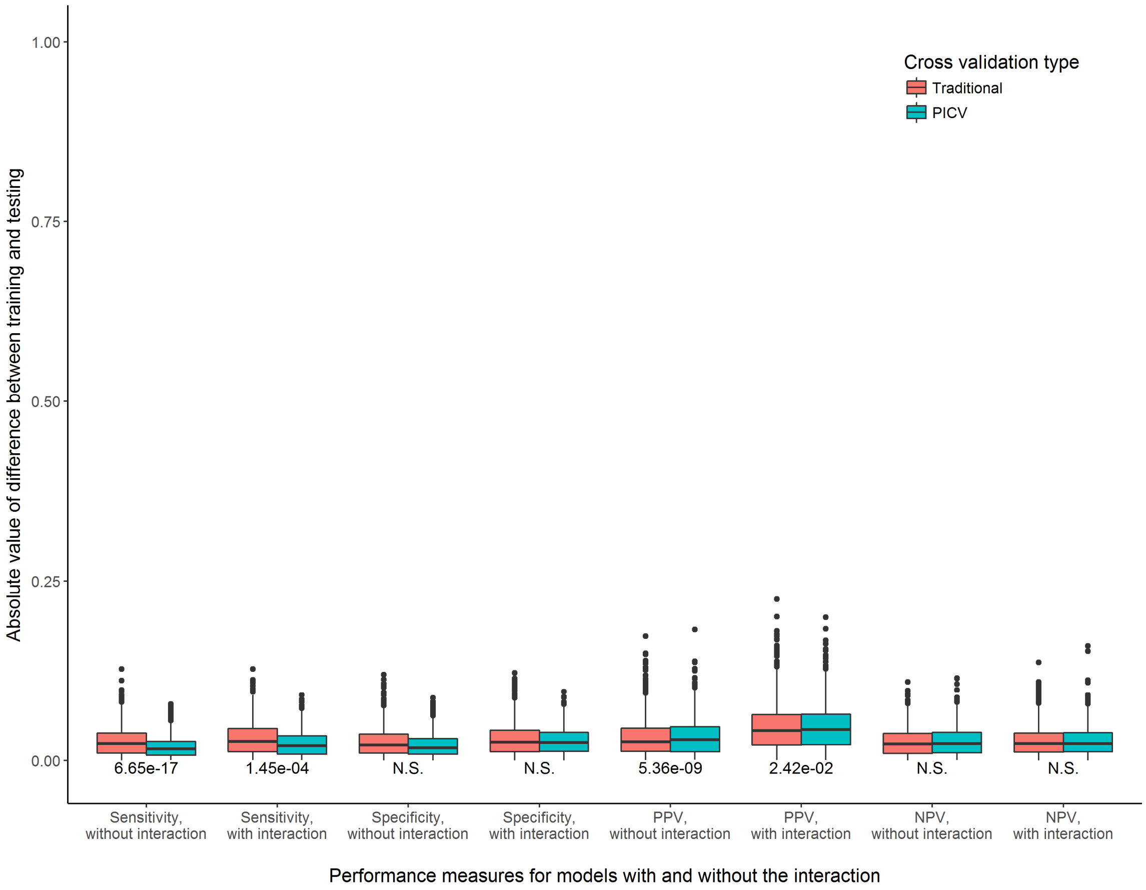
Supplemental Figure 5. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 5, prevalence = 0.5, n = 2000



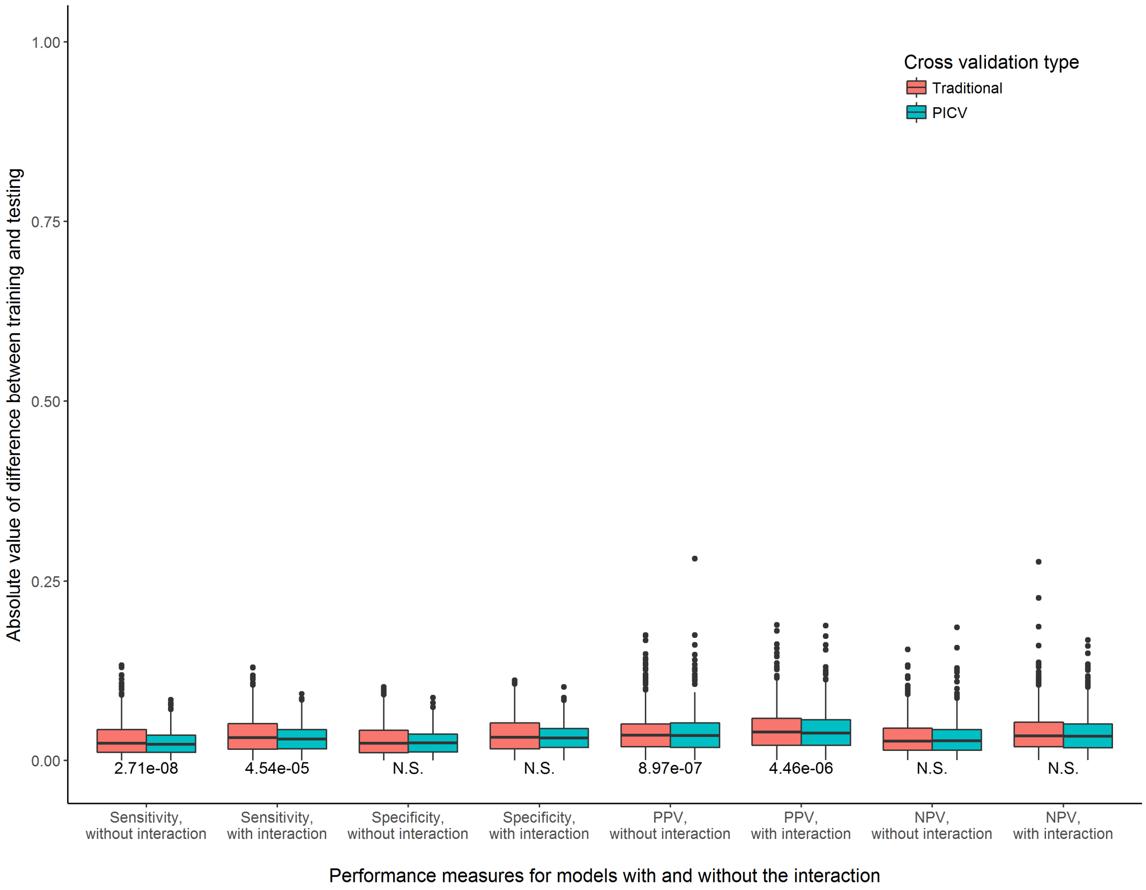
Supplemental Figure 6. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 6, prevalence = 0.5, n = 2000



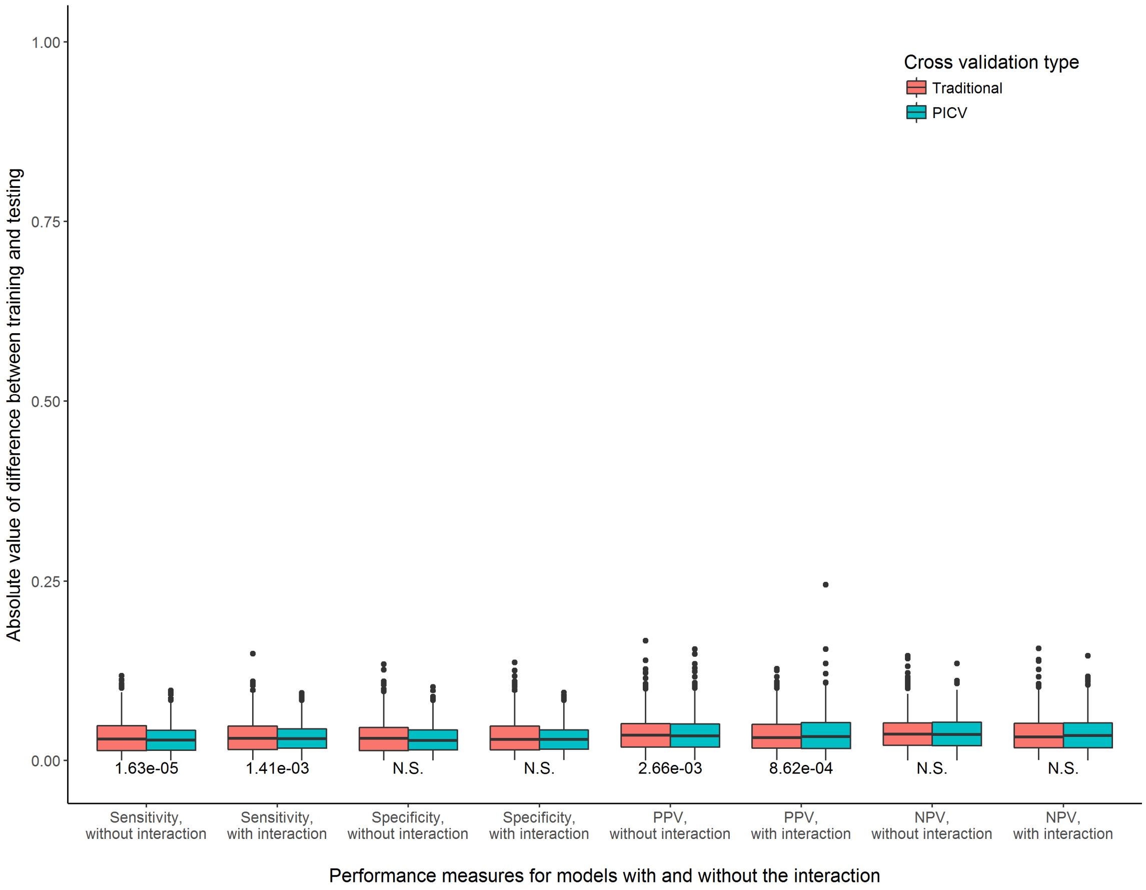
Supplemental Figure 7. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 7, prevalence = 0.5, n = 2000



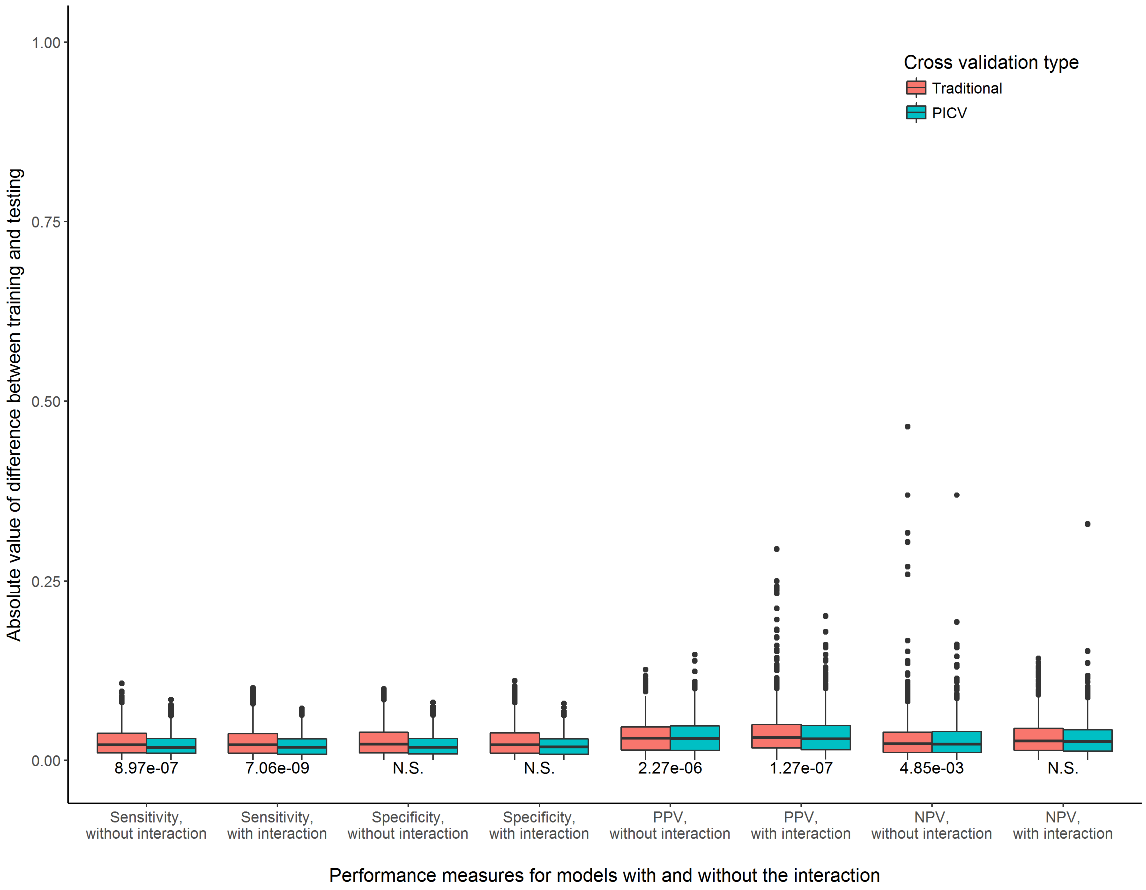
Supplemental Figure 8. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 8, prevalence = 0.5, n = 2000



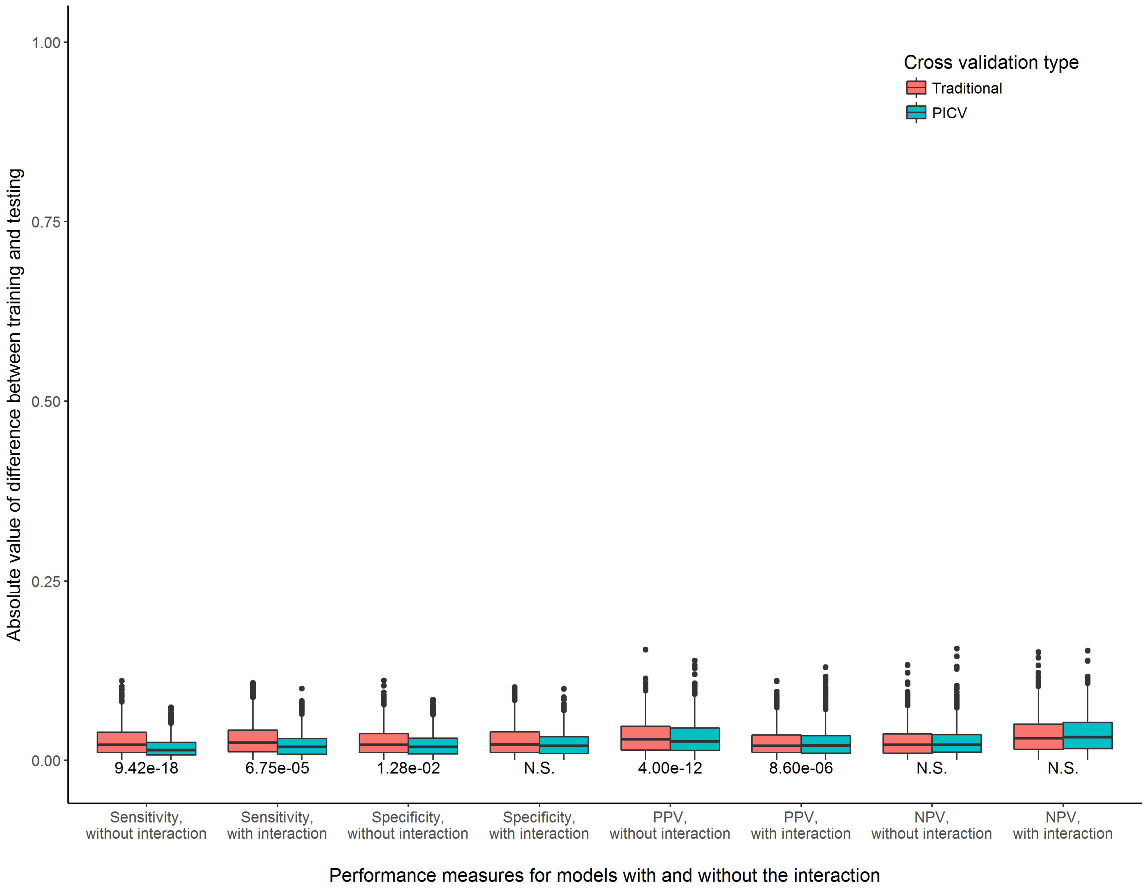
Supplemental Figure 9. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 9, prevalence = 0.5, n = 2000



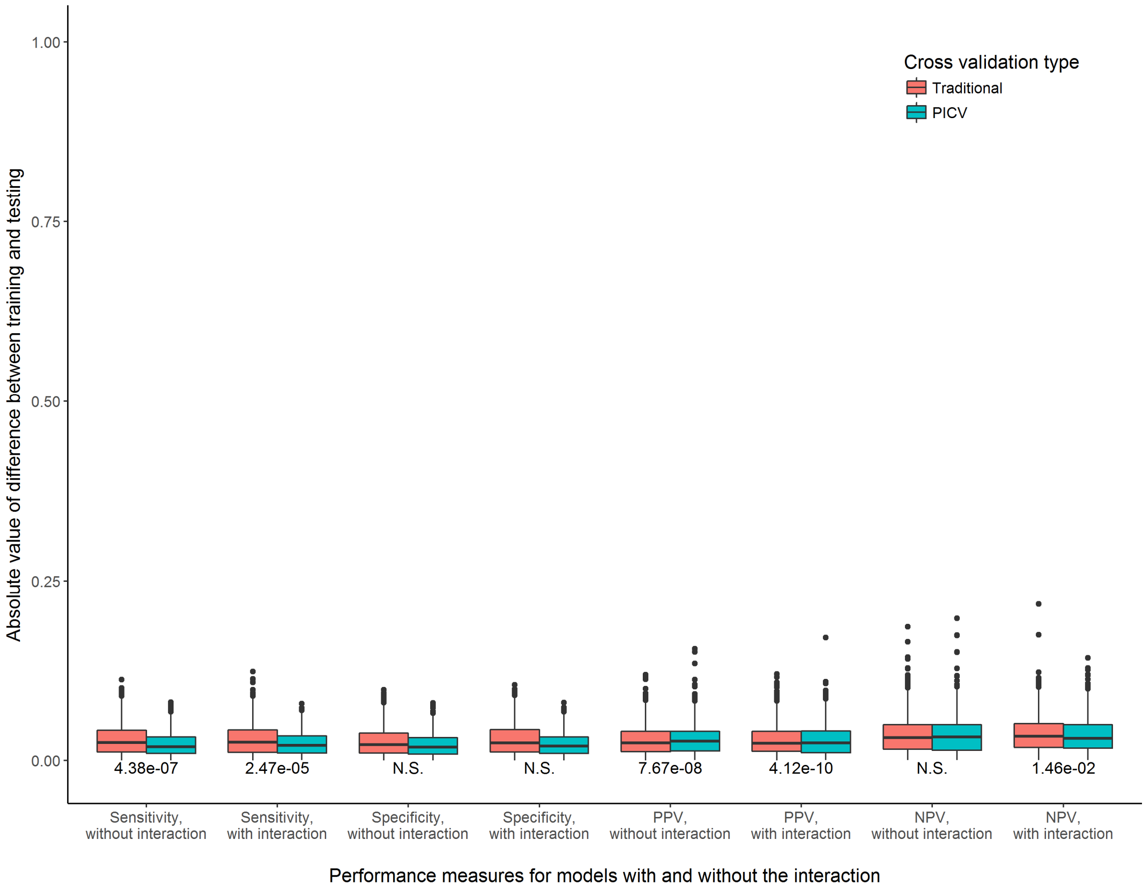
Supplemental Figure 10. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 10, prevalence = 0.5, n = 2000



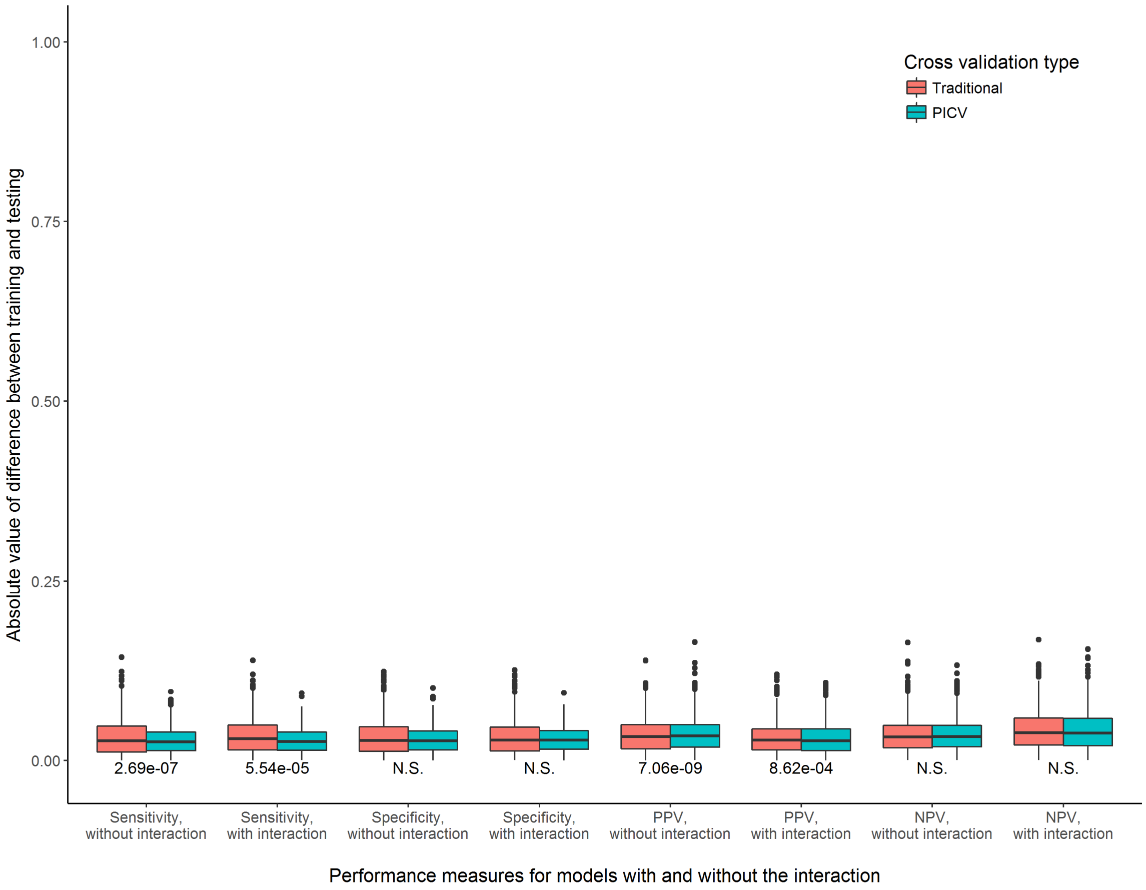
Supplemental Figure 11. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 11, prevalence = 0.5, n = 2000



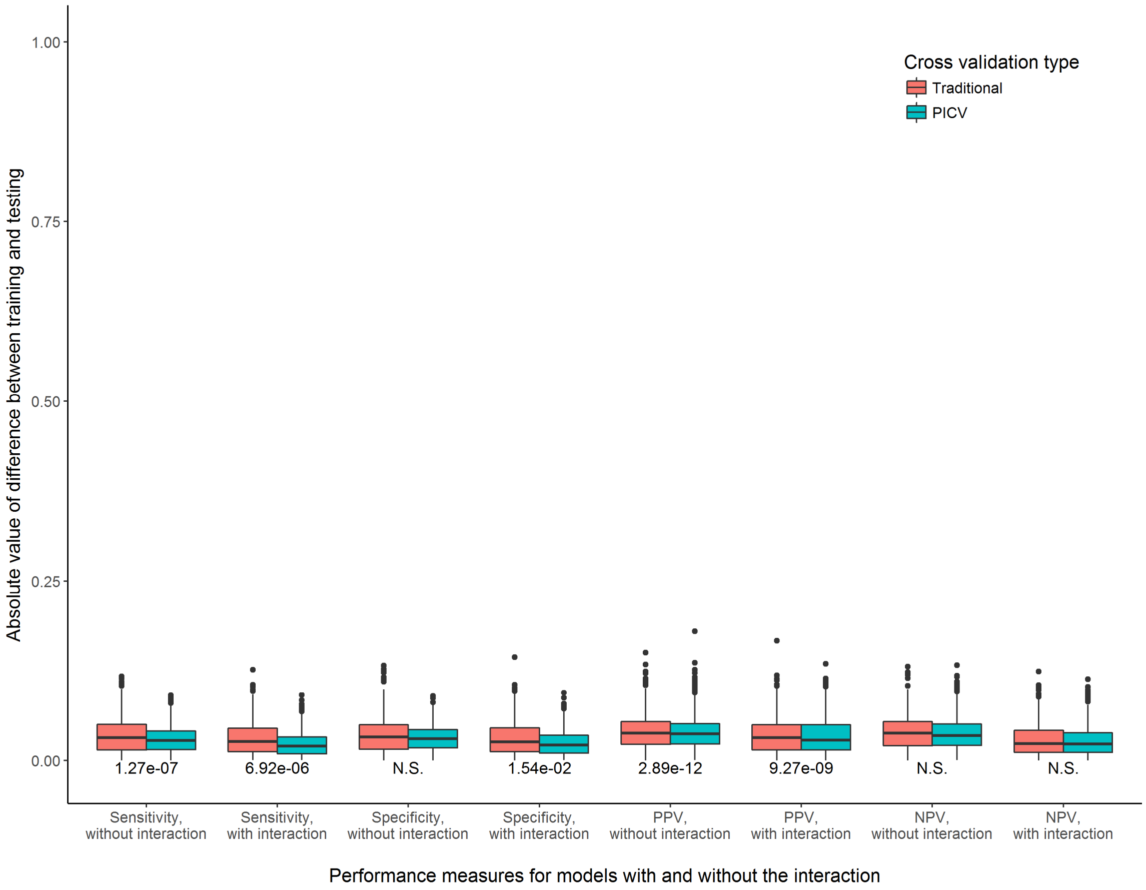
Supplemental Figure 12. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 12, prevalence = 0.5, n = 2000



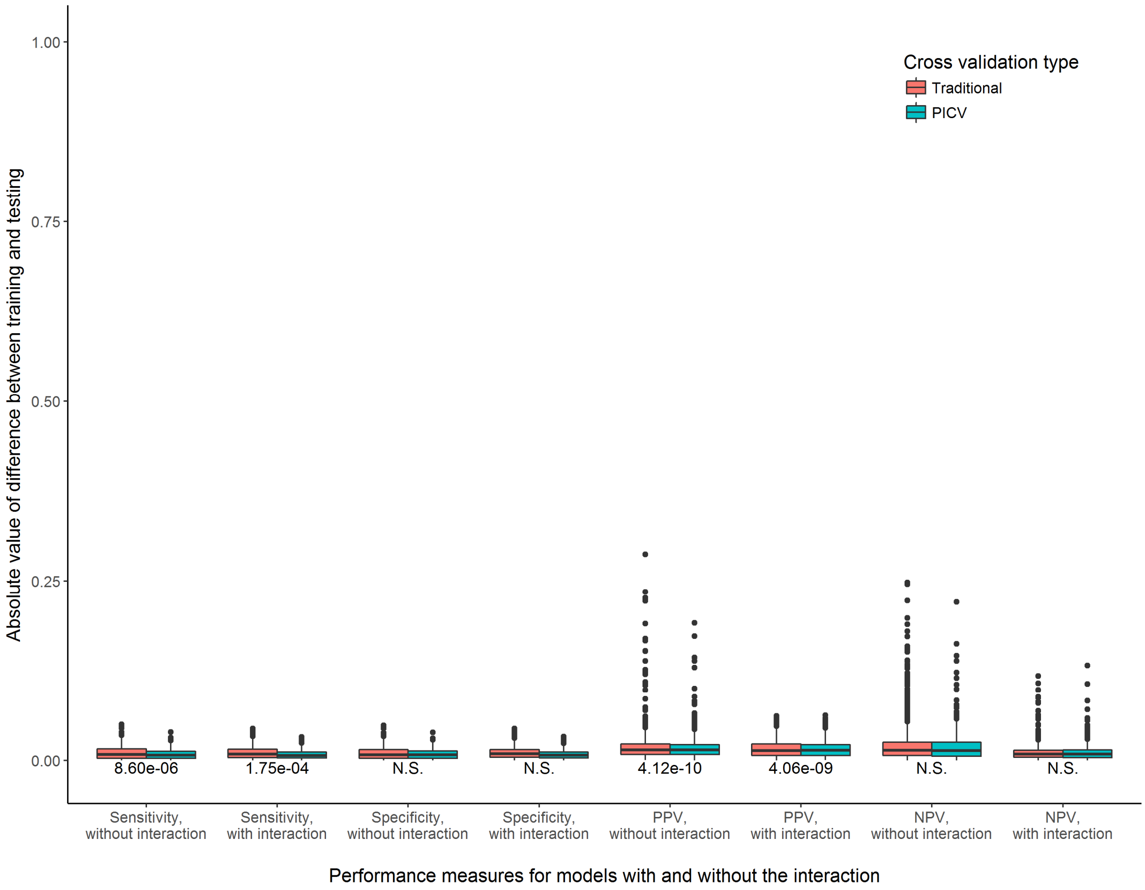
Supplemental Figure 13. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 13, prevalence = 0.5, n = 2000



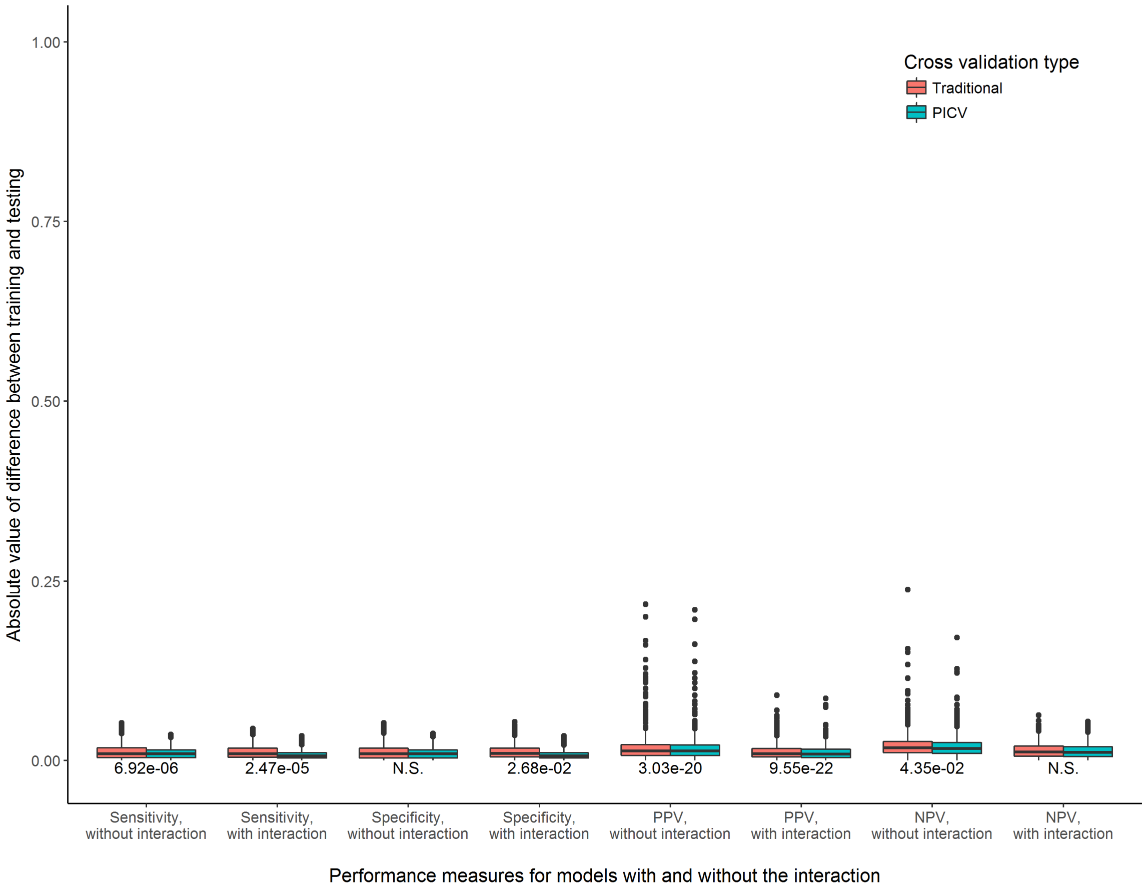
Supplemental Figure 14. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 14, prevalence = 0.5, n = 2000



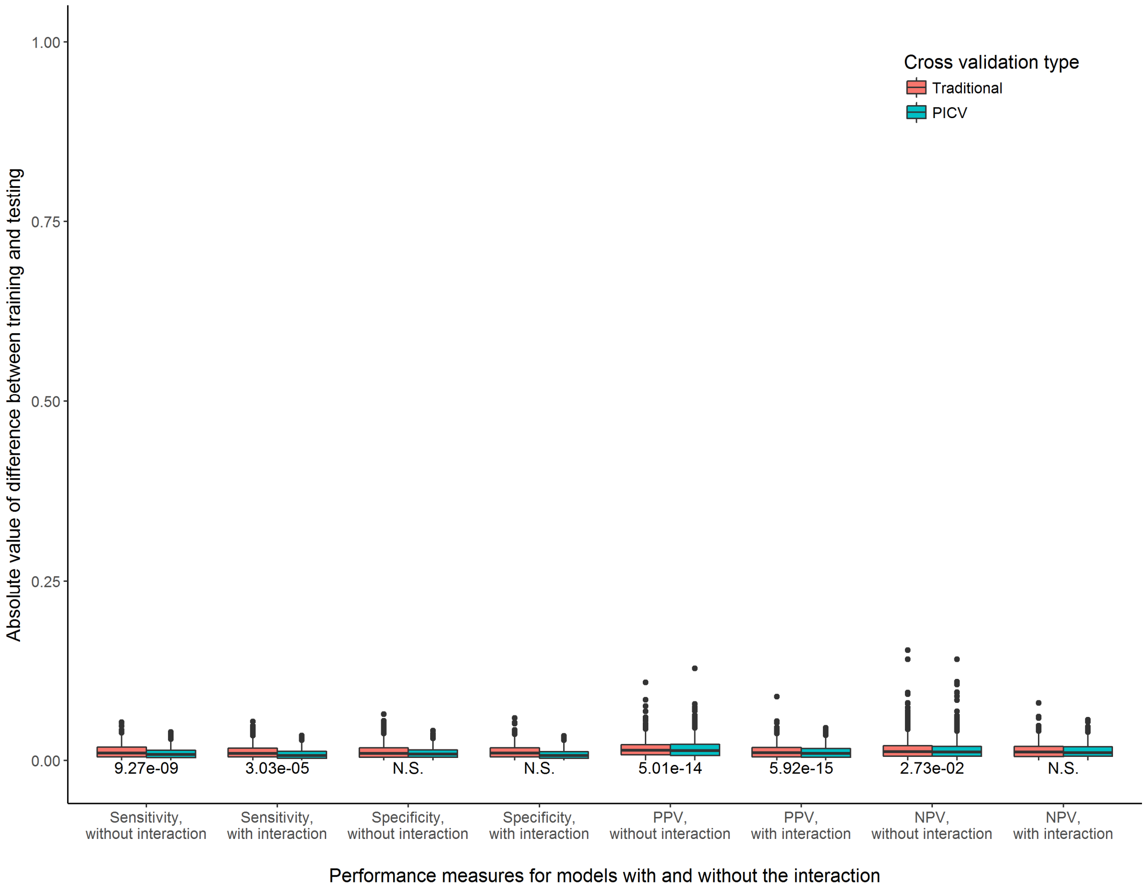
Supplemental Figure 15. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 15, prevalence = 0.5, n = 2000



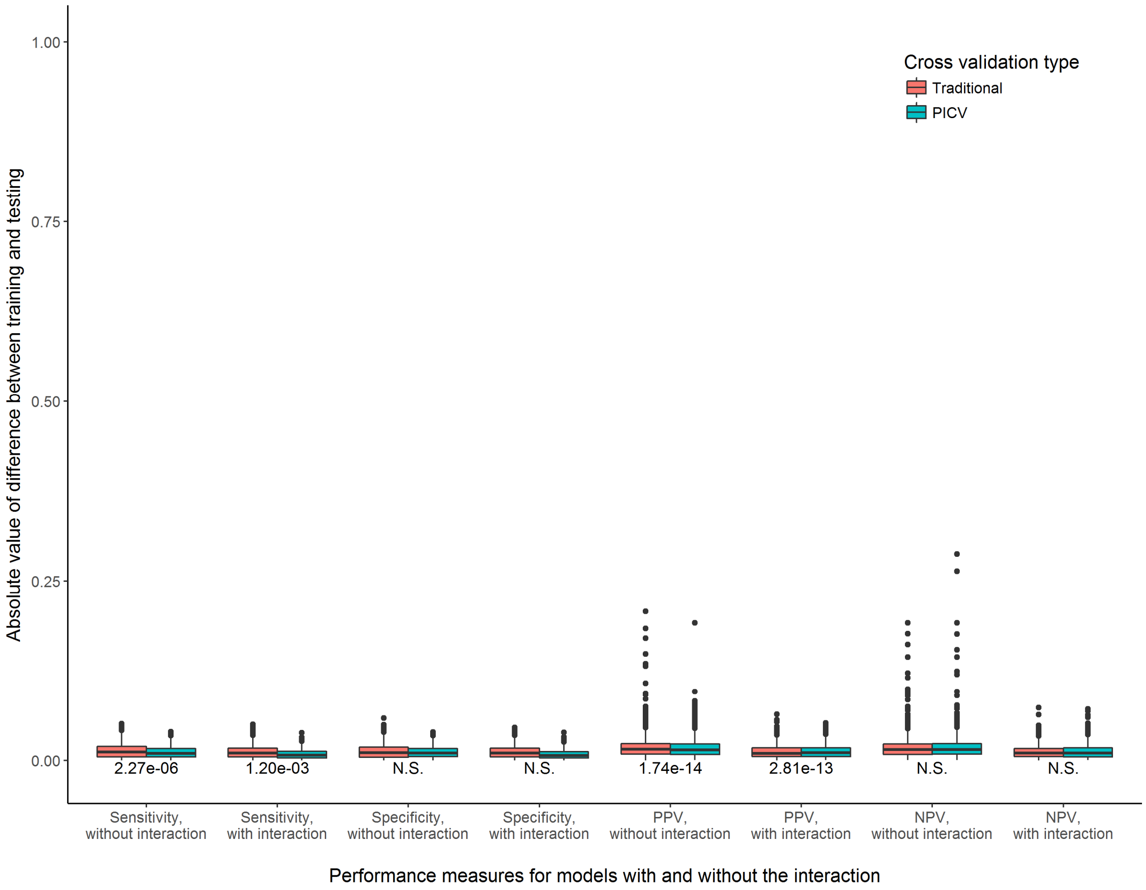
Supplemental Figure 16. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 1, prevalence = 0.5, n = 10000



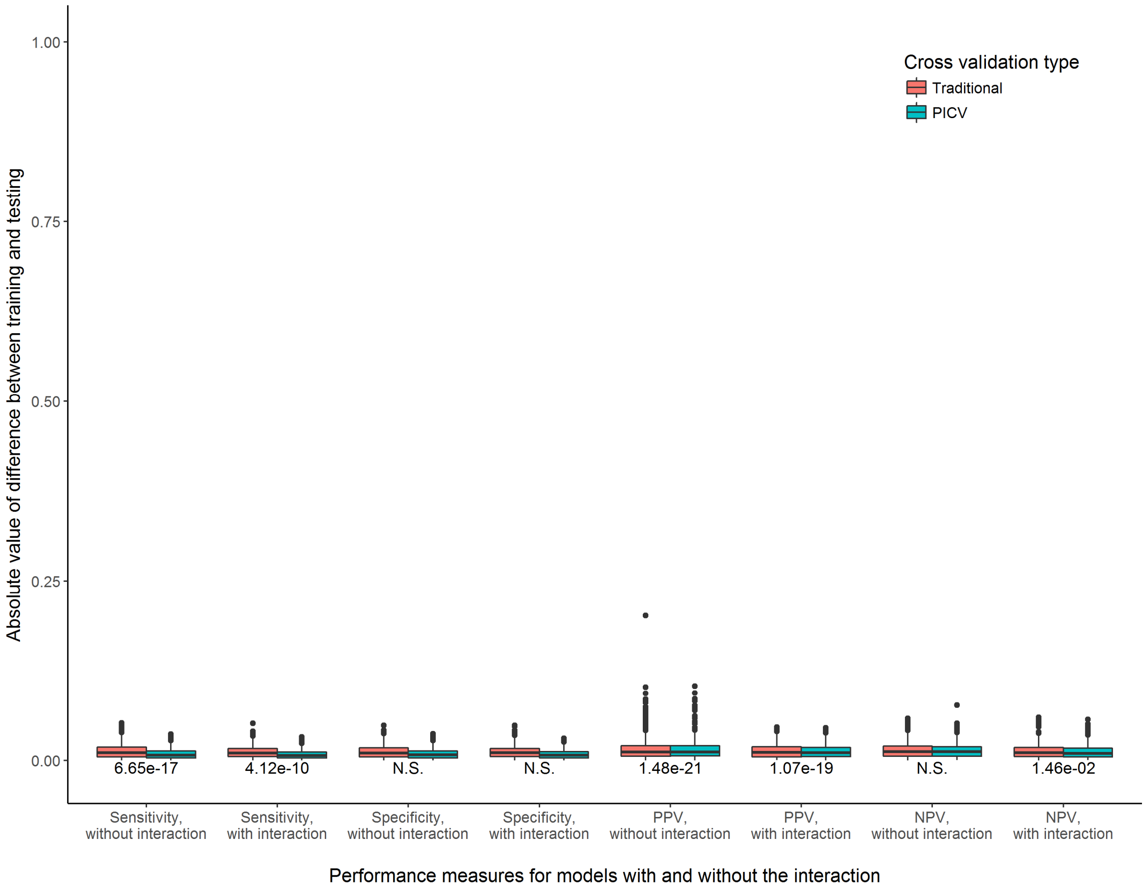
Supplemental Figure 17. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 2, prevalence = 0.5, n = 10000



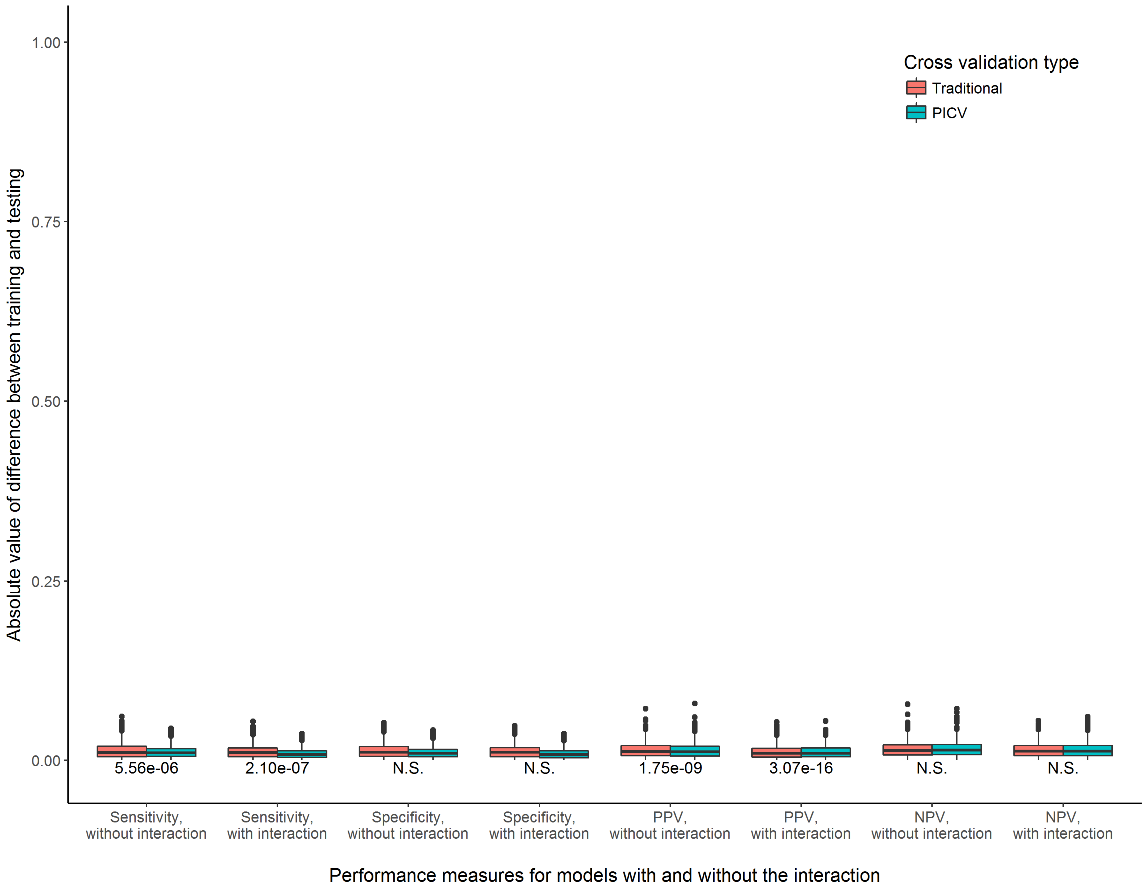
Supplemental Figure 18. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 3, prevalence = 0.5, n = 10000



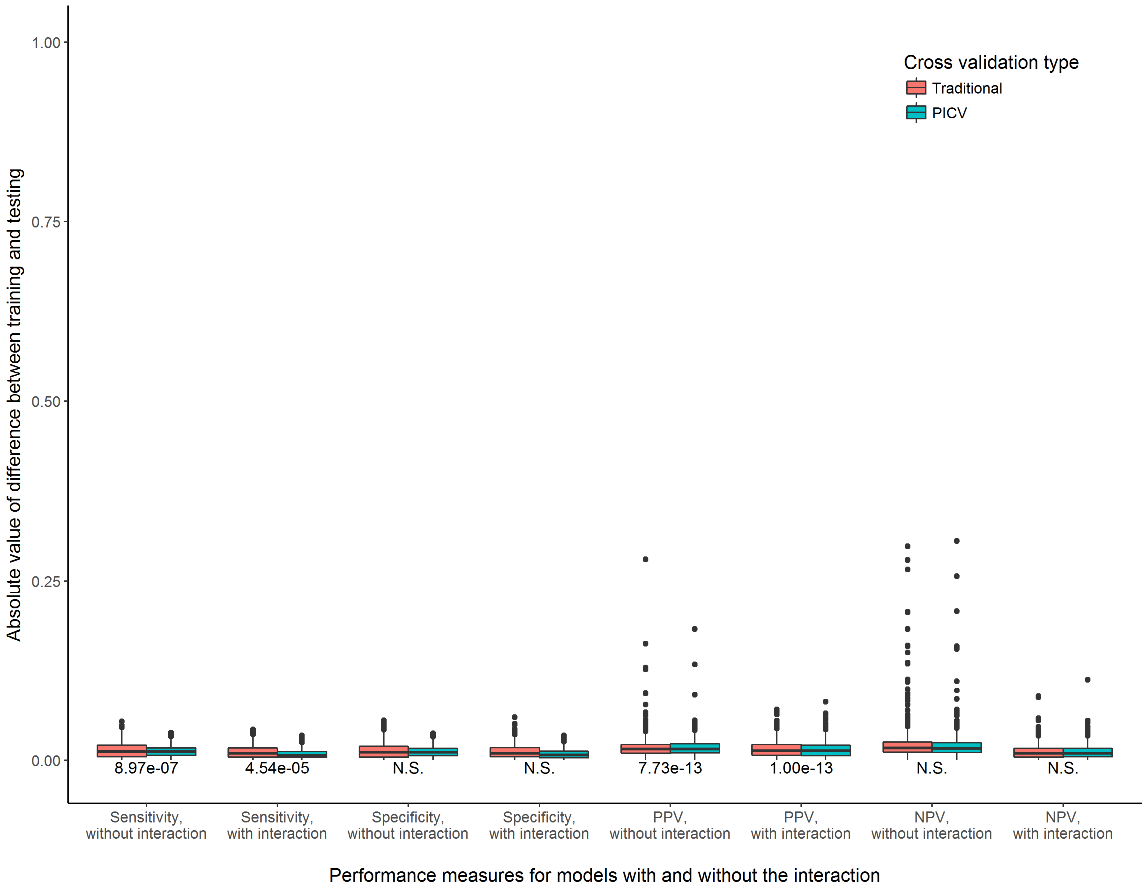
Supplemental Figure 19. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 4, prevalence = 0.5, n = 10000



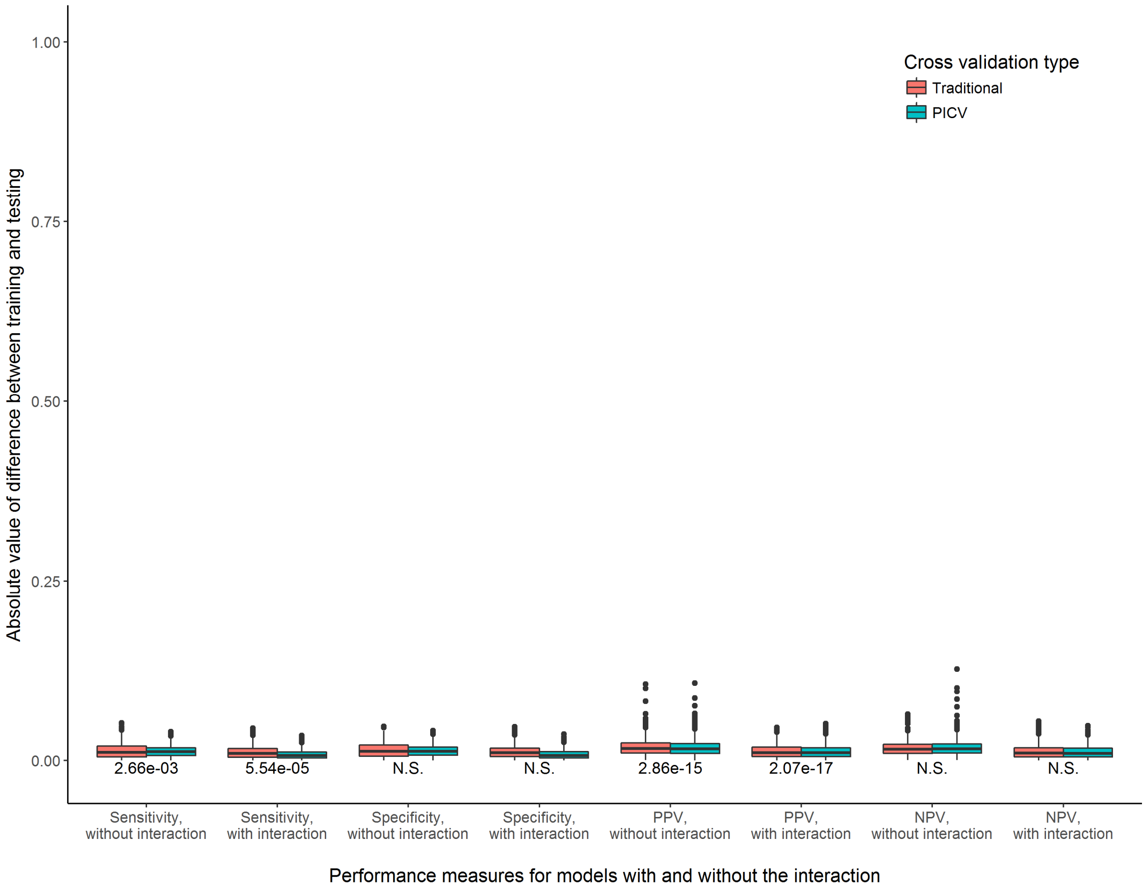
Supplemental Figure 20. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 5, prevalence = 0.5, n = 10000



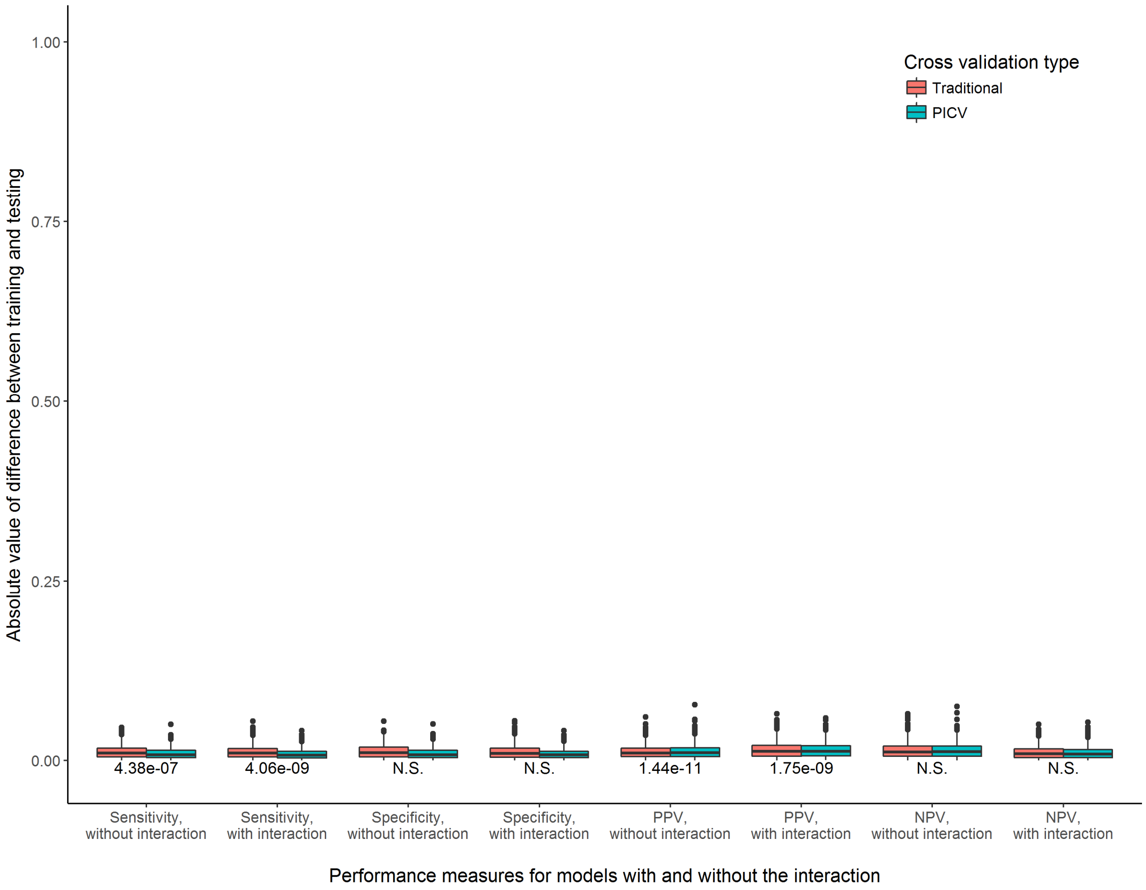
Supplemental Figure 21. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 6, prevalence = 0.5, n = 10000



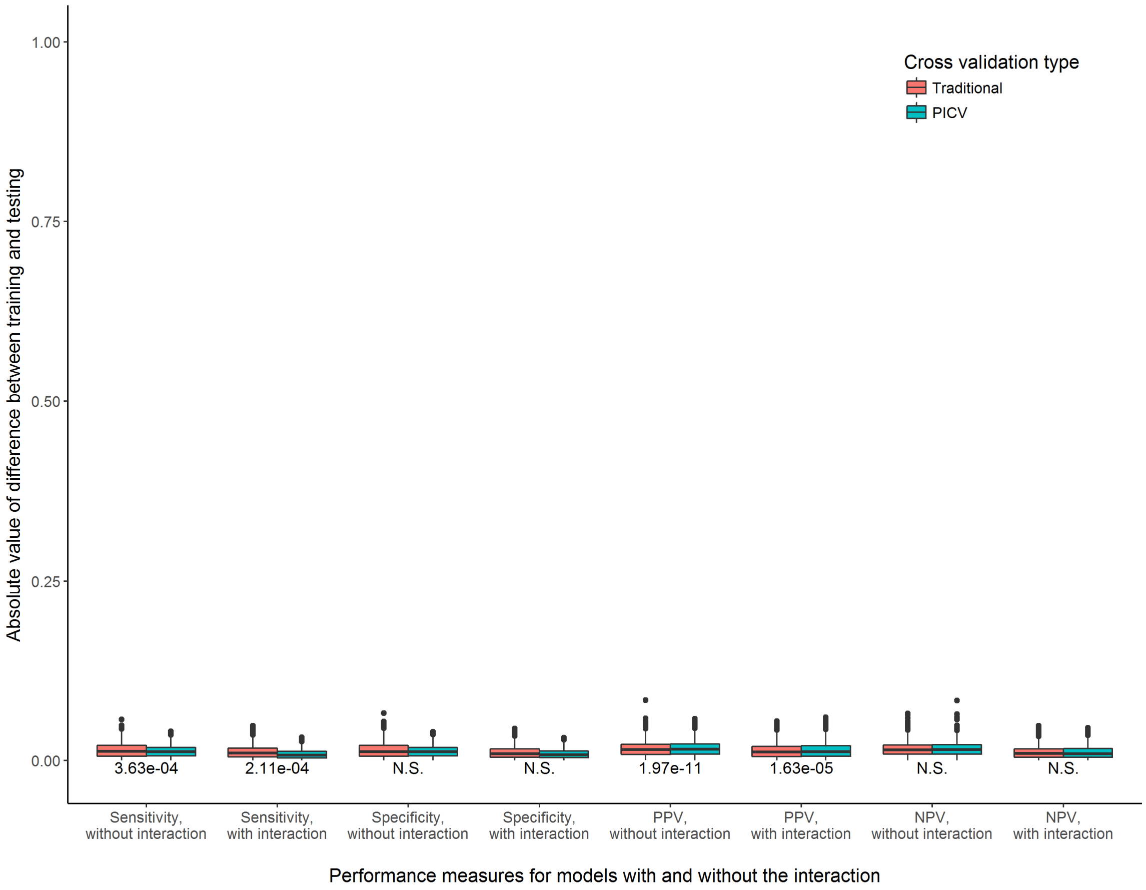
Supplemental Figure 22. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 7, prevalence = 0.5, n = 10000



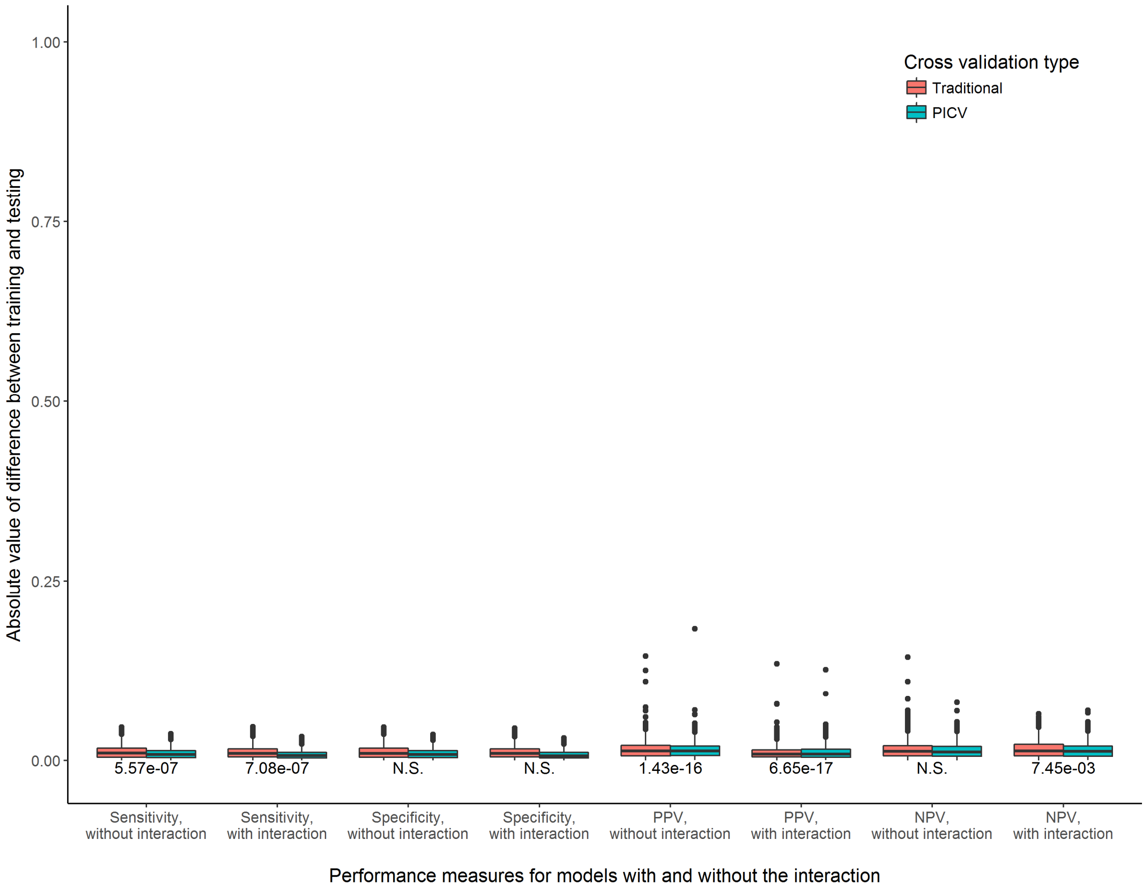
Supplemental Figure 23. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 8, prevalence = 0.5, n = 10000



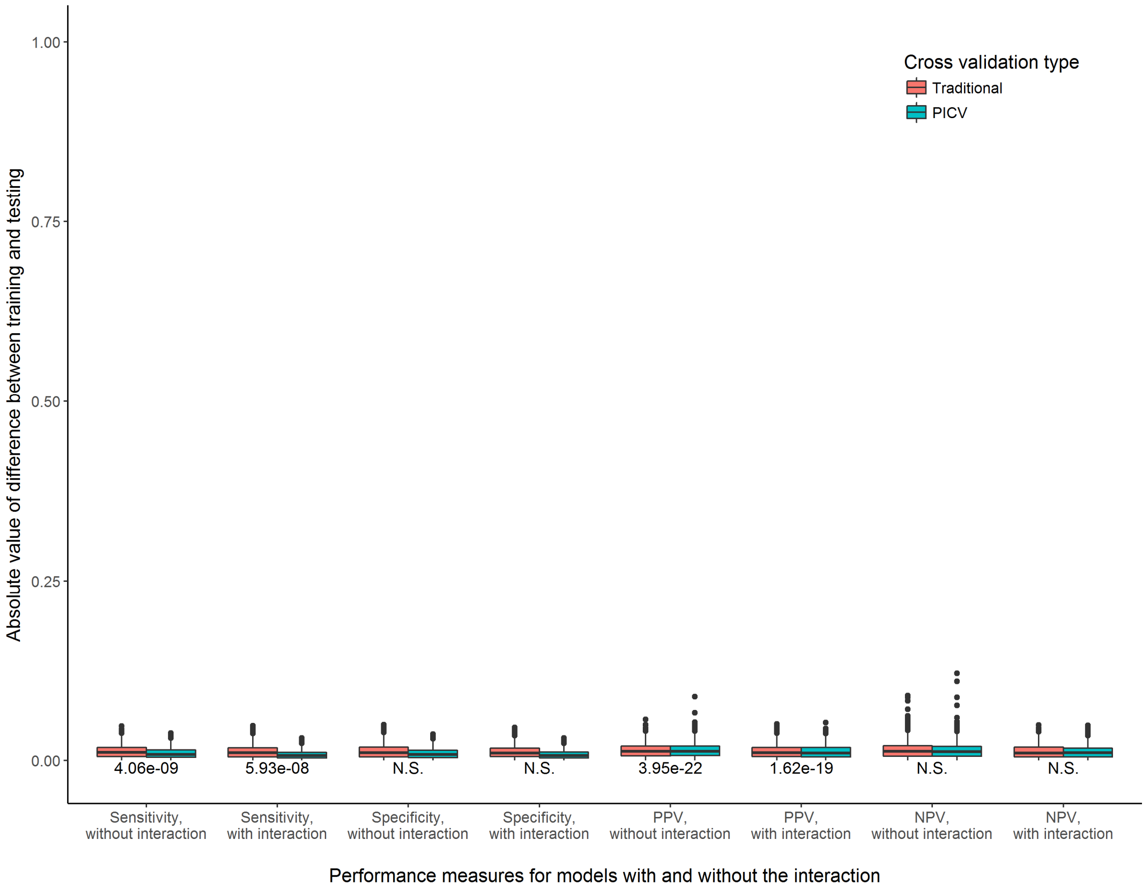
Supplemental Figure 24. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 9, prevalence = 0.5, n = 10000



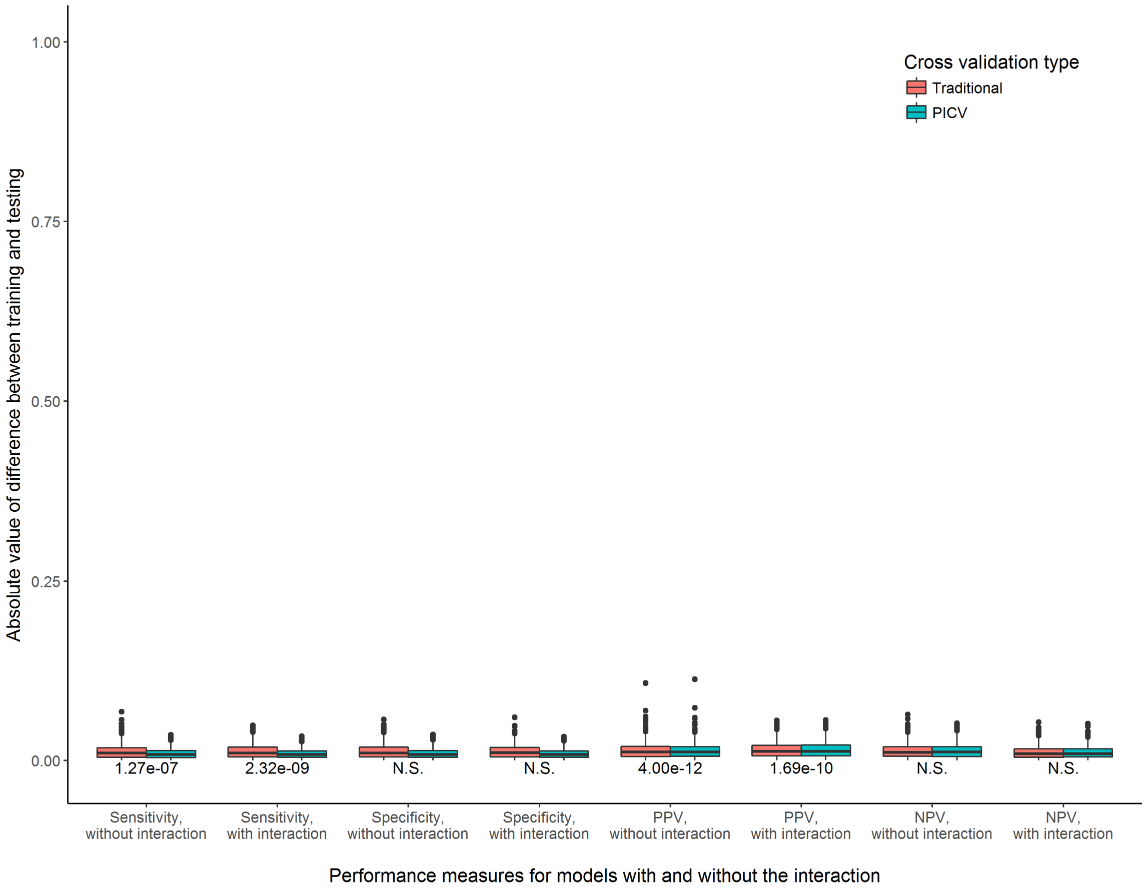
Supplemental Figure 25. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 10, prevalence = 0.5, n = 10000



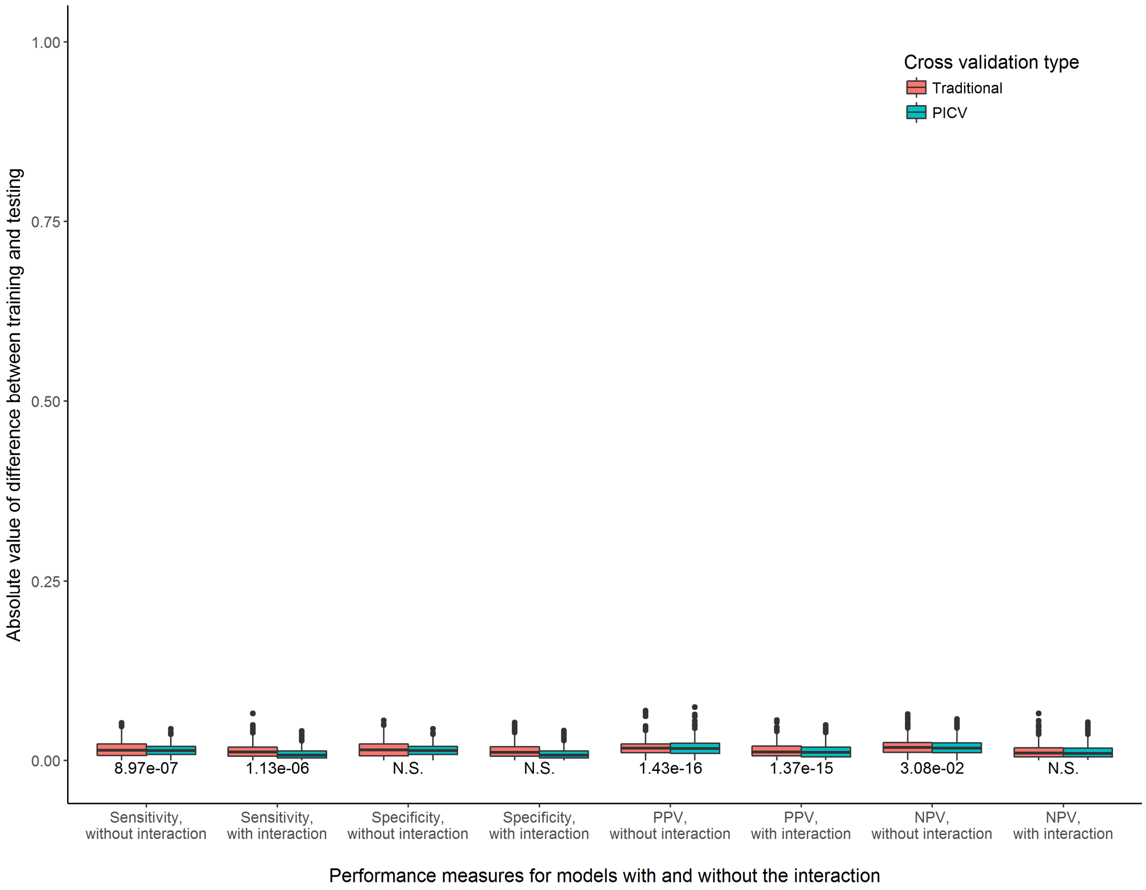
Supplemental Figure 26. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 11, prevalence = 0.5, n = 10000



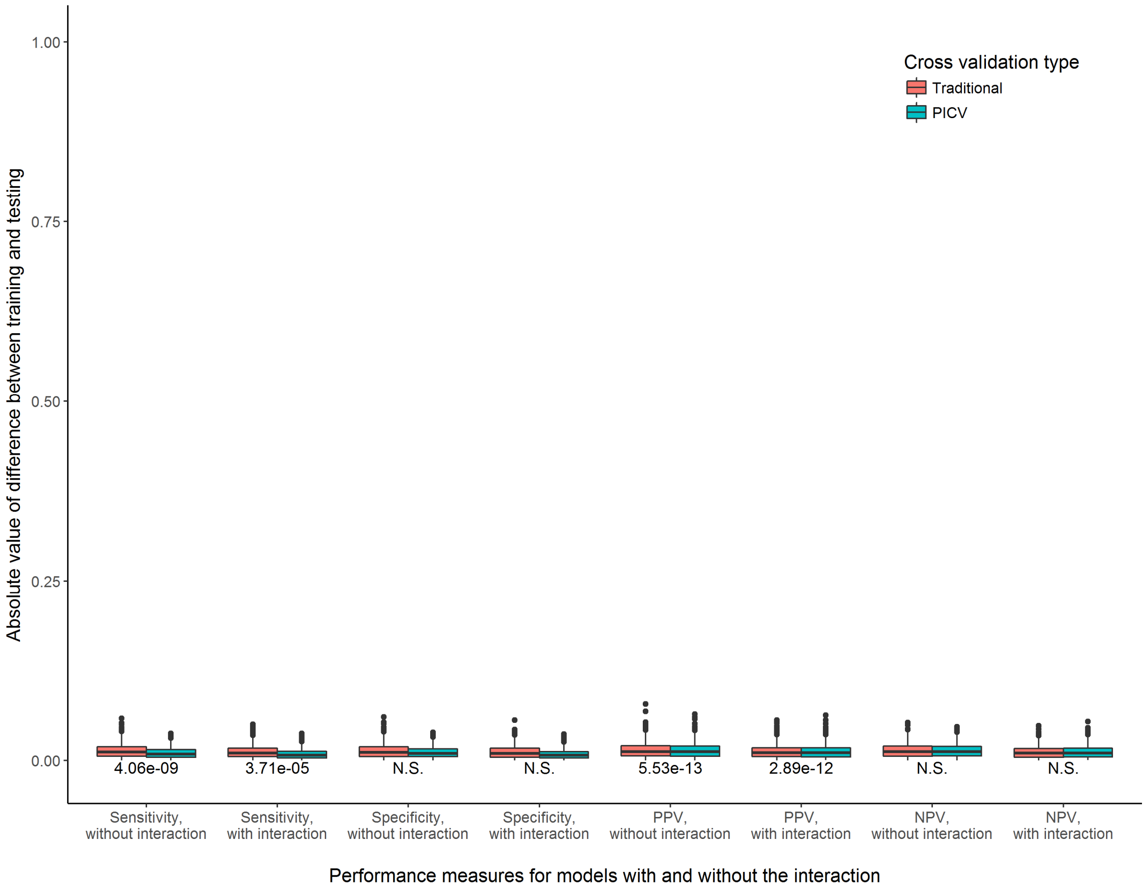
Supplemental Figure 27. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 12, prevalence = 0.5, n = 10000



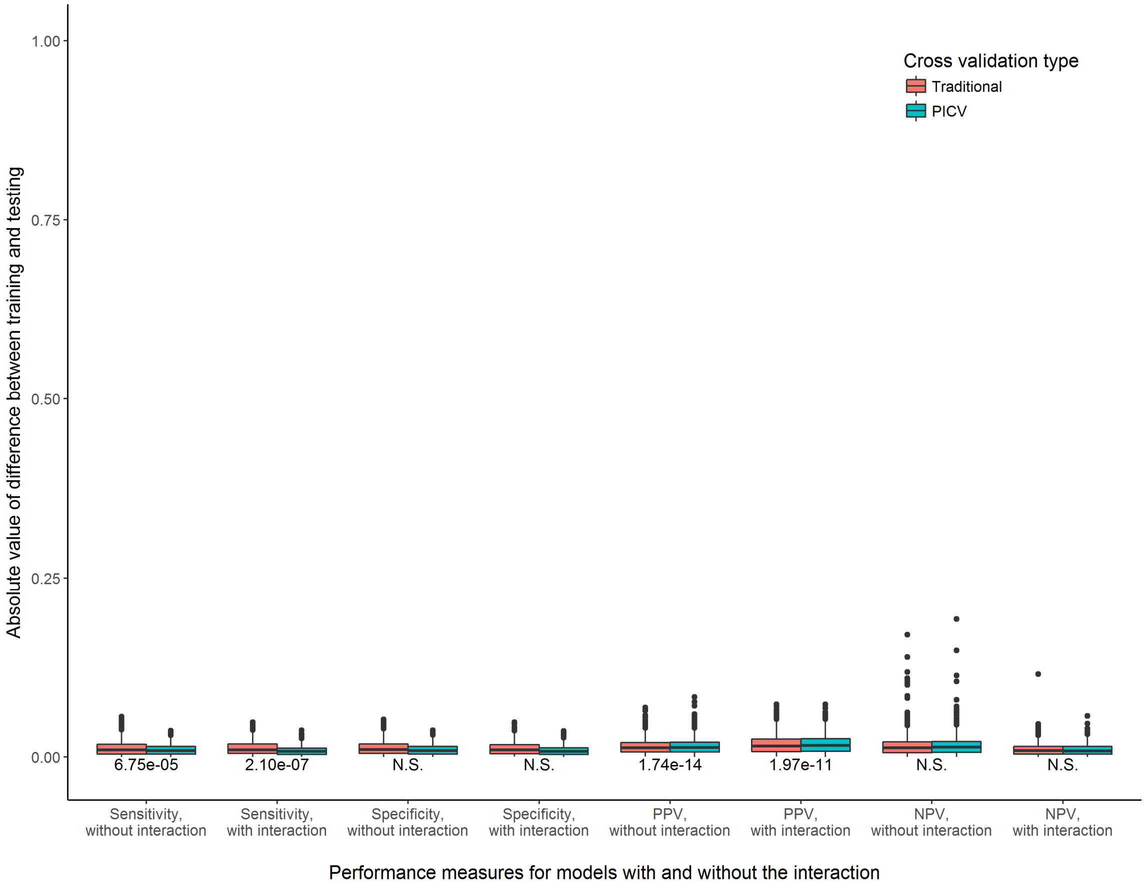
Supplemental Figure 28. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 13, prevalence = 0.5, n = 10000



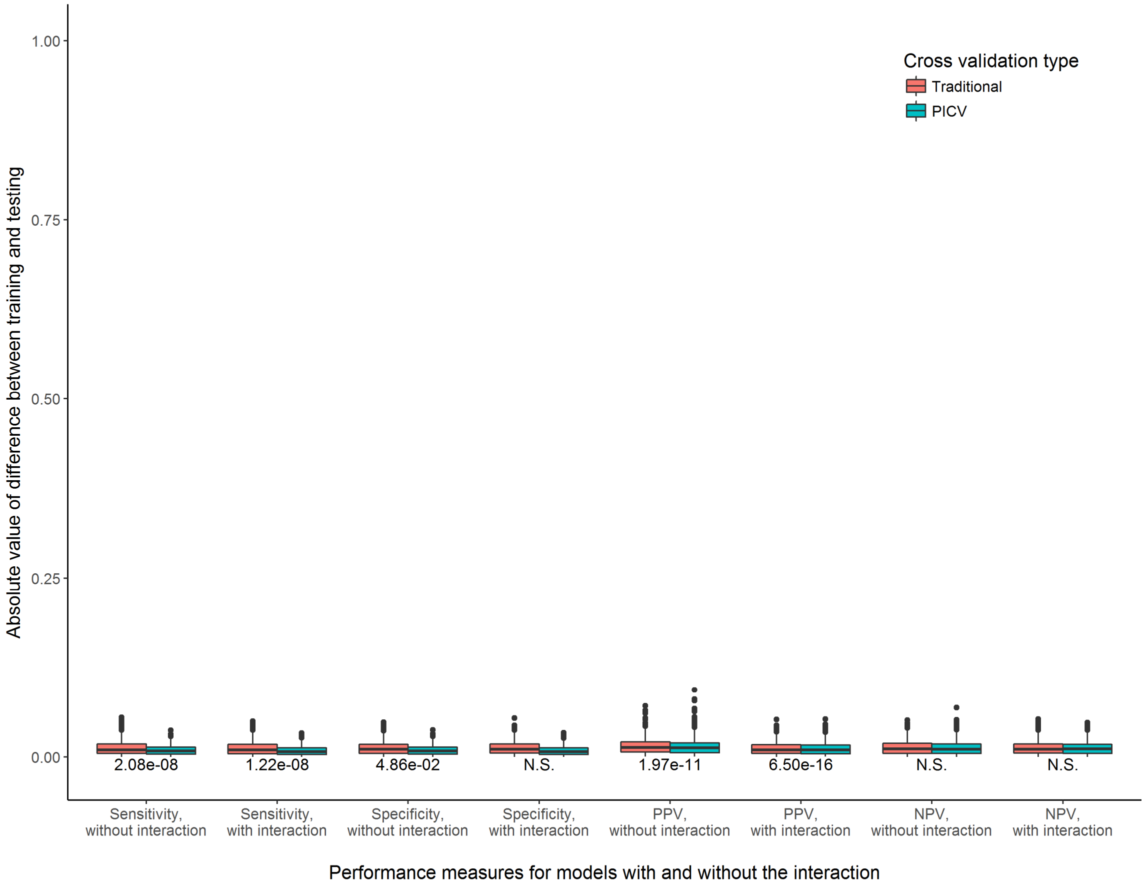
Supplemental Figure 29. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 14, prevalence = 0.5, n = 10000



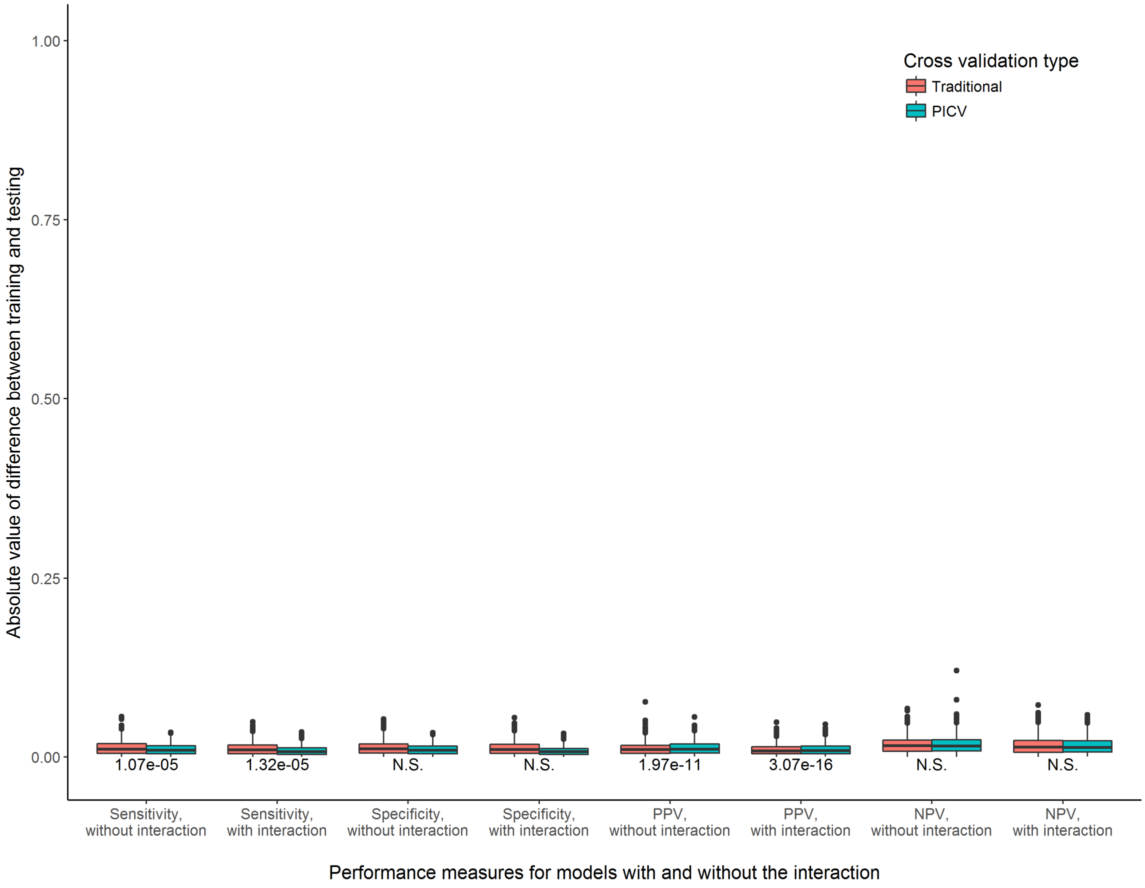
Supplemental Figure 30. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 15, prevalence = 0.5, n = 10000



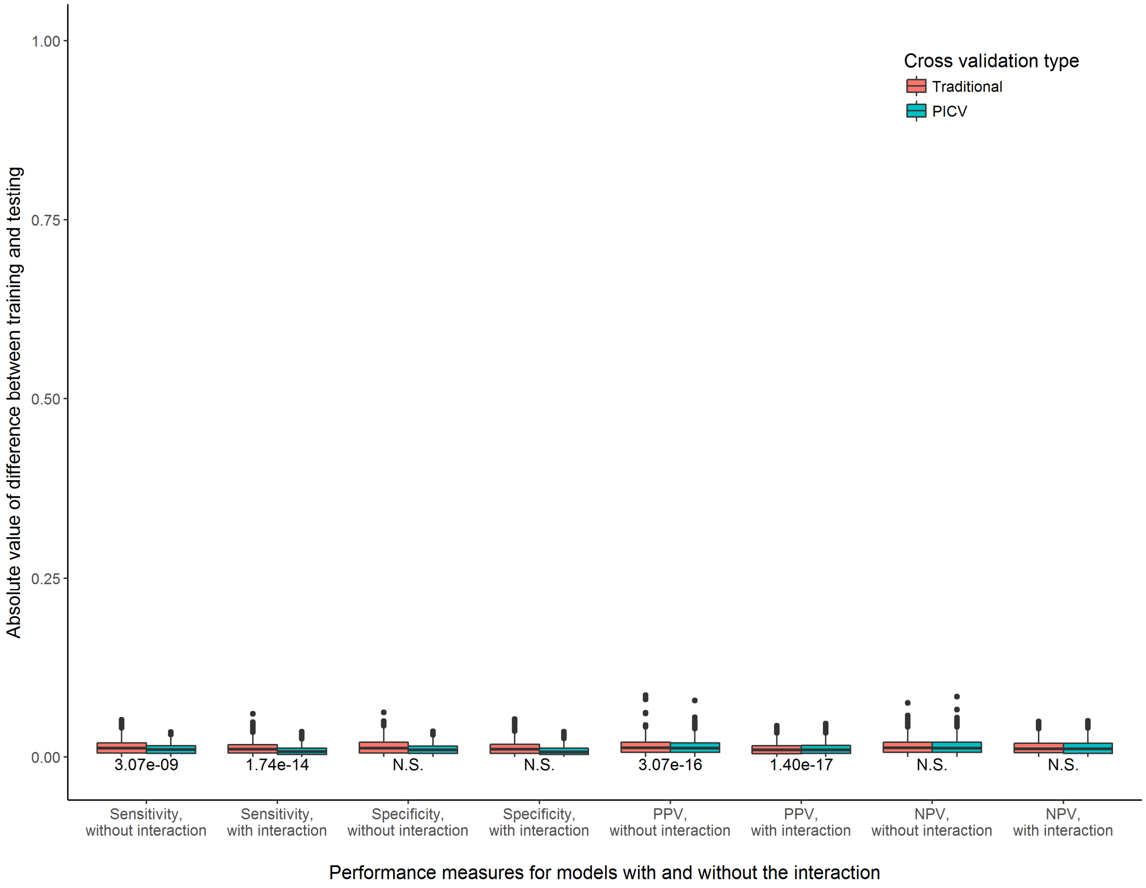
Supplemental Figure 31. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 1, prevalence = 0.1, n = 10000



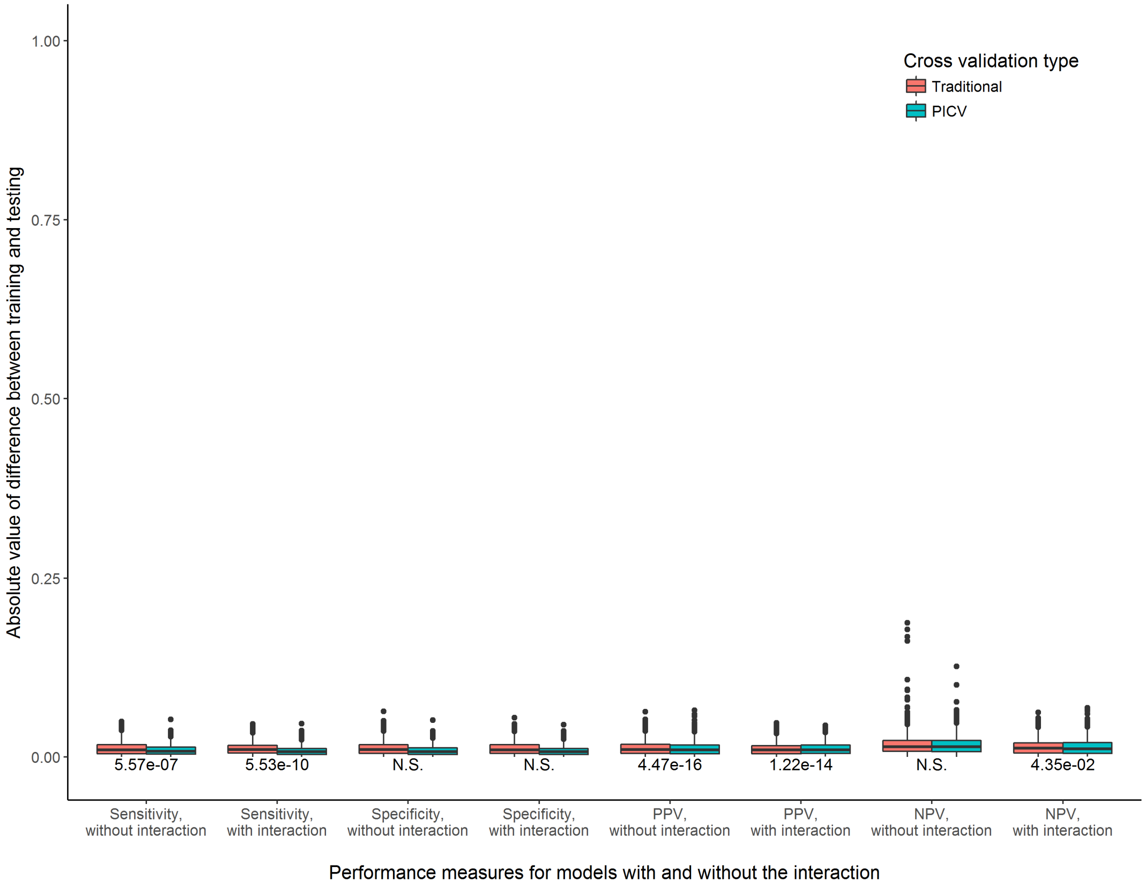
Supplemental Figure 32. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 2, prevalence = 0.1, n = 10000



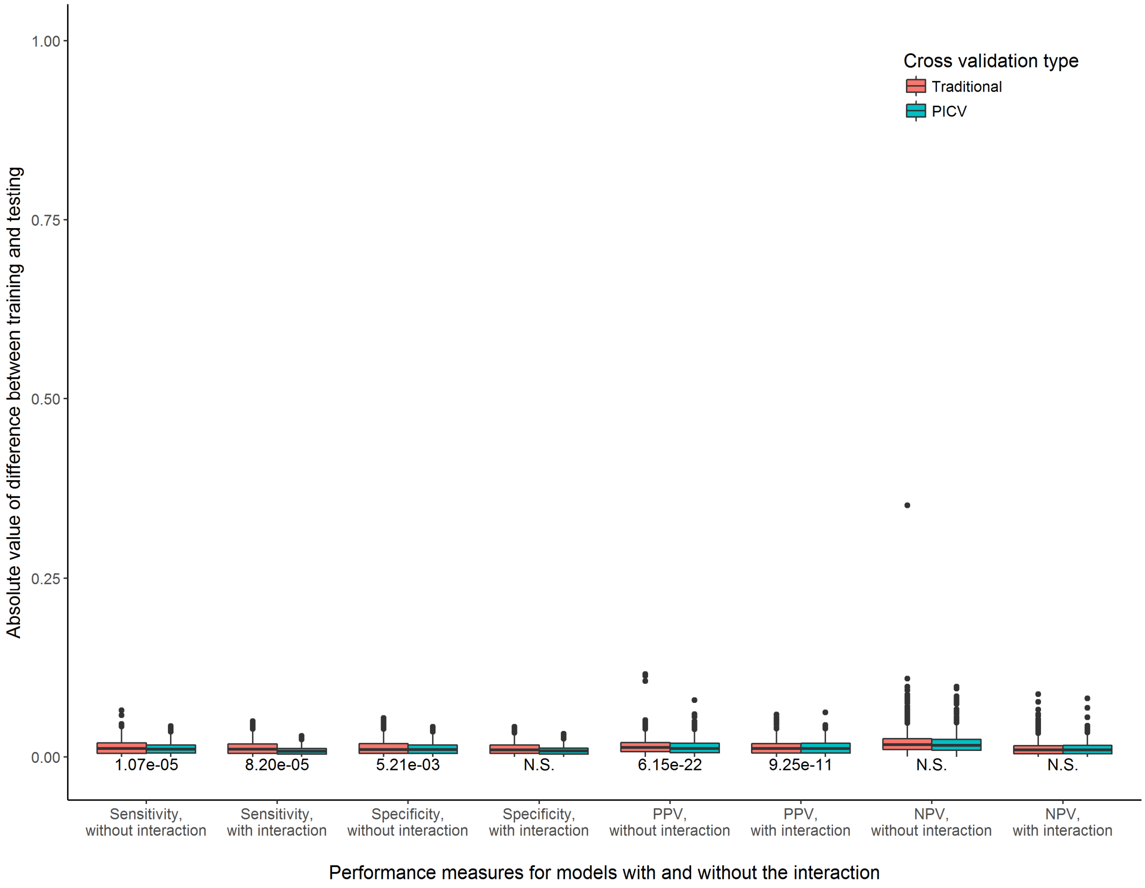
Supplemental Figure 33. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 3, prevalence = 0.1, n = 10000



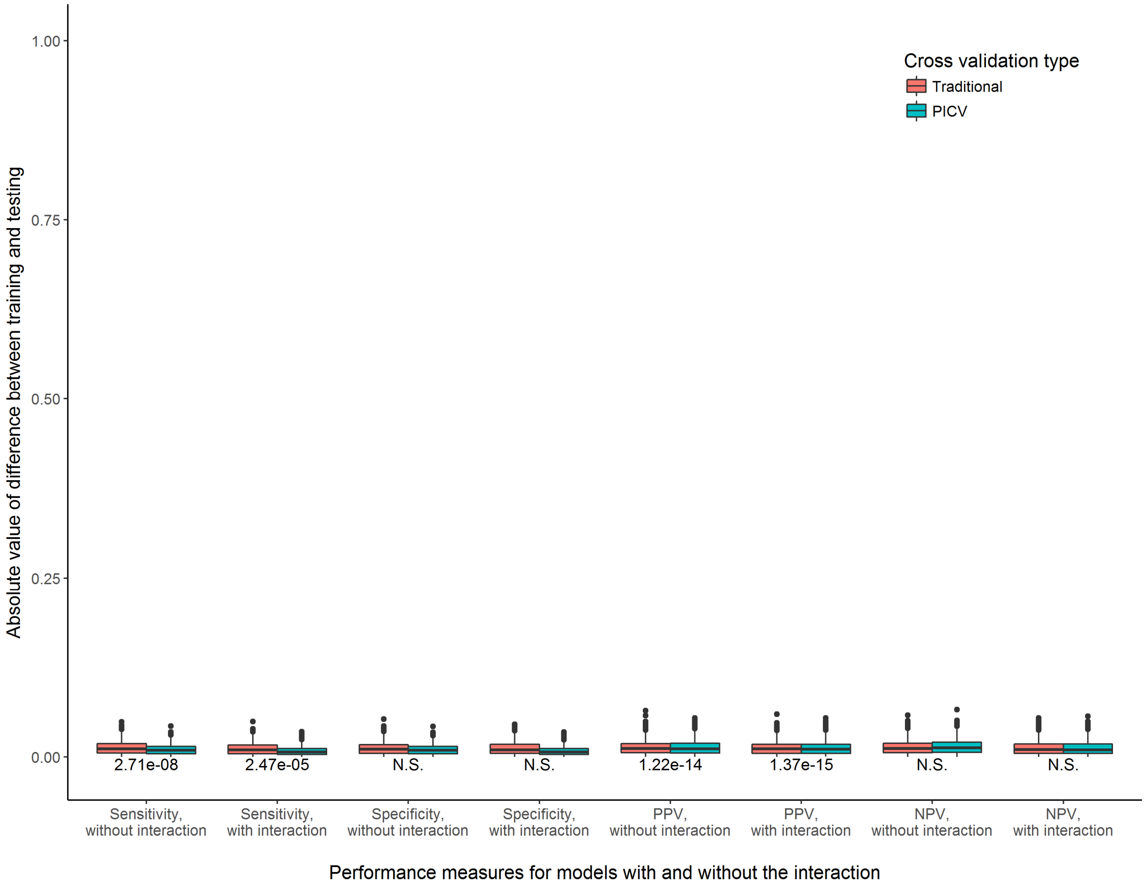
Supplemental Figure 34. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 4, prevalence = 0.1, n = 10000



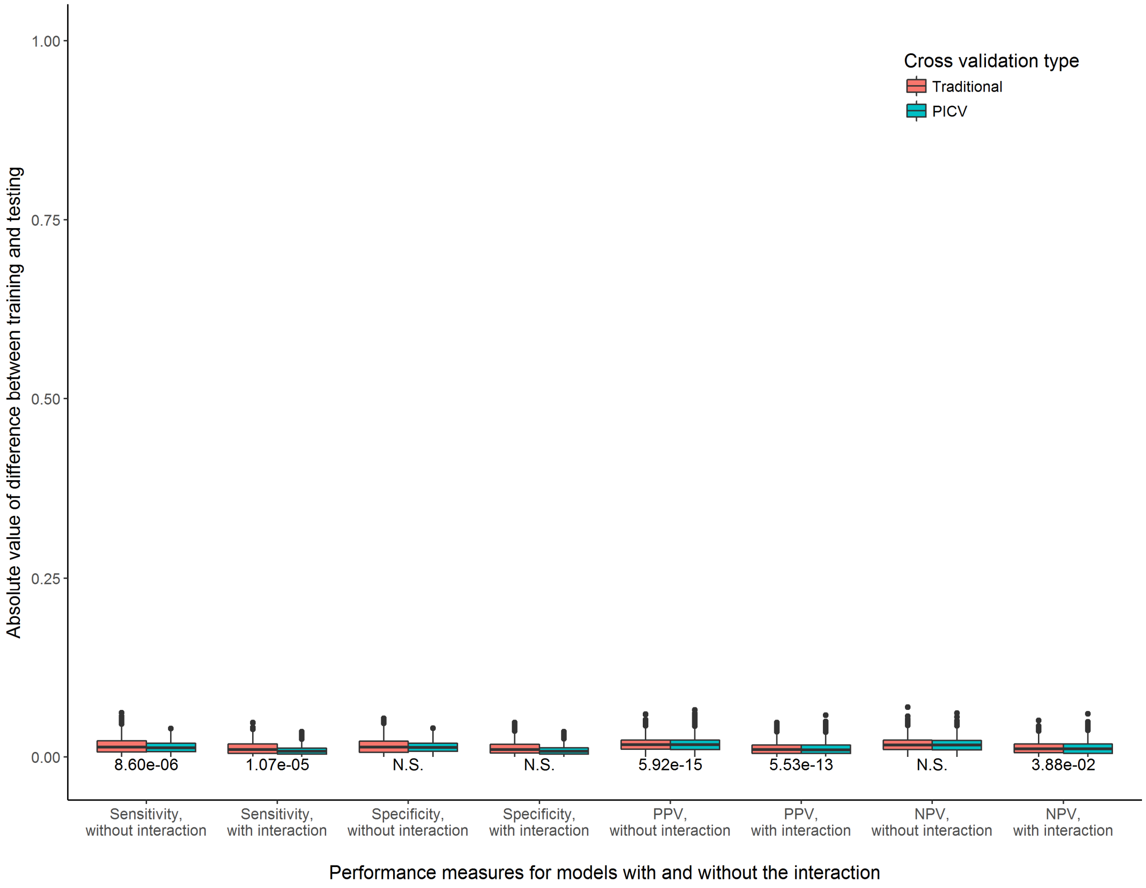
Supplemental Figure 35. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 5, prevalence = 0.1, n = 10000



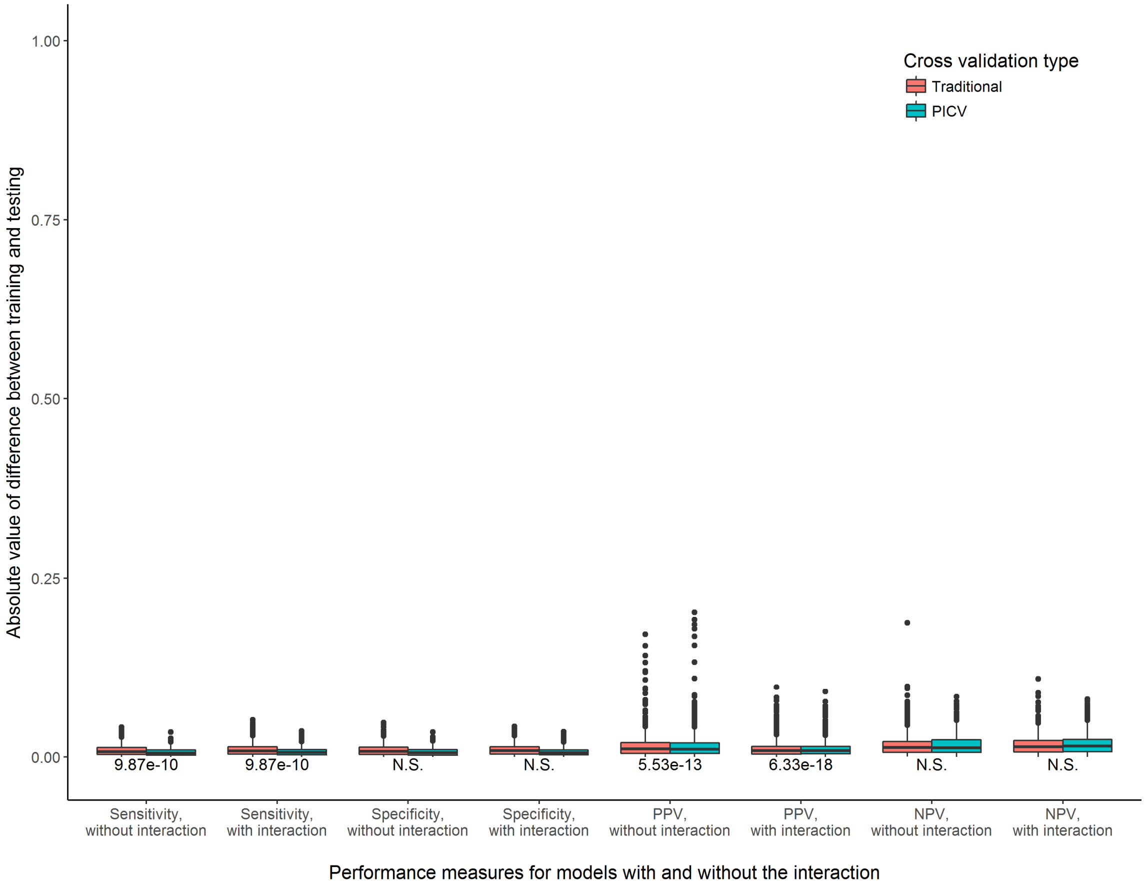
Supplemental Figure 36. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 6, prevalence = 0.1, n = 10000



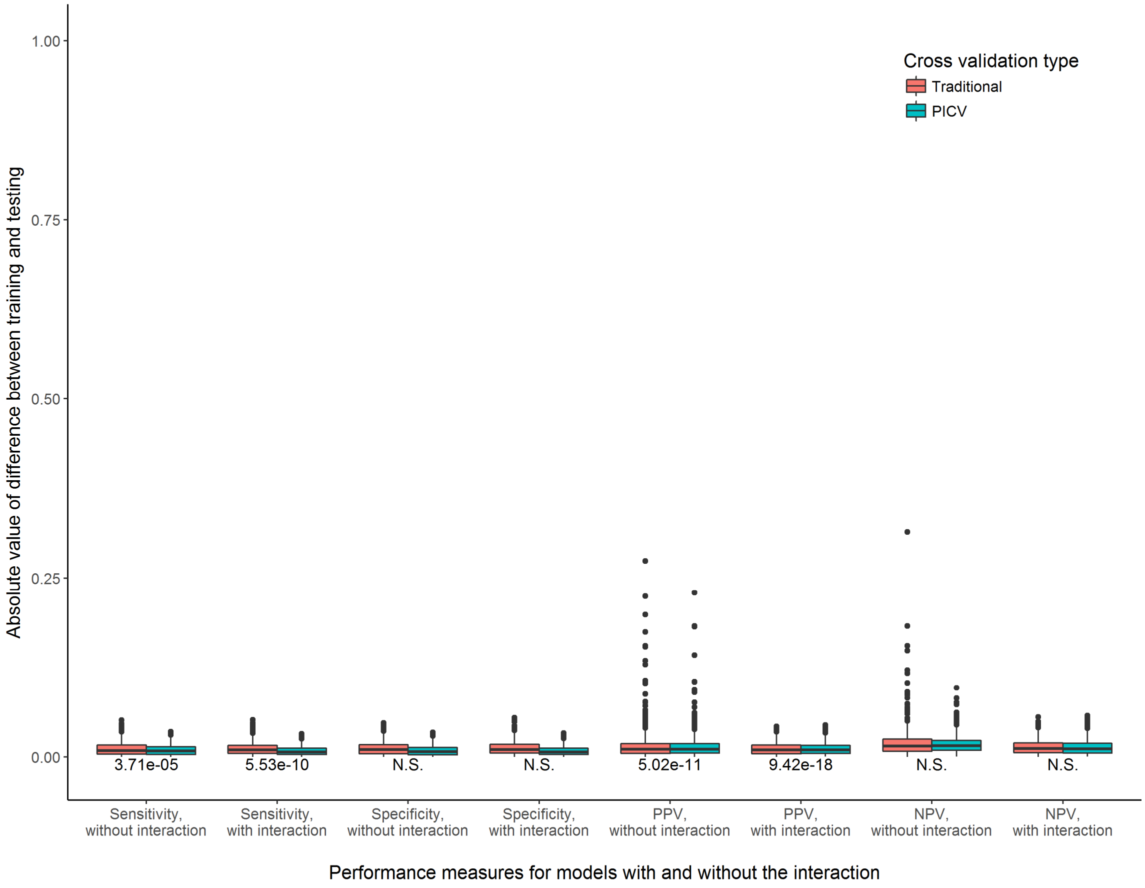
Supplemental Figure 37. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 7, prevalence = 0.1, n = 10000



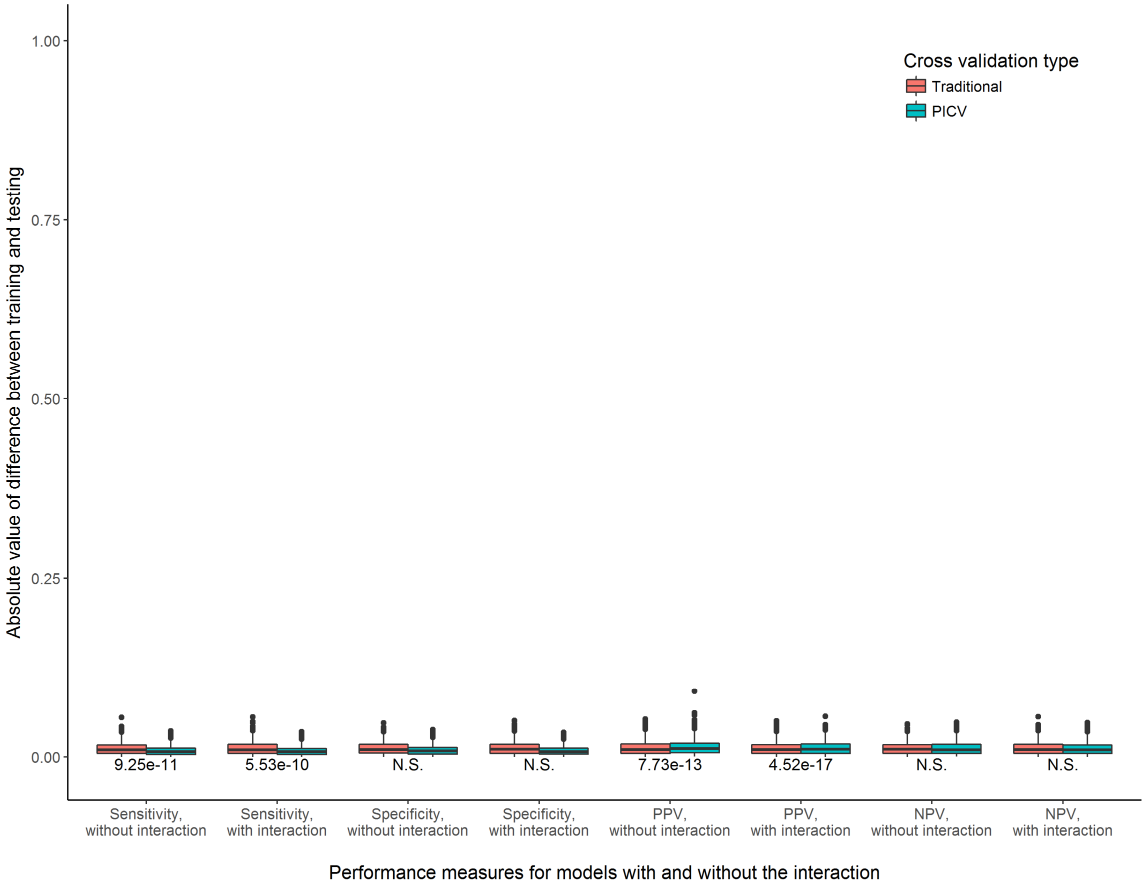
Supplemental Figure 38. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 8, prevalence = 0.1, n = 10000



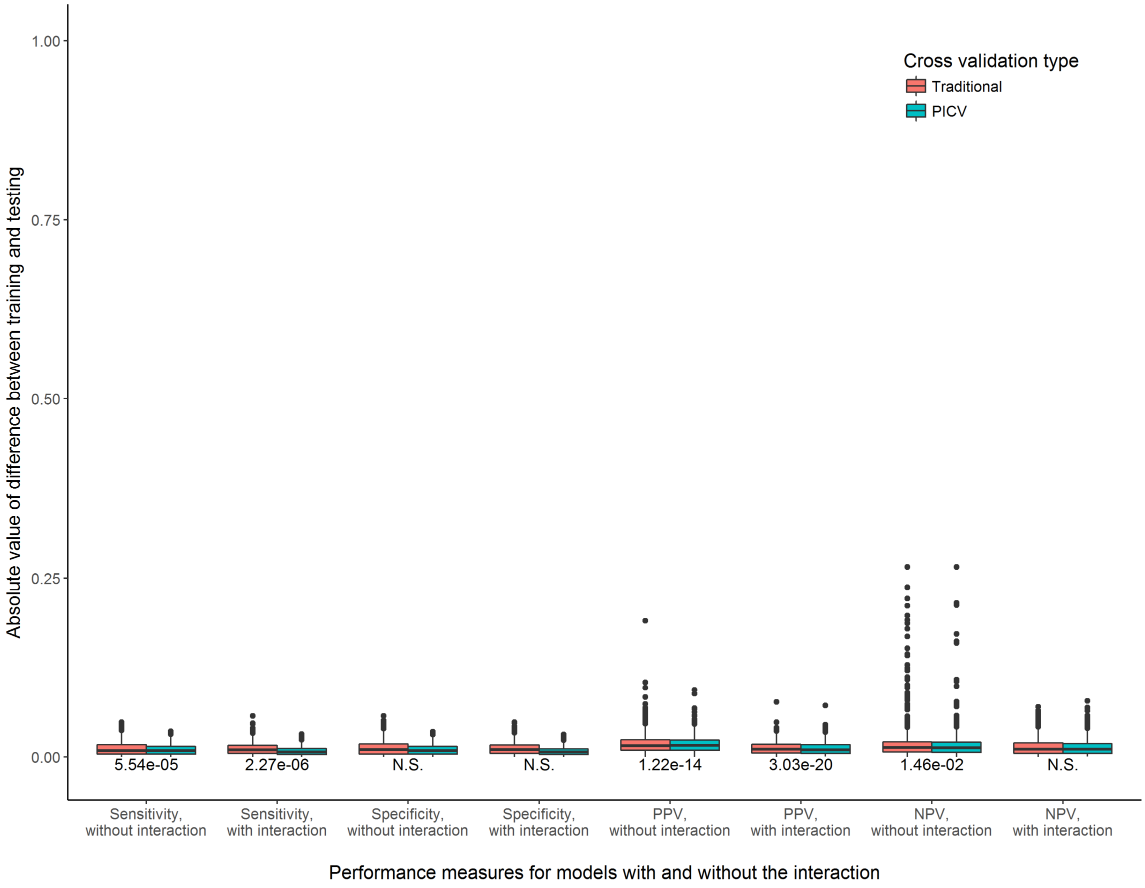
Supplemental Figure 39. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 9, prevalence = 0.1, n = 10000



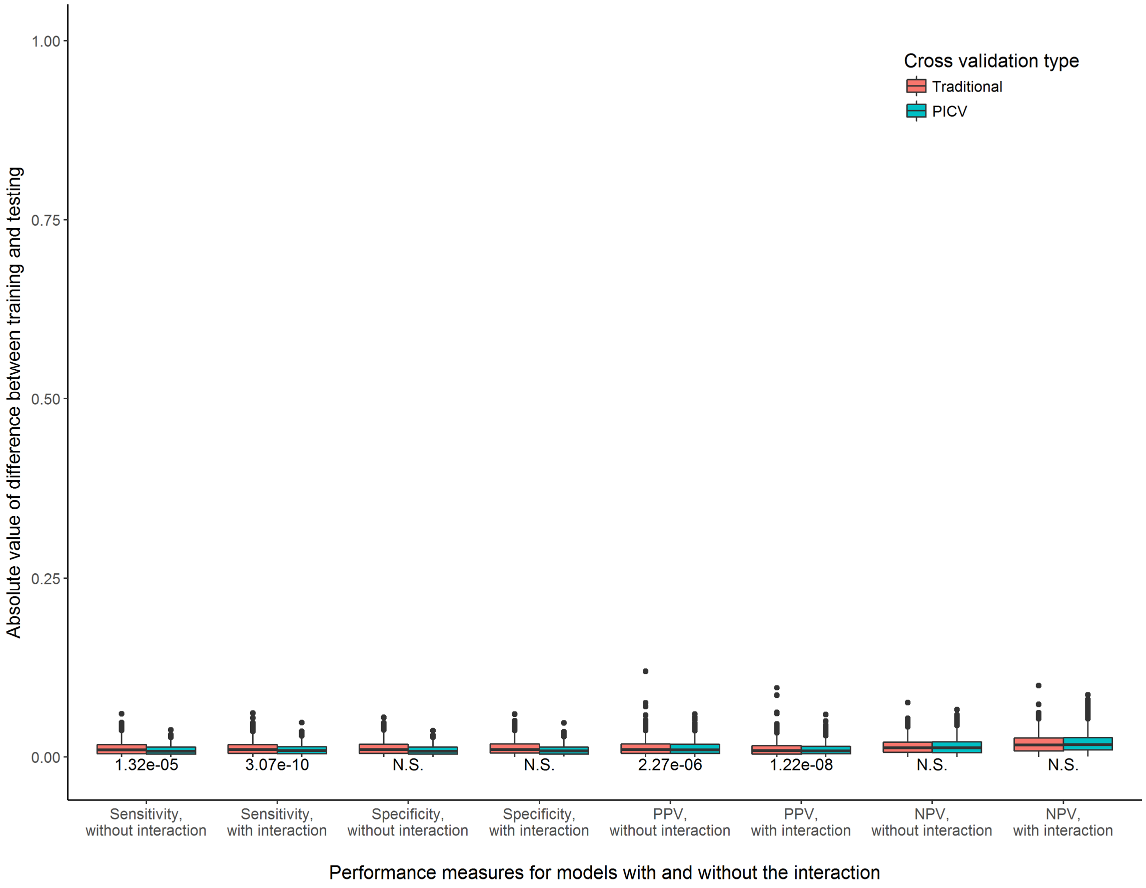
Supplemental Figure 40. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 10, prevalence = 0.1, n = 10000



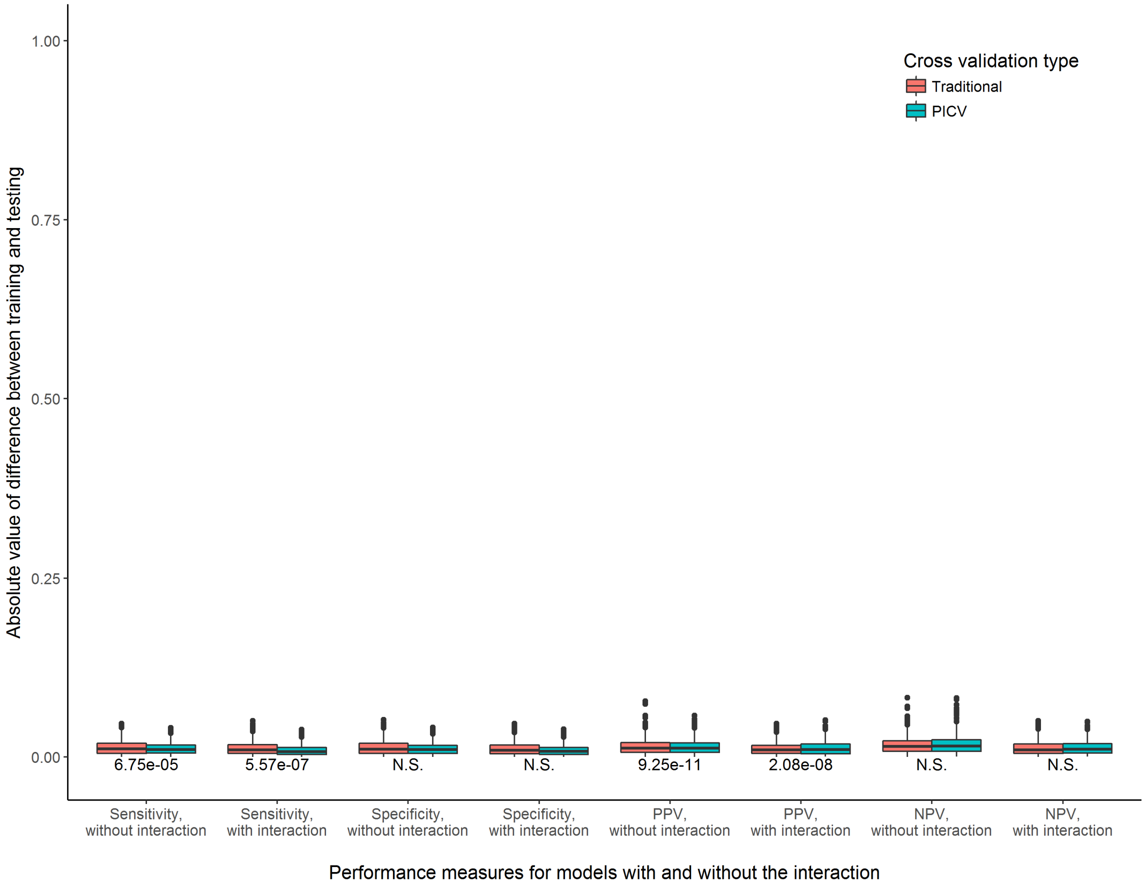
Supplemental Figure 41. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 11, prevalence = 0.1, n = 10000



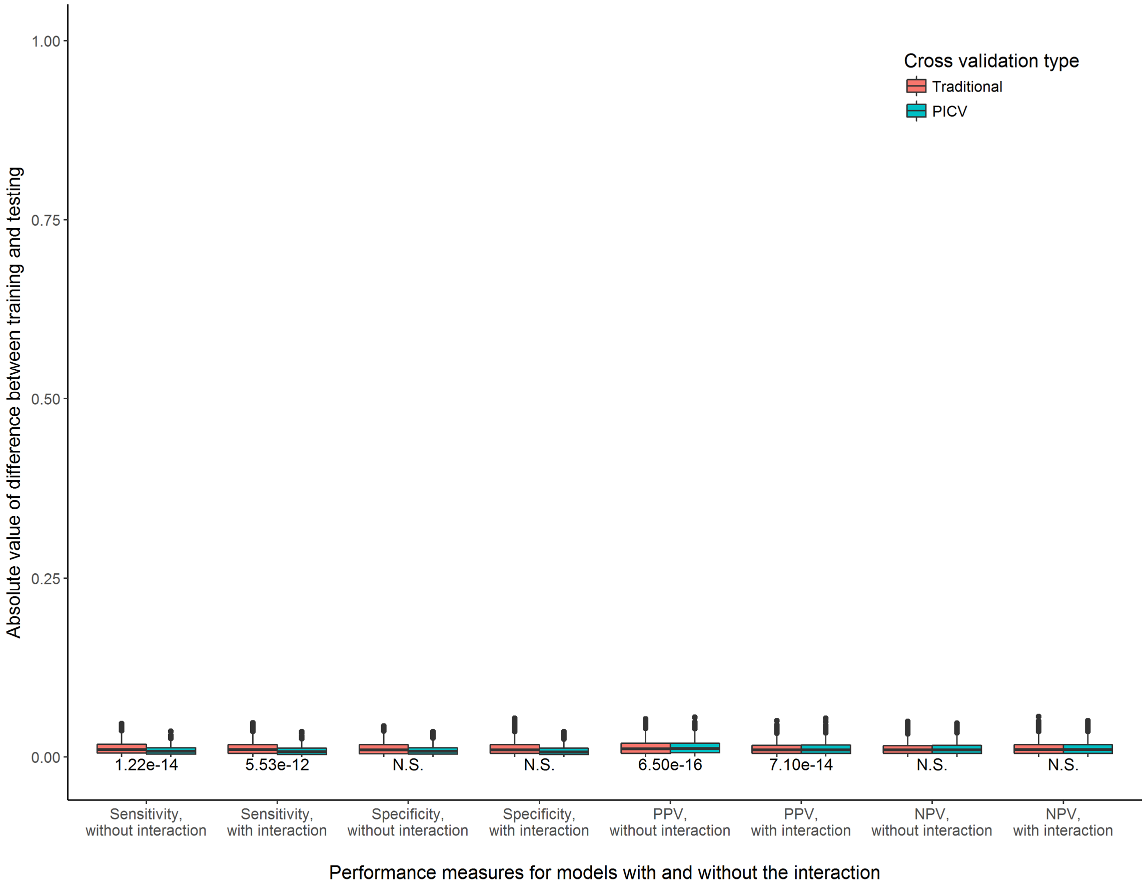
Supplemental Figure 42. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 12, prevalence = 0.1, n = 10000



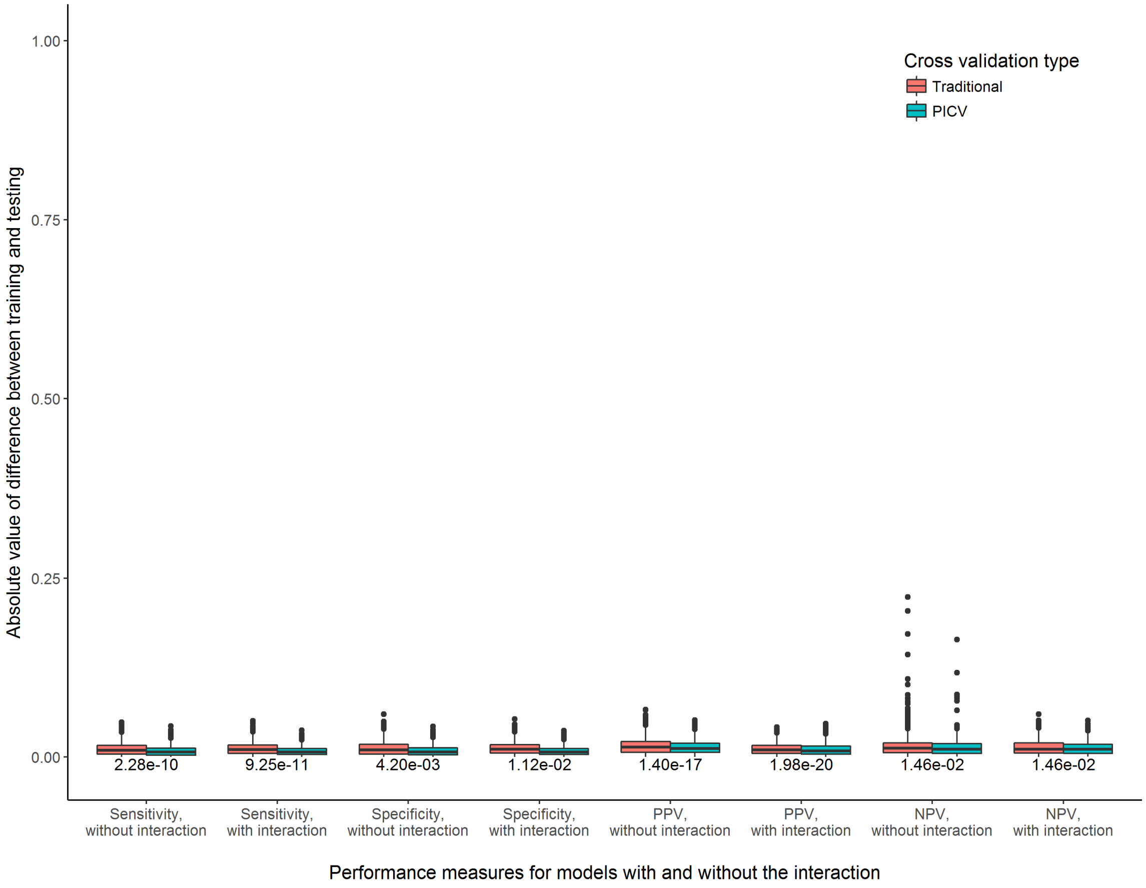
Supplemental Figure 43. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 13, prevalence = 0.1, n = 10000



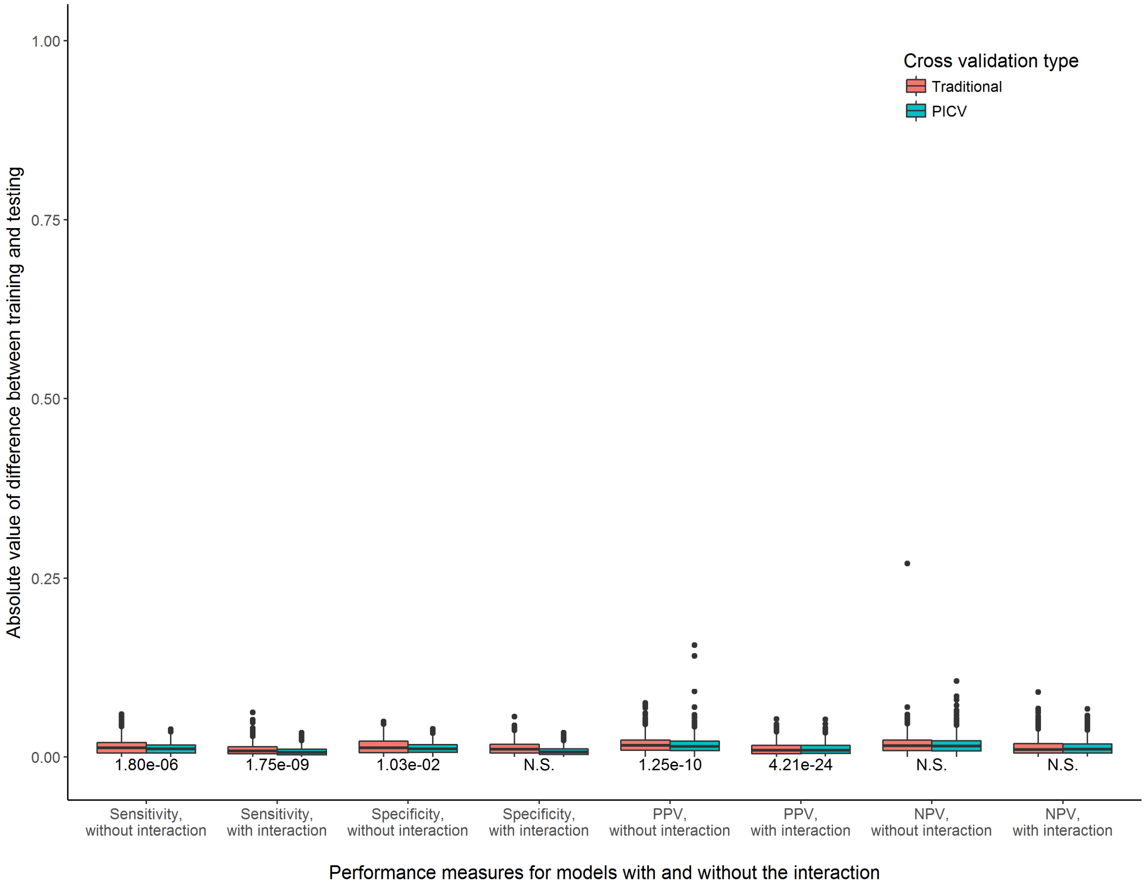
Supplemental Figure 44. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 14, prevalence = 0.1, n = 10000



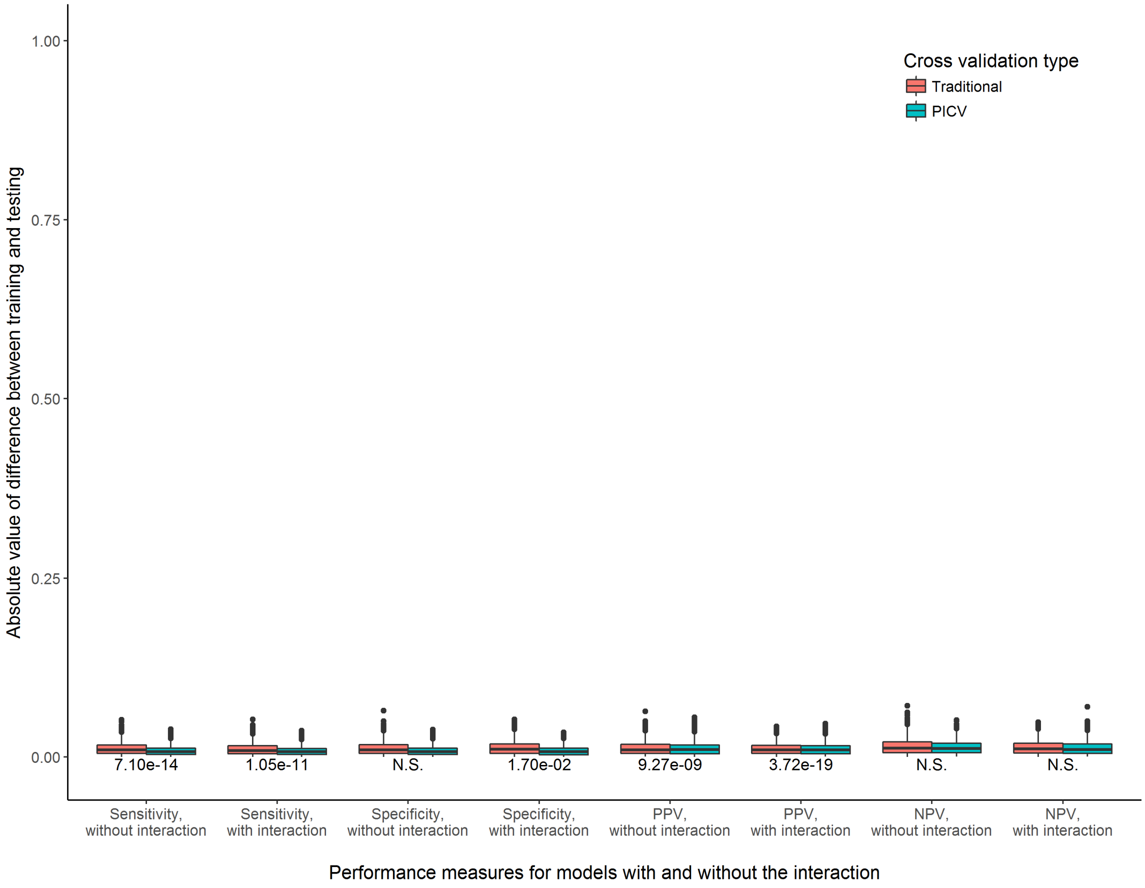
Supplemental Figure 45. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 15, prevalence = 0.1, n = 10000



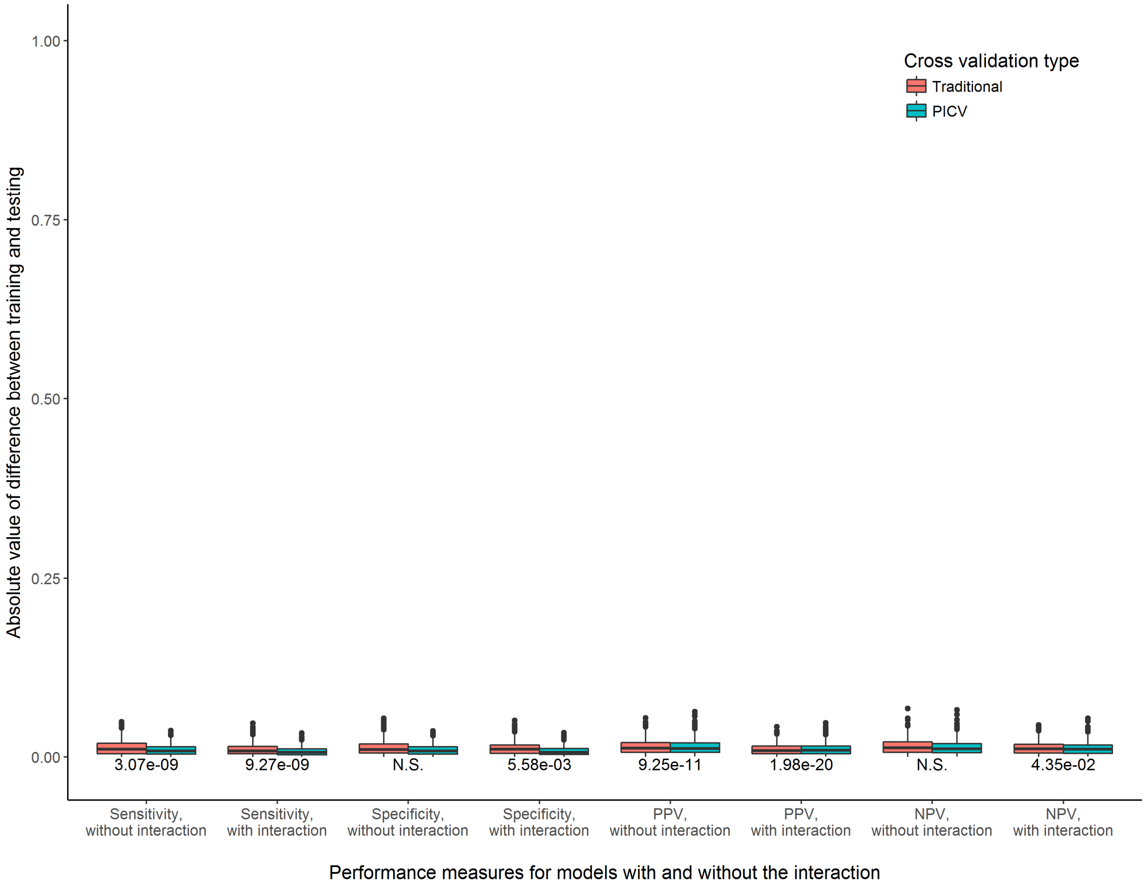
Supplemental Figure 46. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 1, prevalence = 0.02, n = 10000



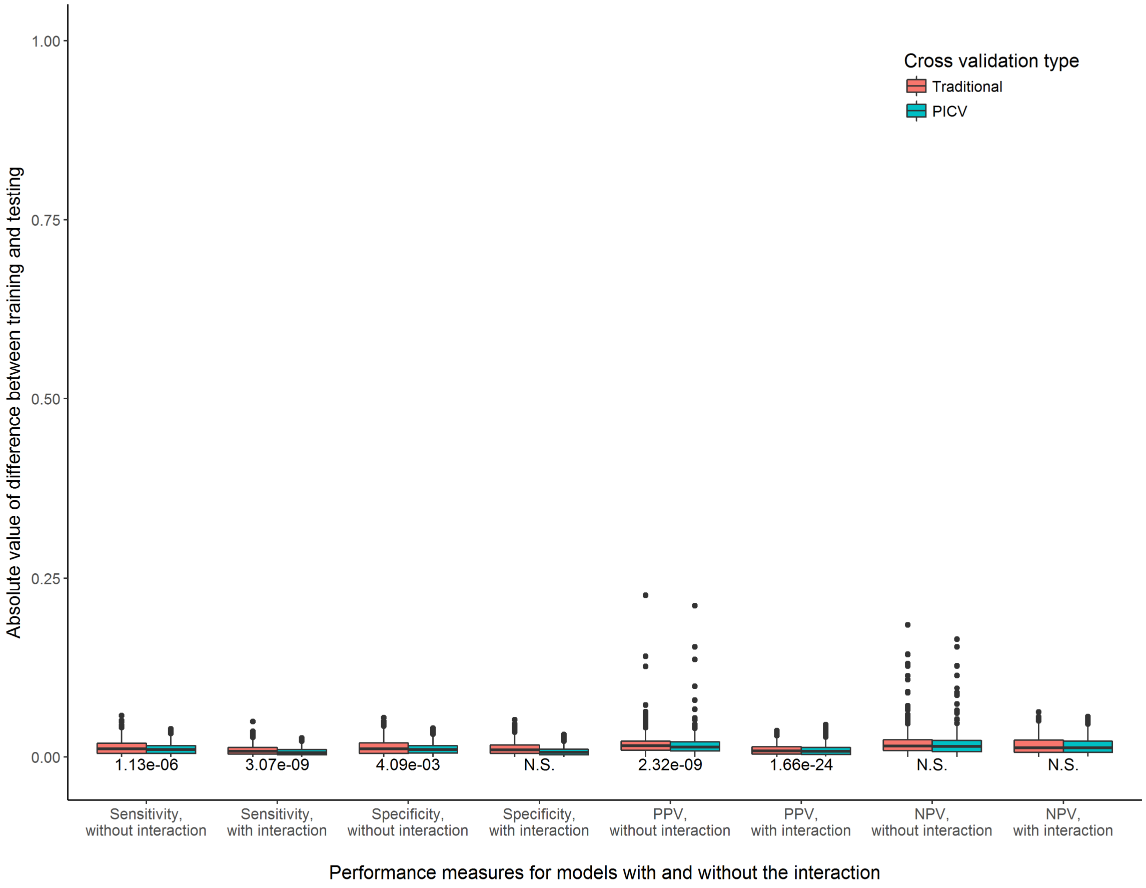
Supplemental Figure 47. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 2, prevalence = 0.02, n = 10000



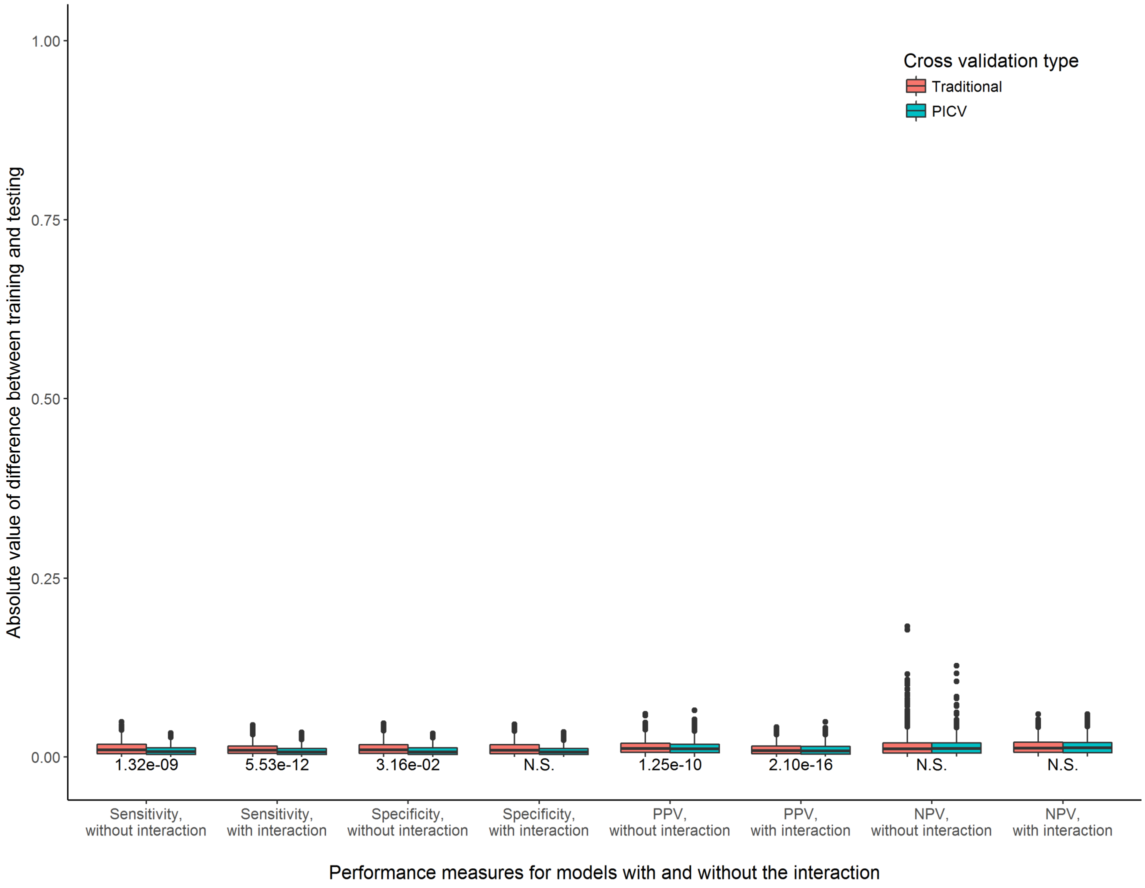
Supplemental Figure 48. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 3, prevalence = 0.02, n = 10000



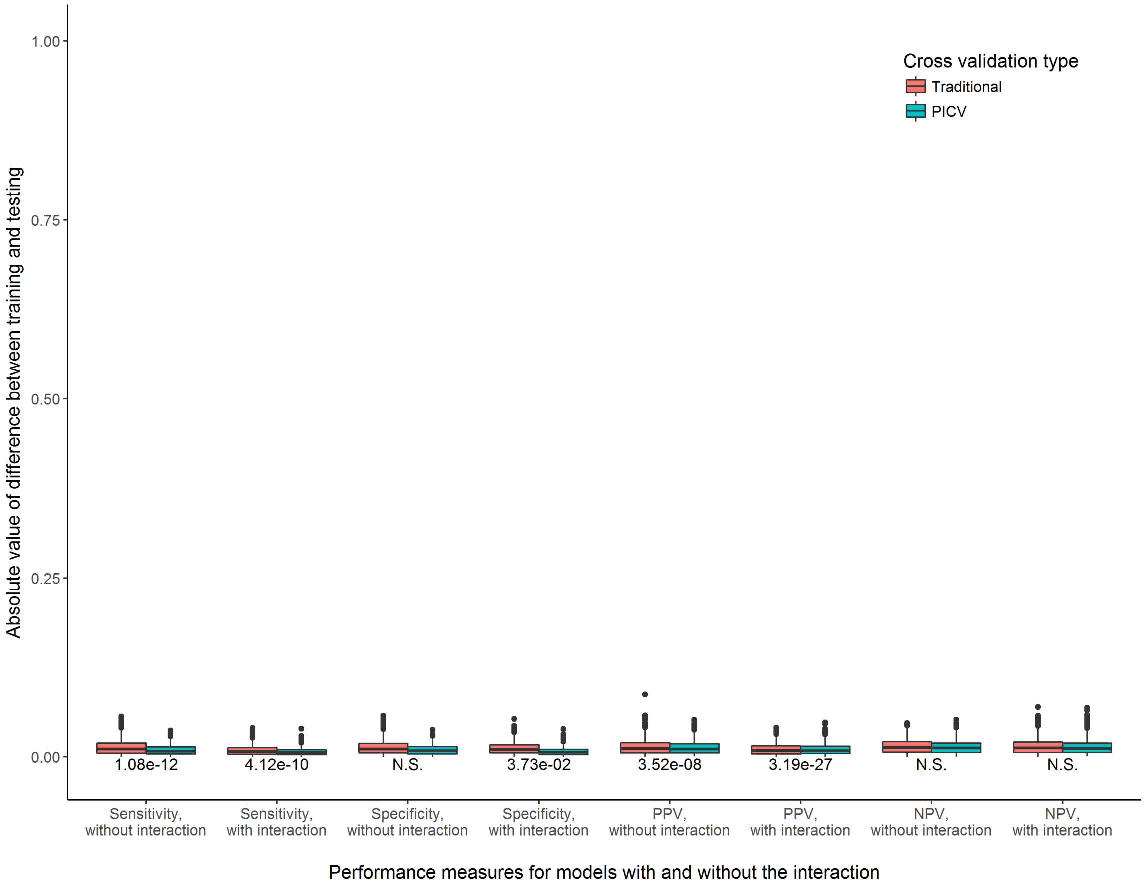
Supplemental Figure 49. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 4, prevalence = 0.02, n = 10000



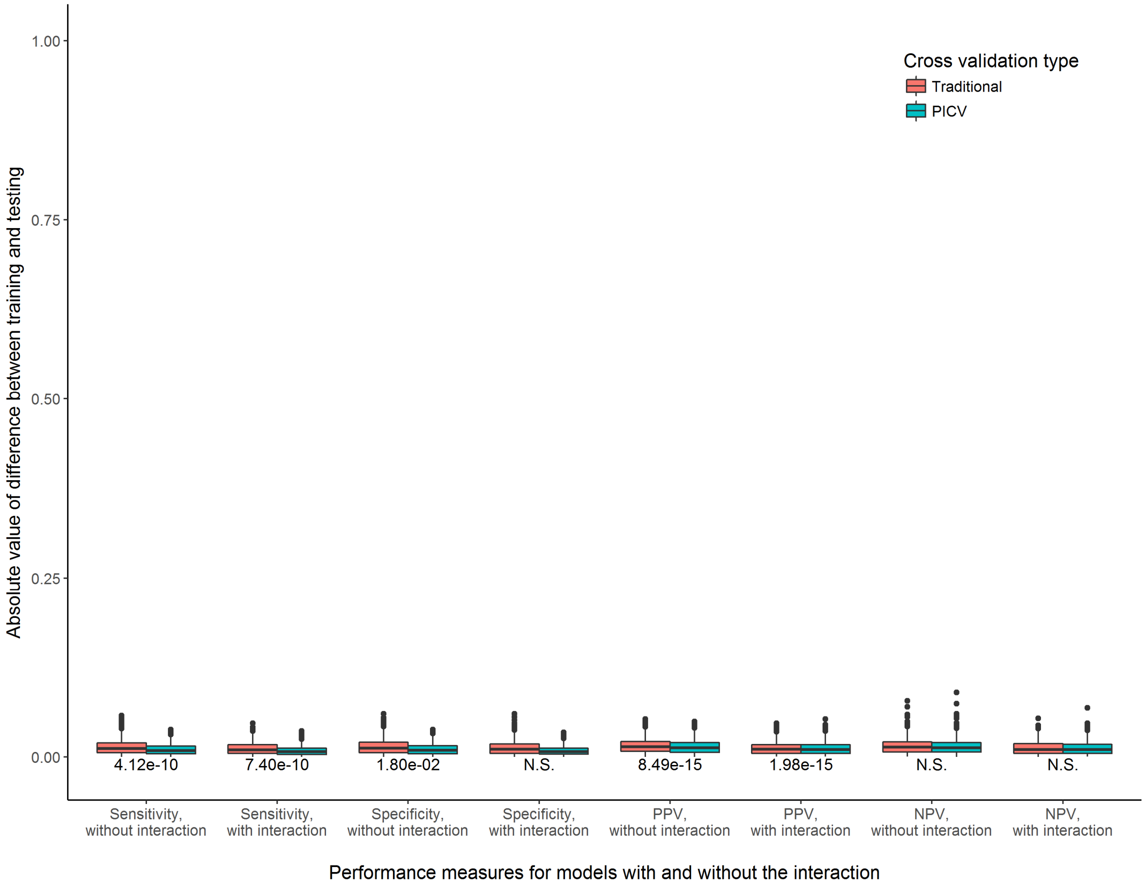
Supplemental Figure 50. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 5, prevalence = 0.02, n = 10000



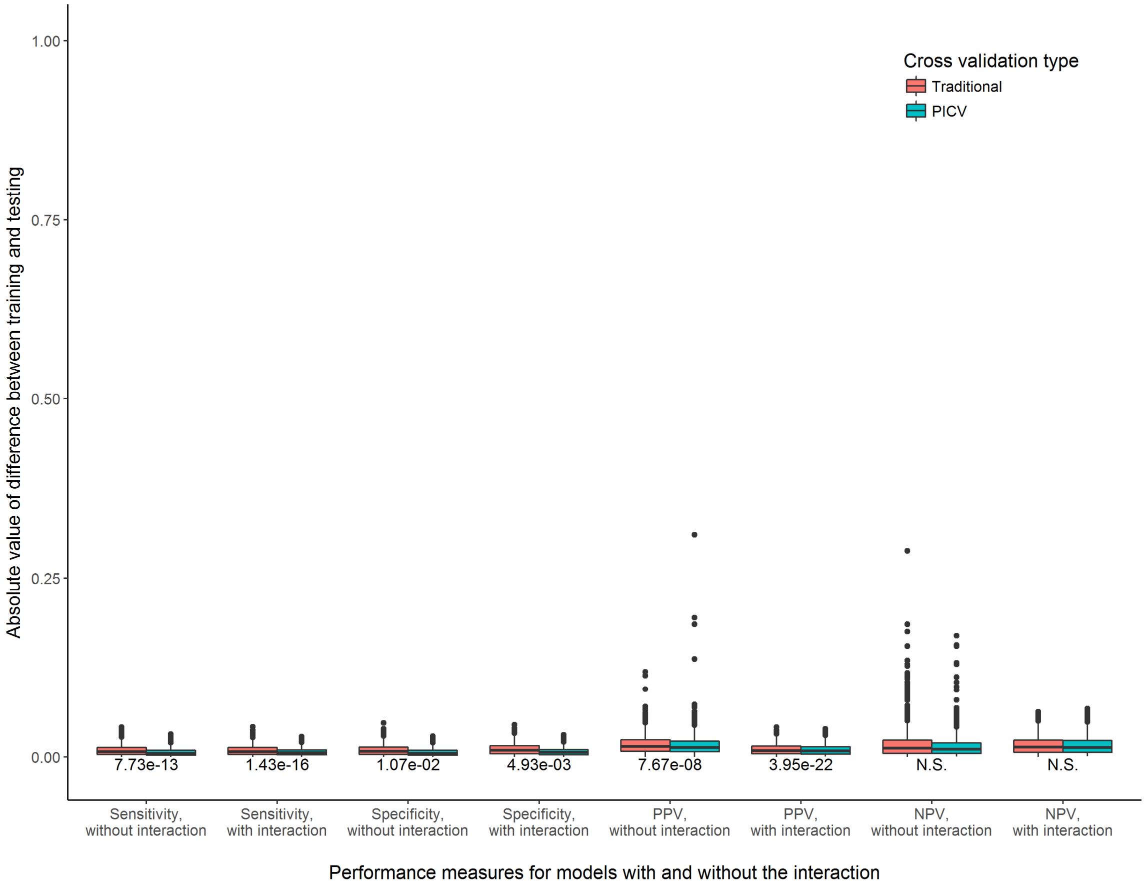
Supplemental Figure 51. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 6, prevalence = 0.02, n = 10000



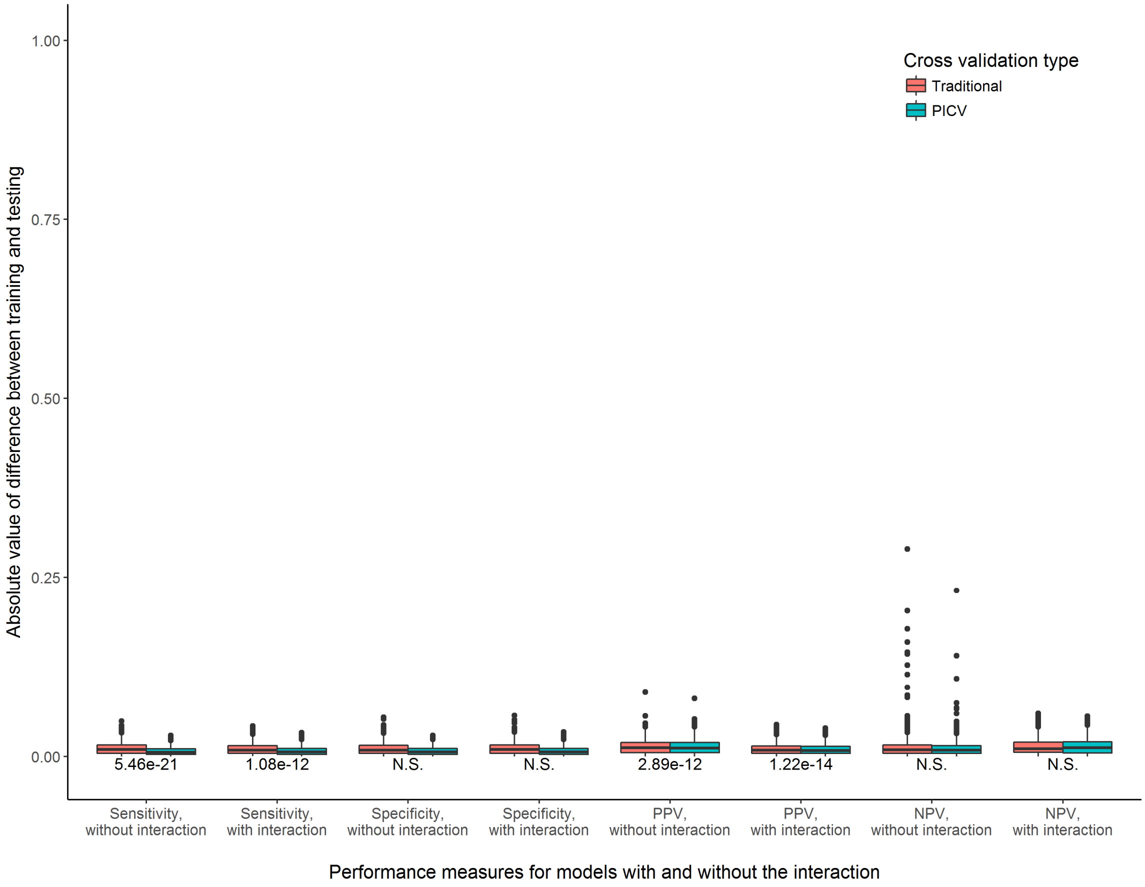
Supplemental Figure 52. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 7, prevalence = 0.02, n = 10000



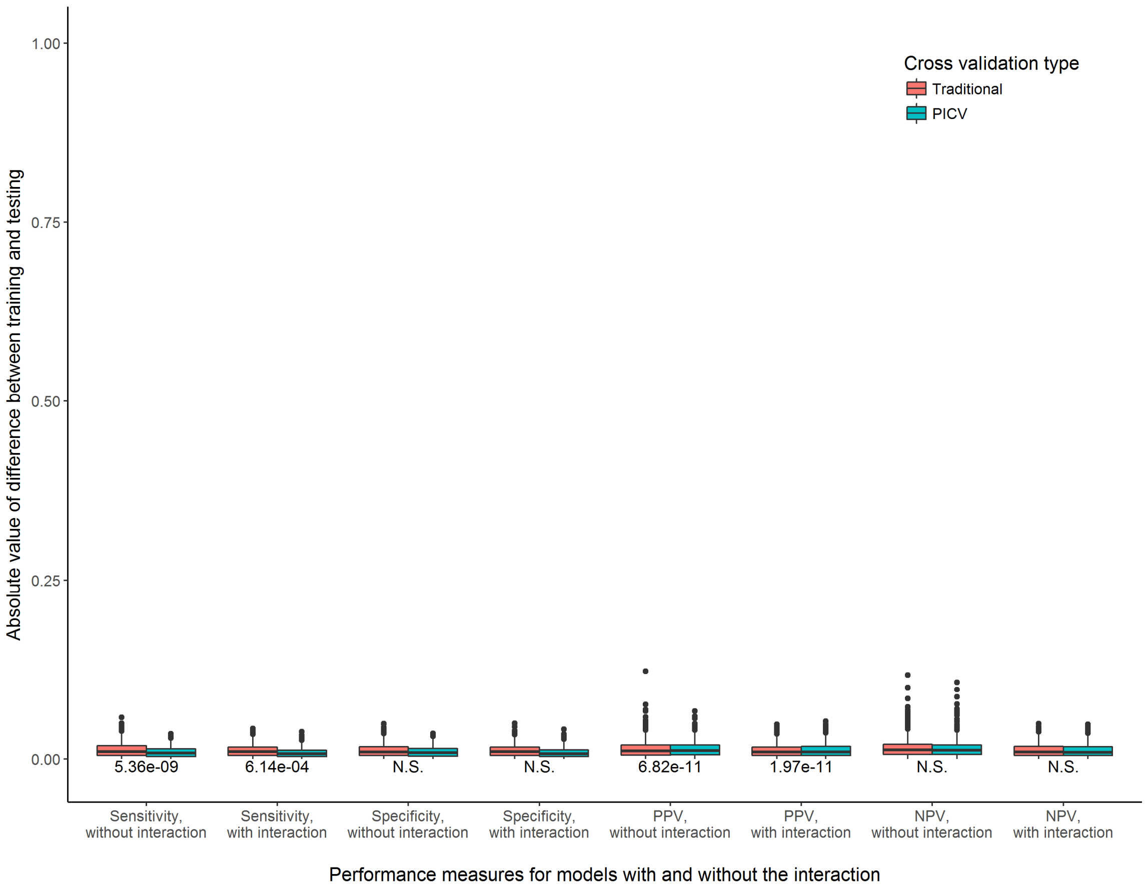
Supplemental Figure 53. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 8, prevalence = 0.02, n = 10000



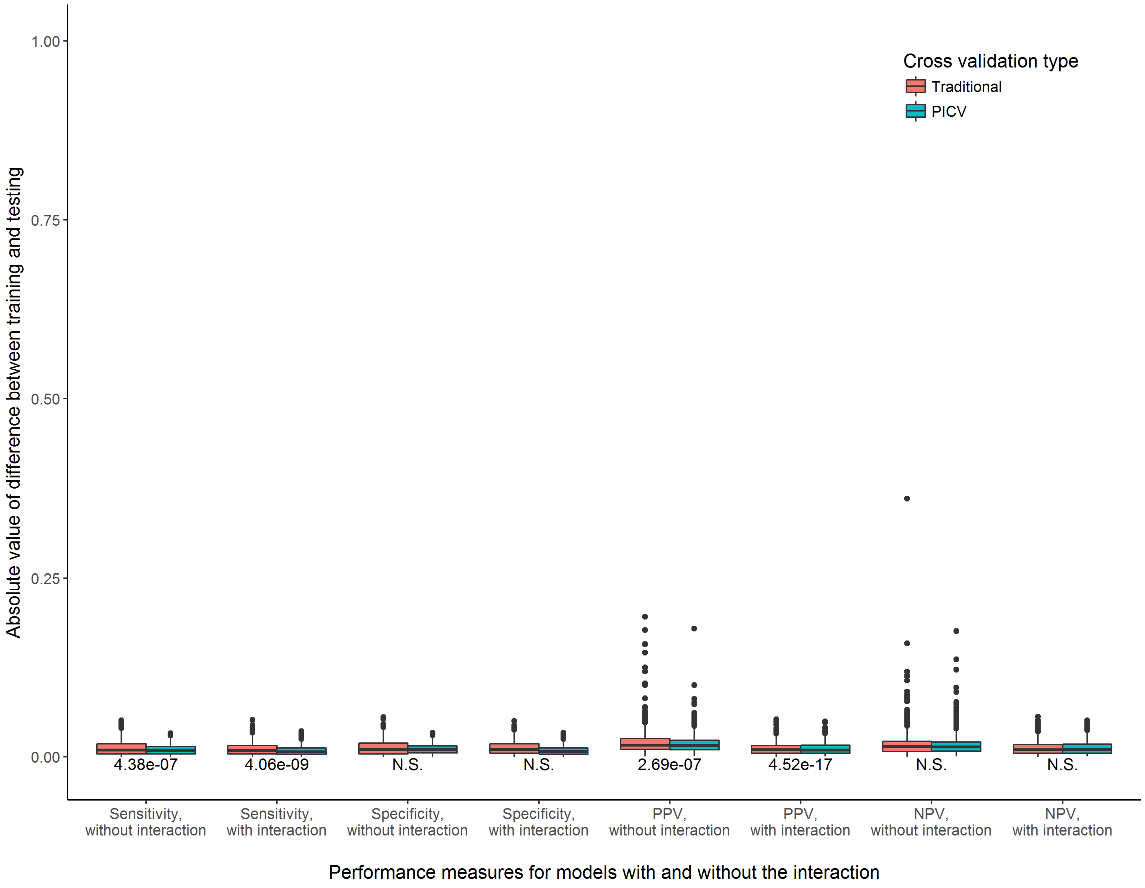
Supplemental Figure 54. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 9, prevalence = 0.02, n = 10000



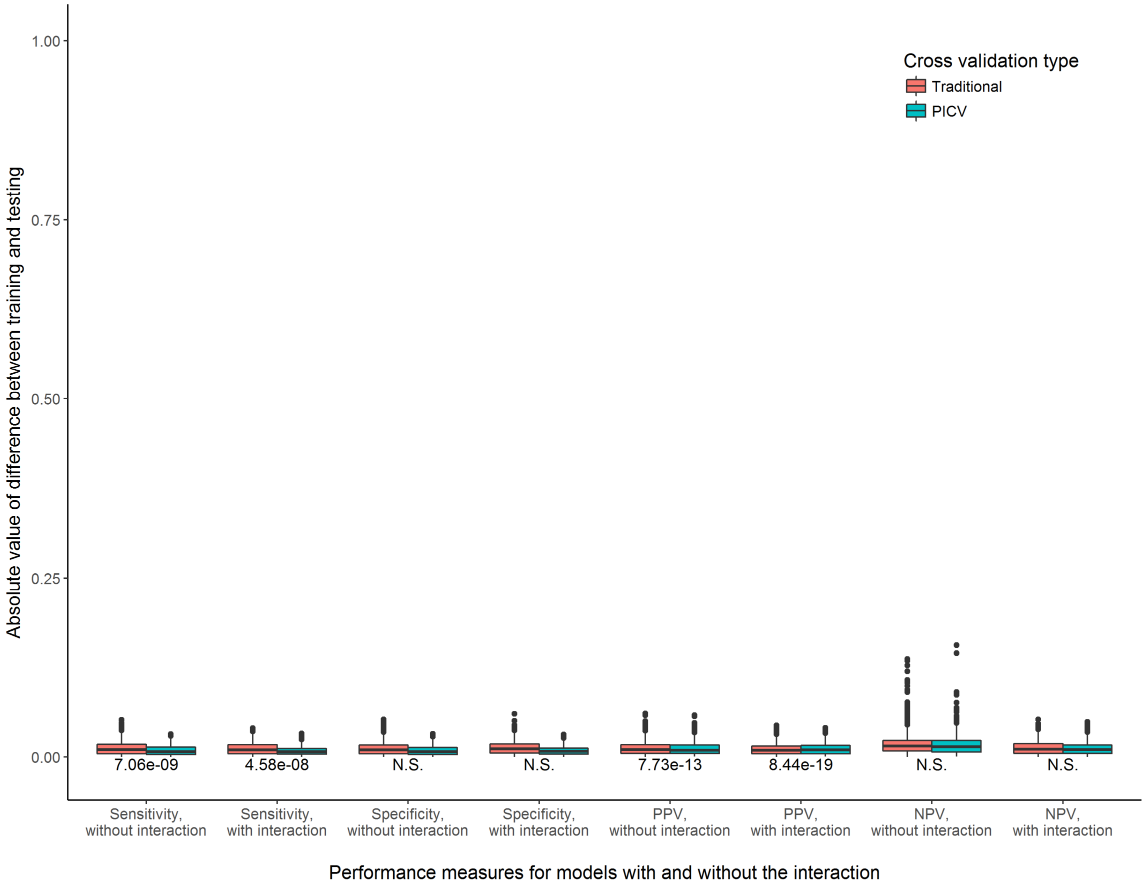
Supplemental Figure 55. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 10, prevalence = 0.02, n = 10000



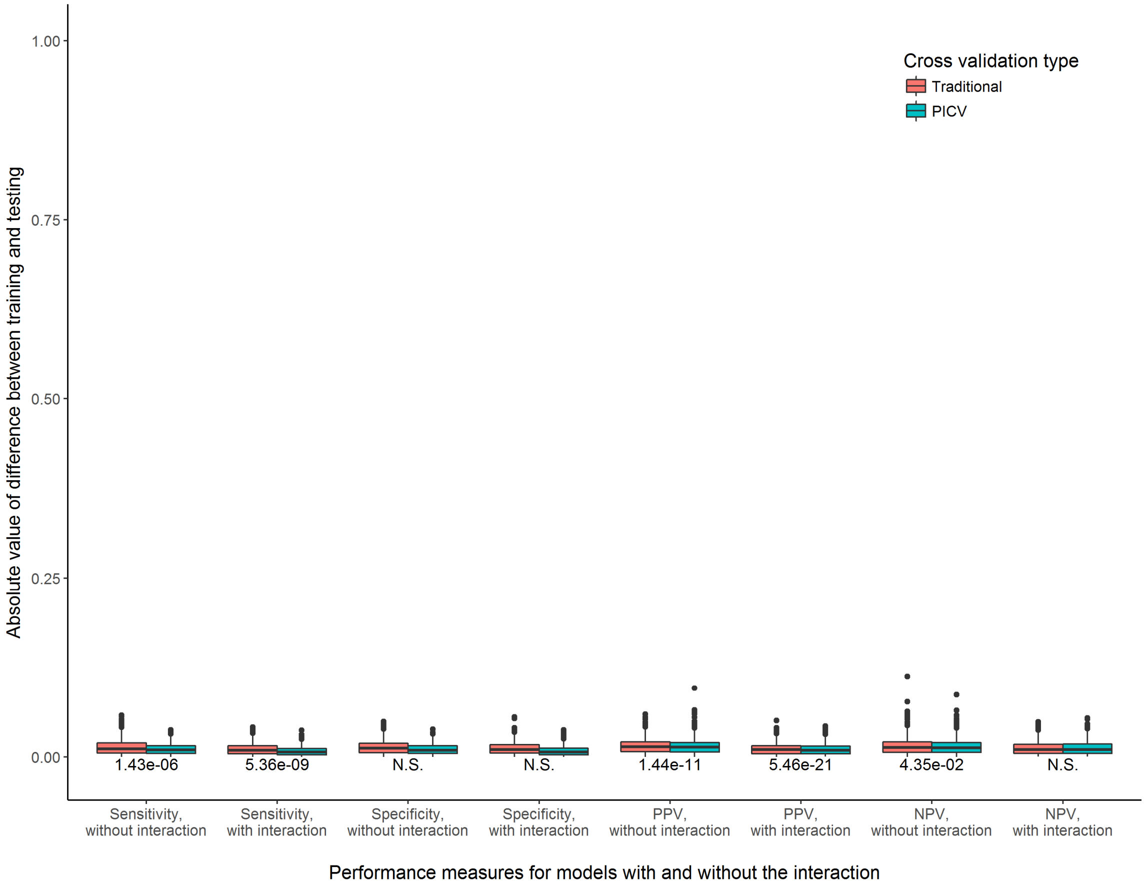
Supplemental Figure 56. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 11, prevalence = 0.02, n = 10000



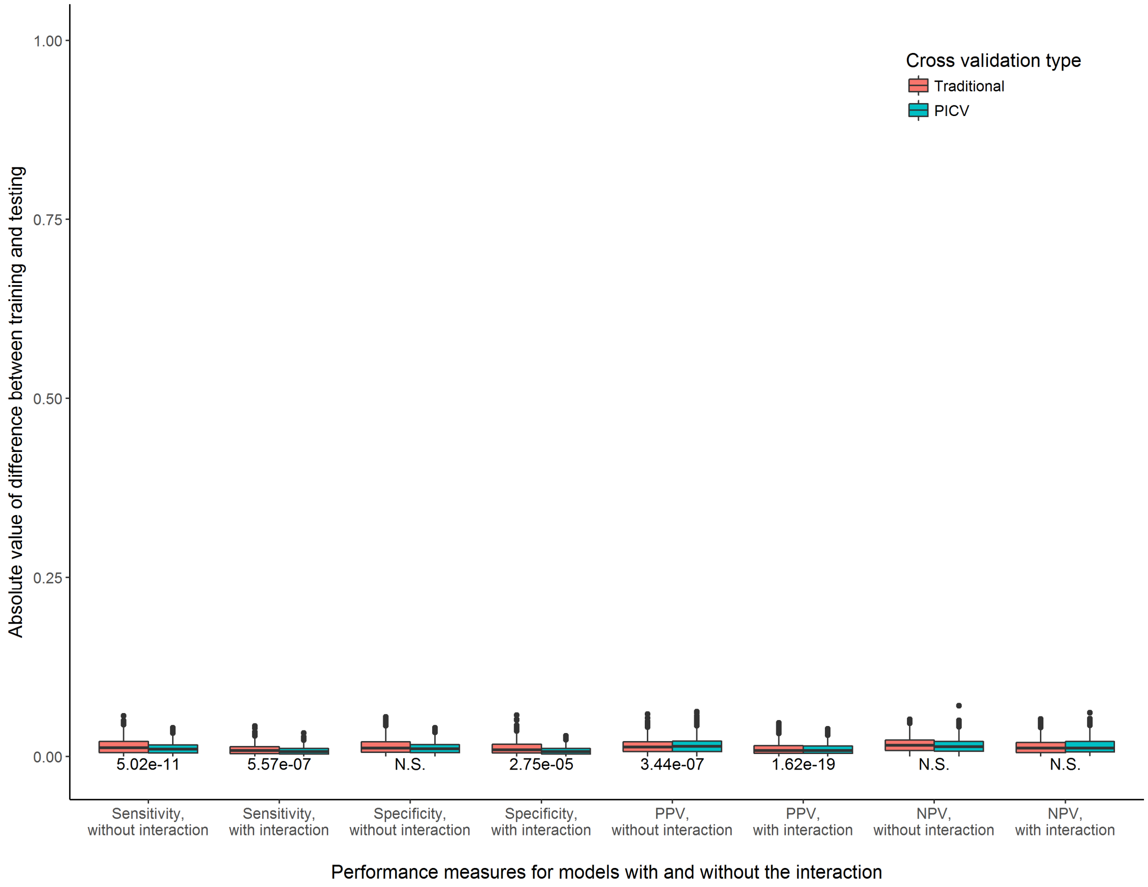
Supplemental Figure 57. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 12, prevalence = 0.02, n = 10000



Supplemental Figure 58. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 13, prevalence = 0.02, n = 10000



Supplemental Figure 59. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 14, prevalence = 0.02, n = 10000

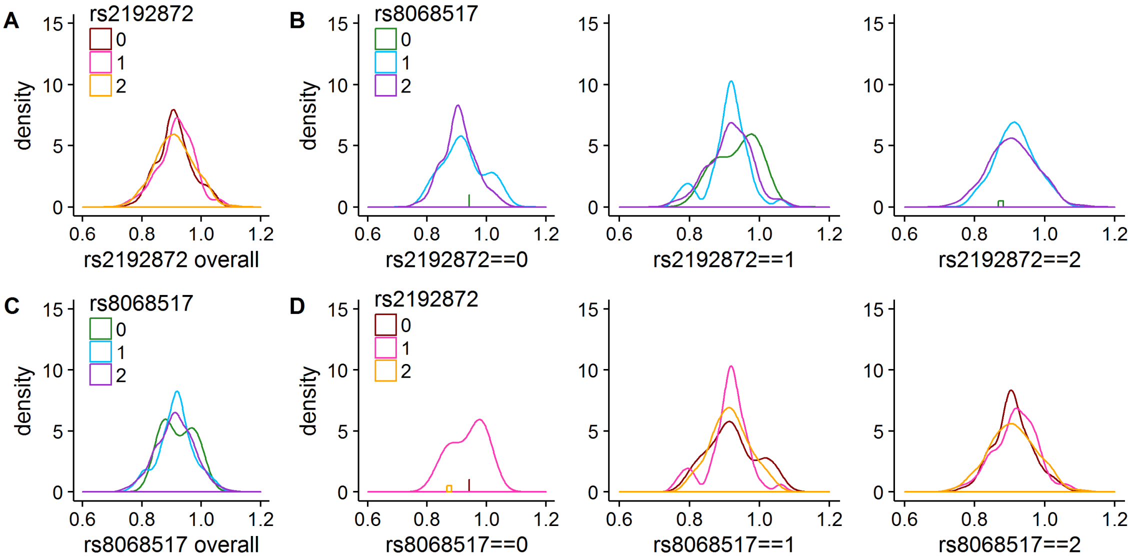


Supplemental Figure 60. Consistency of training and testing performance measures for models with and without the interaction term, comparing a traditional cross validation procedure to PICV. Experimental scenario 15, prevalence = 0.02, n = 10000

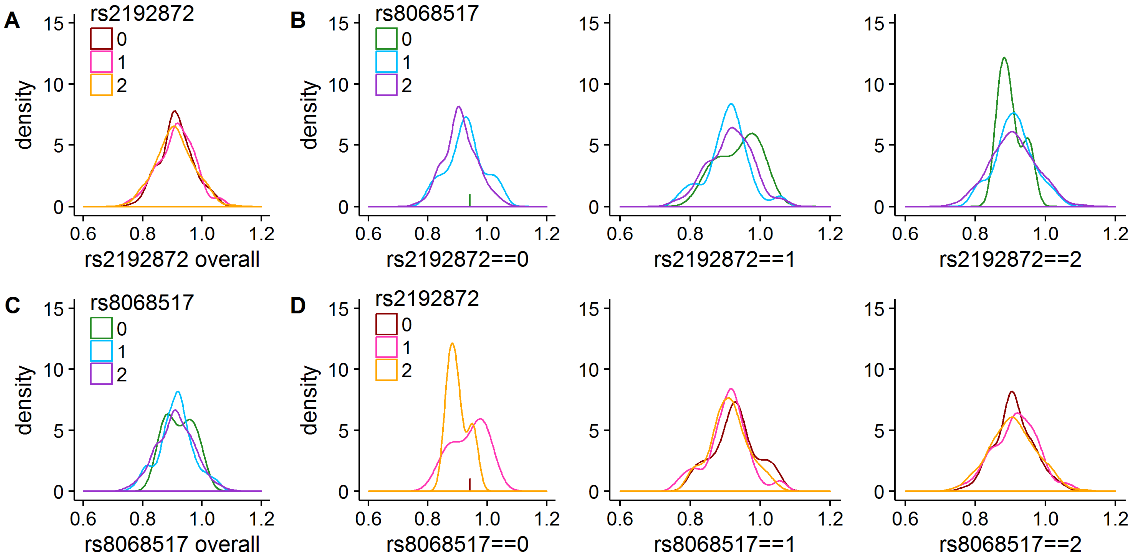
Supplemental Tables and Figures, Chapter 4

Supplemental table 1. Counts and mean values of the phenotype by interaction genotype among individuals included in the simulated analysis, the real analysis, and exclusively in the real analysis but not the simulated analysis. Interaction genotypes listed refer to minor allele counts for rs2192872 and rs8068517, respectively (e.g. 01 refers to the interaction genotype of individuals with 0 copies of the minor allele for rs2192872 and 1 copy of the minor allele for rs8068517).

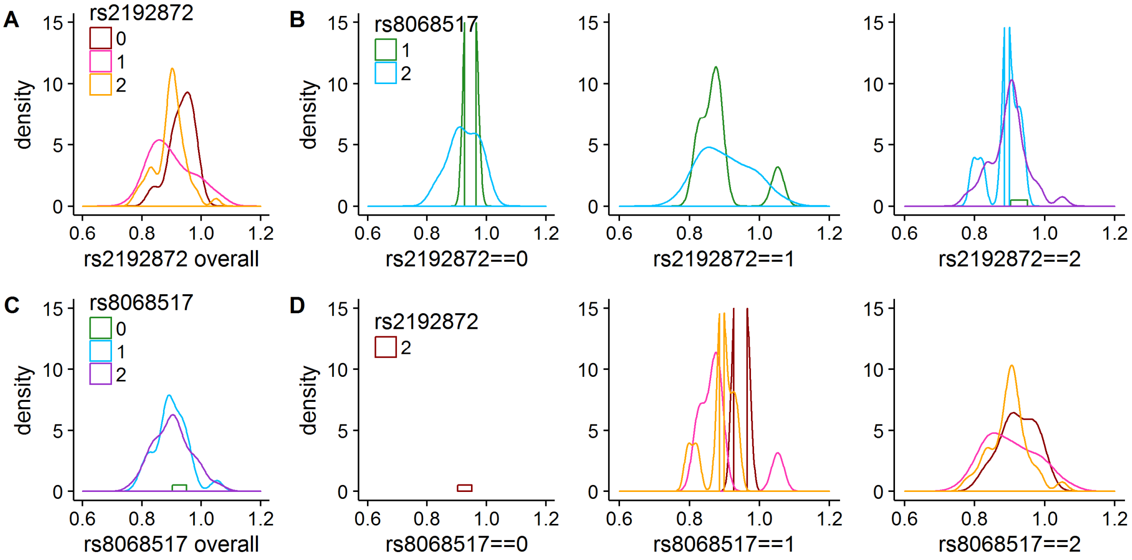
rs2192872: rs8068517 interaction genotype	**00**	**01**	**02**	**10**	**11**	**12**	**20**	**21**	**22**
Simulated (n = 680)	0.942 (1)	0.922 (19)	0.909 (76)	0.938 (5)	0.910 (42)	0.916 (238)	0.874 (2)	0.916 (57)	0.909 (240)
Real (n = 778)	0.942 (1)	0.927 (24)	0.910 (83)	0.938 (5)	0.907 (49)	0.914 (273)	0.900 (4)	0.912 (67)	0.908 (272)
Exclusively real (n = 98)	N/A (0)	0.946 (5)	0.924 (7)	N/A (0)	0.887 (7)	0.903 (35)	0.927 (2)	0.885 (10)	0.898 (32)



Supplemental figure 1. Phenotype distributions by genotype for the 680 subjects included in the analysis of the simulated data. A. Phenotype distributions for individuals with 0, 1, or 2 copies of the minor allele for rs2192872. B. Phenotype distributions for individuals with 0, 1, or 2 copies of the minor allele for rs8068517 amongst those with 0 copies of the minor allele for rs2192872 [left]; 1 copy [center]; 2 copies [right]. C. Phenotype distributions for individuals with 0, 1, or 2 copies of the minor allele for rs8068517. D. Phenotype distributions for individuals with 0, 1, or 2 copies of the minor allele for rs8068517 [left]; 1 copy [center]; 2 copies [right].



Supplemental figure 2. Phenotype distributions by genotype for the 778 subjects included in the analysis of the real data. A. Phenotype distributions for individuals with 0, 1, or 2 copies of the minor allele for rs2192872. B. Phenotype distributions for individuals with 0, 1, or 2 copies of the minor allele for rs8068517 amongst those with 0 copies of the minor allele for rs2192872 [left]; 1 copy [center]; 2 copies [right]. C. Phenotype distributions for individuals with 0, 1, or 2 copies of the minor allele for rs8068517. D. Phenotype distributions for individuals with 0, 1, or 2 copies of the minor allele for rs2192872 amongst those with 0 copies of the minor allele for rs8068517 [left]; 1 copy [center]; 2 copies [right].



Supplemental figure 3. Phenotype distributions by genotype for the 98 subjects exclusively included in the analysis of the real data and not included in the analysis of the simulated data. A. Phenotype distributions for individuals with 0, 1, or 2 copies of the minor allele for rs2192872. B. Phenotype distributions for individuals with 0, 1, or 2 copies of the minor allele for rs8068517 amongst those with 0 copies of the minor allele for rs2192872 [left]; 1 copy [center]; 2 copies [right]. C. Phenotype distributions for individuals with 0, 1, or 2 copies of the minor allele for rs8068517. D. Phenotype distributions for individuals with 0, 1, or 2 copies of the minor allele for rs2192872 amongst those with 0 copies of the minor allele for rs8068517 [left]; 1 copy [center]; 2 copies [right].

BIBLIOGRAPHY

1. Adams, N. (2010). Dataset Shift in Machine Learning. Journal of the Royal Statistical Society: Series A (Statistics in Society), 173(1), 274–274. https://doi.org/10.1111/j.1467-985X.2009.00624_10.x

2. Agirre, X., Vilas-Zornoza, A., Jiménez-Velasco, A., Martin-Subero, J. I., Cordeu, L., Gárate, L., … Prósper, F. (2009). Epigenetic silencing of the tumor suppressor microRNA Hsa-miR-124a regulates CDK6 expression and confers a poor prognosis in acute lymphoblastic leukemia. Cancer Research, 69(10), 4443–4453. https://doi.org/10.1158/0008-5472.CAN-08-4025

3. Arlot, S., & Celisse, A. (2010). Cross validation. Statistics Surveys, 4(0), 40–79. https://doi.org/10.1214/09-SS054

4. Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. Nature, 533(7604), 452–454. https://doi.org/10.1038/533452a

5. Barandela, R., Sánchez, J. S., García, V., & Rangel, E. (2003). Strategies for learning in class imbalance problems. Pattern Recognition, 36(3), 849–851. https://doi.org/10.1016/S0031-3203(02)00257-1

6. Barnard, G. A. (1963). New methods of quality control. Journal of the Royal Statistical Society, 126(2), 255–258.

7. Baroukh, N. N., & Van Obberghen, E. (2009). Function of microRNA-375 and microRNA-124a in pancreas and brain. FEBS Journal. https://doi.org/10.1111/j.1742-4658.2009.07353.x

8. Bates, J. M., & Granger, C. W. J. (1969). The Combination of Forecasts. OR, 20(4), 451. https://doi.org/10.2307/3008764

9. Bauchner, H., Golub, R. M., & Fontanarosa, P. B. (2016). Data sharing: An ethical and scientific imperative. JAMA - Journal of the American Medical Association, 315(12), 1237–1239. https://doi.org/10.1001/jama.2016.2420

10. Beaulieu-Jones, B. K., & Greene, C. S. (2017). Reproducibility of computational workflows is automated using continuous analysis. Nature Biotechnology, 35(4), 342–346. https://doi.org/10.1038/nbt.3780

11. Beekman, M., Nederstigt, C., Suchiman, H. E. D., Kremer, D., van der Breggen, R., Lakenberg, N., … Slagboom, P. E. (2010). Genome-wide association study (GWAS)-identified disease risk alleles do not compromise human longevity. Proceedings of the National Academy of Sciences of the United States of America, 107(42), 18046–18049. https://doi.org/10.1073/pnas.1003540107

12. Begley, C. G., & Ioannidis, J. P. A. (2015). Reproducibility in science: Improving the standard for basic and preclinical research. Circulation Research. https://doi.org/10.1161/CIRCRESAHA.114.303819

13. Begum, F., Ghosh, D., Tseng, G. C., & Feingold, E. (2012). Comprehensive literature review and statistical considerations for GWAS meta-analysis. Nucleic Acids Research. https://doi.org/10.1093/nar/gkr1255

14. Bellman, R. (1957). Dynamic Programming. Princeton University Press. Retrieved from https://books.google.it/books?id=wdtoPwAACAAJ&hl=en

15. Ben Gacem, R., Ben Abdelkrim, O., Ziadi, S., Ben Dhiab, M., & Trimeche, M. (2014). Methylation of miR-124a-1, miR-124a-2, and miR-124a-3 genes correlates with aggressive and advanced breast cancer disease. Tumor Biology, 35(5), 4047–4056. https://doi.org/10.1007/s13277-013-1530-4

16. Bergstra J., & Yoshua Bengio. (2012). Random Search for Hyper-Parameter Optimization. Journal of Machine Learning Research, 13, 281–305. https://doi.org/10.1162/153244303322533223

17. Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for Hyper-Parameter Optimization. In Advances in Neural Information Processing Systems (NIPS) (pp. 2546–2554). https://doi.org/2012arXiv1206.2944S

18. Berulava, T., Rahmann, S., Rademacher, K., Klein-Hitpass, L., & Horsthemke, B. (2015). N6-Adenosine Methylation in MiRNAs. PLoS ONE, 10(2). https://doi.org/10.1371/journal.pone.0118438

19. Boettiger, C. (2015). An introduction to Docker for reproducible research. ACM SIGOPS Operating Systems Review, 49(1), 71–79. https://doi.org/10.1145/2723872.2723882

20. Botezatu, A., Goia-Rusanu, C. D., Iancu, I. V., Huica, I., Plesa, A., Socolov, D., … Anton, G. (2011). Quantitative analysis of the relationship between microRNA-124a, -34b and -203 gene methylation and cervical oncogenesis. Molecular Medicine Reports, 4(1), 121–128. https://doi.org/10.3892/mmr.2010.394

21. Breiman, L. (1998). Arcing classifier (with discussion and a rejoinder by the author). The Annals of Statistics, 26(3), 801–849. https://doi.org/10.1214/aos/1024691079

22. Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), 123–140. https://doi.org/10.1007/BF00058655

23. Brochu, E., Cora, V. M., & De Freitas, N. (2010). A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. ArXiv, 49. https://doi.org/1012.2599

24. Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. In Proceedings - International Conference on Pattern Recognition (pp. 3121–3124). https://doi.org/10.1109/ICPR.2010.764

25. Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., … Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. Proceedings of the National Academy of Sciences of the United States of America, 97(1), 262–267. https://doi.org/10.1073/pnas.97.1.262

26. Bush, W. S., Dudek, S. M., & Ritchie, M. D. (2009). Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. Pacific Symposium on Biocomputing, 368–379. https://doi.org/10.1016/j.bbi.2008.05.010

27. Buzdugan, L., Kalisch, M., Navarro, A., Schunk, D., Fehr, E., & Bühlmann, P. (2016). Assessing statistical significance in multivariable genome wide association analysis. Bioinformatics, 32(13), 1990–2000. https://doi.org/10.1093/bioinformatics/btw128

28. Cantor, R. M., Lange, K., & Sinsheimer, J. S. (2010). Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. American Journal of Human Genetics. https://doi.org/10.1016/j.ajhg.2009.11.017

29. Carlborg, Ö., & Haley, C. S. (2004). Epistasis: Too often neglected in complex trait studies? Nature Reviews Genetics. https://doi.org/10.1038/nrg1407

30. CDC. (2015). Chronic Disease Overview | Publications | Chronic Disease Prevention and Health Promotion | CDC. Retrieved December 19, 2017, from https://www.cdc.gov/chronicdisease/overview/index.htm.

31. Chakraborty, C., Doss, C. G. P., Bandyopadhyay, S., & Agoramoorthy, G. (2014). Influence of miRNA in insulin signaling pathway and insulin resistance: Micro-molecules with a major role in type-2 diabetes. Wiley Interdisciplinary Reviews: RNA, 5(5), 697–712. https://doi.org/10.1002/wrna.1240

32. Chamberlain, R., & Schommer, J. (2014). Using Docker to support Reproducible Research (submission to WSSSPE2). Figshare, 1–4. https://doi.org/10.6084/m9.figshare.1101910

33. Chanock, S. J., Manolio, T., Boehnke, M., Boerwinkle, E., Hunter, D. J., Thomas, G., … Collins, F. S. (2007). Replicating genotype-phenotype associations. Nature. https://doi.org/10.1038/447655a

34. Chen, T., Hao, Y. J., Zhang, Y., Li, M. M., Wang, M., Han, W., … Zhou, Q. (2015). M6A RNA methylation is regulated by microRNAs and promotes reprogramming to pluripotency. Cell Stem Cell, 16(3), 289–301. https://doi.org/10.1016/j.stem.2015.01.016

35. Chen, X., He, D., Dong, X. Da, Dong, F., Wang, J., Wang, L., … Tu, L. L. (2013). MicroRNA-124a is epigenetically regulated and acts as a tumor suppressor by controlling multiple targets in uveal melanoma. Investigative Ophthalmology and Visual Science, 54(3), 2248–2256. https://doi.org/10.1167/iovs.12-10977

36. Church, D. M., Schneider, V. A., Graves, T., Auger, K., Cunningham, F., Bouk, N., … Hubbard, T. (2011). Modernizing reference genome assemblies. PLoS Biology, 9(7). https://doi.org/10.1371/journal.pbio.1001091

37. Ciccacci, C., Di Fusco, D., Cacciotti, L., Morganti, R., D'Amato, C., Greco, C., … Borgiani, P. (2013). MicroRNA genetic variations: Association with type 2 diabetes. Acta Diabetologica, 50(6), 867–872. https://doi.org/10.1007/s00592-013-0469-7

38. Claerbout, J., & Karrenbach, M. (2005). Electronic documents give reproducible research a new meaning. SEG Technical Program Expanded Abstracts, 11(1), 601. https://doi.org/10.1190/1.1822162

39. Consortium, GTex (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science, 348(6235), 648–660. https://doi.org/10.1126/science.1262110

40. Cornelis, M. C., Agrawal, A., Cole, J. W., Hansel, N. N., Barnes, K. C., Beaty, T. H., … Yu, K. (2010). The gene, environment association studies consortium (GENEVA): Maximizing the knowledge obtained from GWAS by collaboration across studies of multiple conditions. Genetic Epidemiology, 34(4), 364–372. https://doi.org/10.1002/gepi.20492

41. Deng, G., Kakar, S., & Kim, Y. S. (2011). MicroRNA-124a and microRNA-34b/c are frequently methylated in all histological types of colorectal cancer and polyps, and in the adjacent normal mucosa. Oncology Letters, 2(1), 175–180. https://doi.org/10.3892/ol.2010.222

42. Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. Multiple Classifier Systems, 1857, 1–15. https://doi.org/10.1007/3-540-45014-9

43. Donoho, D. L. (2010). An invitation to reproducible computational research. Biostatistics, 11(3), 385–388. https://doi.org/10.1093/biostatistics/kxq028

44. Drummond, D. C. (2009). Replicability is not reproducibility: Nor is it good science. Proceedings of the Evaluation Methods for Machine Learning Workshop 26th International Conference for Machine Learning, (2005), 1–4.

45. Dudley, J. T., & Butte, A. J. (2010). In silico research in the era of cloud computing. Nature Biotechnology. https://doi.org/10.1038/nbt1110-1181

46. Dudoit, S., & Fridlyand, J. (2003). Bagging to improve the accuracy of a clustering procedure. Bioinformatics, 19(9), 1090–1099. https://doi.org/10.1093/bioinformatics/btg038

47. Eleftherohorinou, H., Wright, V., Hoggart, C., Hartikainen, A.-L., Jarvelin, M.-R., Balding, D. J., … Levin, M. (2009). Pathway analysis of GWAS provides new insights into genetic susceptibility to 3 inflammatory diseases. PloS One, 4(11), e8068. https://doi.org/10.1371/journal.pone.0008068

48. Ericson, G., & Rohm, W. A. (2017). Machine learning algorithm cheat sheet for Microsoft Azure Machine Learning Studio. Retrieved December 19, 2017, from https://docs.microsoft.com/en-us/azure/machine-learning/studio/algorithm-cheat-sheet

49. Estabrooks, A., Jo, T., & Japkowicz, N. (2004). A Multiple Resampling Method for Learning from Imbalanced Data Sets. Computational Intelligence, 20(1), 18–36. https://doi.org/10.1111/j.0824-7935.2004.t01-1-00228.x

50. Fernández-Hernando, C., Suárez, Y., Rayner, K. J., & Moore, K. J. (2011). MicroRNAs in lipid metabolism. Current Opinion in Lipidology, 22(2), 86–92. https://doi.org/10.1097/MOL.0b013e3283428d9d

51. Friedman, D. S., Wolfs, R. C. W., O'Colmain, B. J., Klein, B. E., Taylor, H. R., West, S., … Eye Diseases Prevalence Research Group. (2004). Prevalence of open-angle glaucoma among adults in the United States. Archives of Ophthalmology (Chicago, Ill. : 1960), 122(4), 532–538. https://doi.org/10.1001/archopht.122.4.532

52. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews. https://doi.org/10.1109/TSMCC.2011.2161285

53. Gentleman, R., & Temple Lang, D. (2007). Statistical analyses and reproducible research. Journal of Computational and Graphical Statistics. https://doi.org/10.1198/106186007X178663

54. Greene, C. S., Penrod, N. M., Williams, S. M., & Moore, J. H. (2009). Failure to replicate a genetic association may provide important clues about genetic architecture. PLoS ONE, 4(6). https://doi.org/10.1371/journal.pone.0005639

55. Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. Journal of Machine Learning Research (JMLR), 3(3), 1157–1182. https://doi.org/10.1016/j.aca.2011.07.027

56. Guzella, T. S., & Caminhas, W. M. (2009). A review of machine learning approaches to Spam filtering. Expert Systems with Applications. https://doi.org/10.1016/j.eswa.2009.02.037

57. Hand, D. J. (2010). Machine Learning: An Algorithmic Perspective by Stephen Marsland. International Statistical Review, 78(2), 325–325. https://doi.org/10.1111/j.1751-5823.2010.00118_11.x

58. Hawkins, D. M. (2004). The Problem of Overfitting. Journal of Chemical Information and Computer Sciences. https://doi.org/10.1021/ci0342472

59. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

60. Henckaerts, L., Van Steen, K., Verstreken, I., Cleynen, I., Franke, A., Schreiber, S., … Vermeire, S. (2009). Genetic Risk Profiling and Prediction of Disease Course in Crohn's Disease Patients. Clinical Gastroenterology and Hepatology, 7(9). https://doi.org/10.1016/j.cgh.2009.05.001

61. Hernandez, D. G., Nalls, M. A., Moore, M., Chong, S., Dillman, A., Trabzuni, D., … Cookson, M. R. (2012). Integration of GWAS SNPs and tissue specific expression profiling reveal discrete eQTLs for human traits in blood and brain. Neurobiology of Disease, 47(1), 20–28. https://doi.org/10.1016/j.nbd.2012.03.020

62. Hines, W. C., Su, Y., Kuhn, I., Polyak, K., & Bissell, M. J. (2014). Sorting out the FACS: A devil in the details. Cell Reports. https://doi.org/10.1016/j.celrep.2014.02.021

63. Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian Model Averaging: A Tutorial. Statistical Science, 14(4), 382–417. https://doi.org/10.2307/2676803

64. Howe, B. (2012). Virtual appliances, cloud computing, and reproducible research. Computing in Science and Engineering, 14(4), 36–41. https://doi.org/10.1109/MCSE.2012.62

65. Inza, I., Larrañaga, P., Blanco, R., & Cerrolaza, A. J. (2004). Filter versus wrapper gene selection approaches in DNA microarray domains. Artificial Intelligence in Medicine, 31(2), 91–103. https://doi.org/10.1016/j.artmed.2004.01.007

66. Irvin, M. R., Zhi, D., Joehanes, R., Mendelson, M., Aslibekyan, S., Claas, S. A., … Arnett, D. K. (2014). Epigenome-wide association study of fasting blood lipids in the genetics of lipid-lowering drugs and diet network study. Circulation, 130(7), 565–572. https://doi.org/10.1161/CIRCULATIONAHA.114.009158

67. Jia, P., Wang, L., Meltzer, H. Y., & Zhao, Z. (2011). Pathway-based analysis of GWAS datasets: Effective but caution required. International Journal of Neuropsychopharmacology. https://doi.org/10.1017/S1461145710001446

68. Jia, P., Wang, L., Meltzer, H. Y., & Zhao, Z. (2010). Common variants conferring risk of schizophrenia: A pathway analysis of GWAS data. Schizophrenia Research, 122(1–3), 38–42. https://doi.org/10.1016/j.schres.2010.07.001

69. Jia, P., & Zhao, Z. (2014). Network-assisted analysis to prioritize GWAS results: Principles, methods and perspectives. Human Genetics. https://doi.org/10.1007/s00439-013-1377-1

70. Johnson, R. C., Nelson, G. W., Troyer, J. L., Lautenberger, J. A., Kessing, B. D., Winkler, C. A., & O'Brien, S. J. (2010). Accounting for multiple comparisons in a genome-wide association study (GWAS). BMC Genomics, 11(1). https://doi.org/10.1186/1471-2164-11-724

71. Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science. https://doi.org/10.1126/science.aaa8415

72. Kluyver, T., Ragan-kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., … Willing, C. (2016). Jupyter Notebooks—a publishing format for reproducible computational workflows. Positioning and Power in Academic Publishing: Players, Agents and Agendas. https://doi.org/10.3233/978-1-61499-649-1-87

73. Kong, L., Zhu, J., Han, W., Jiang, X., Xu, M., Zhao, Y., … Zhao, L. (2011). Significance of serum microRNAs in pre-diabetes and newly diagnosed type 2 diabetes: A clinical study. Acta Diabetologica, 48(1), 61–69. https://doi.org/10.1007/s00592-010-0226-0

74. Kraft, P., Zeggini, E., & Ioannidis, J. (2010). Replication in genome-wide association studies. Stat Sci, 24(4), 561–573. https://doi.org/10.1214/09-STS290.Replication

75. Krizhevsky, A., Sutskever, I., & Geoffrey E., H. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems 25 (NIPS2012), 1–9. https://doi.org/10.1109/5.726791

76. Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., … Robles, V. (2006). Machine learning in bioinformatics. Briefings in Bioinformatics. https://doi.org/10.1093/bib/bbk007

77. Leamer, E. E. (1978). Specification Searches: Ad Hoc Inference with Nonexperimental Data. Wiley.

78. Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. Nat Rev Genet, 16(6), 321–332. https://doi.org/10.1038/nrg3920

79. Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., … Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science, 326(5950), 289–293. https://doi.org/10.1126/science.1181369

80. Lin, W. J., & Chen, J. J. (2013). Class-imbalanced classifiers for high-dimensional data. Briefings in Bioinformatics. https://doi.org/10.1093/bib/bbs006

81. Little, R. J. a, & Rubin, D. B. (2002). Statistical Analysis with Missing Data. Statistical analysis with missing data Second edition. https://doi.org/10.2307/1533221

82. Liu, N., & Pan, T. (2016). N6-methyladenosine-encoded epitranscriptomics. Nature Structural and Molecular Biology. https://doi.org/10.1038/nsmb.3162

83. Liu, X.-Y., Wu, J., & Zhou, Z.-H. (2009). Exploratory Undersampling for Class Imbalance Learning. IEEE Transactions on Systems, Man and Cybernetics, 39(2), 539–550. https://doi.org/10.1109/TSMCB.2008.2007853

84. Loeliger, J. (2009). Version Control with Git. Environmental Science and Technology. Retrieved from http://books.google.ca/books?id=78lsu1nMYm0C

85. Longo, D. L., & Drazen, J. M. (2016). Data Sharing. New England Journal of Medicine, 374(3), 276–277. https://doi.org/10.1056/NEJMe1516564

86. López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences, 250, 113–141. https://doi.org/10.1016/j.ins.2013.07.007

87. Luo, J., Wu, M., Gopukumar, D., & Zhao, Y. (2016). Big Data Application in Biomedical Research and Health Care: A Literature Review. Biomedical Informatics Insights, 1. https://doi.org/10.4137/BII.S31559

88. Lynn, F. C. (2009). Meta-regulation: microRNA regulation of glucose and lipid metabolism. Trends in Endocrinology & Metabolism, 20(9), 452–459. https://doi.org/10.1016/j.tem.2009.05.007

89. Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. Nature. https://doi.org/10.1038/nature08494

90. Marchini, J., Donnelly, P., & Cardon, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. Nature Genetics, 37(4), 413–417. https://doi.org/10.1038/ng1537

91. McClellan, J., & King, M.-C. (2010). Genetic Heterogeneity in Human Disease. Cell, 141(2), 210–217. https://doi.org/10.1016/j.cell.2010.03.032

92. McKinney, B. A., & Pajewski, N. M. (2012). Six degrees of epistasis: Statistical network models for GWAS. Frontiers in Genetics. https://doi.org/10.3389/fgene.2011.00109

93. Merkel, D. (2014). Docker: lightweight Linux containers for consistent development and deployment. Linux Journal. https://doi.org/10.1097/01.NND.0000320699.47006.a3

94. Mesirov, J. P. (2010). Accessible reproducible research. Science. https://doi.org/10.1126/science.1179653

95. Millman, K. J., & Pérez, F. (2014). Implementing Reproducible Research. Journal of Statistical Software, 61(October), 149–184. https://doi.org/10.1201/b16868

96. Moore, J. H. (2003). The ubiquitous nature of epistasis in determining susceptibility to common human diseases. In Human Heredity (Vol. 56, pp. 73–82). https://doi.org/10.1159/000073735

97. Moore, J. H., Asselbergs, F. W., & Williams, S. M. (2010). Bioinformatics challenges for genome-wide association studies. Bioinformatics. https://doi.org/10.1093/bioinformatics/btp713

98. Moore, J. H., & Williams, S. M. (2005). Traversing the conceptual divide between biological and statistical epistasis: Systems biology and a more modern synthesis. BioEssays. https://doi.org/10.1002/bies.20236

99. Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie Du Sert, N., … Ioannidis, J. P. A. (2017). A manifesto for reproducible science. Nature Human Behaviour. https://doi.org/10.1038/s41562-016-0021

100. National Center for Health Statistics. (2017). Heart Disease. Retrieved December 19, 2017, from https://www.cdc.gov/nchs/fastats/heart-disease.htm

101. Niu, B., Cai, Y.-D., Lu, W.-C., Li, G.-Z., & Chou, K.-C. (2006). Predicting protein structural class with AdaBoost Learner. Protein and Peptide Letters, 13(c), 489–492. https://doi.org/10.2174/092986606776819619

102. Olson, R. S. (2017). A System for Accessible Artificial Intelligence. ArXiv.

103. Olson, R. S., La Cava, W., Orzechowski, P., Urbanowicz, R. J., & Moore, J. H. (2017). PMLB: A Large Benchmark Suite for Machine Learning Evaluation and Comparison. Journal of Machine Learning Research, x. Retrieved from https://arxiv.org/pdf/1703.00512.pdf

104. Panagiotou, O. A., Ioannidis, J. P. A., Hirschhorn, J. N., Abecasis, G. R., Frayling, T. M., McCarthy, M. I., … Kesheng, W. (2012). What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. International Journal of Epidemiology, 41(1), 273–286. https://doi.org/10.1093/ije/dyr178

105. Pandey, A., Davis, N. A., White, B. C., Pajewski, N. M., Savitz, J., Drevets, W. C., & Mckinney, B. A. (2012). Epistasis network centrality analysis yields pathway replication across two GWAS cohorts for bipolar disorder. Translational Psychiatry, 2. https://doi.org/10.1038/tp.2012.80

106. Patil, P., Peng, R. D., & Leek, J. (2016). A statistical definition for reproducibility and replicability. bioRxiv. https://doi.org/10.1101/066803

107. Peng, R. D. (2011). Reproducible research in computational science. Science. https://doi.org/10.1126/science.1213847

108. Peng, R. D. (2009). Reproducible research and Biostatistics. Biostatistics, 10(3), 405–408. https://doi.org/10.1093/biostatistics/kxp014

109. Peng, R. D., Dominici, F., & Zeger, S. L. (2006). Reproducible epidemiologic research. American Journal of Epidemiology. https://doi.org/10.1093/aje/kwj093

110. Phillips, P., Lithgow, G. J., & Driscoll, M. (2017). A long journey to reproducible results. Nature. https://doi.org/10.1038/548387a

111. Price, A. L., Zaitlen, N. A., Reich, D., & Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. Nature Reviews Genetics. https://doi.org/10.1038/nrg2813

112. Refaeilzadeh, P., Tang, L., & Liu., H. (2009). “Cross-Validation.” In Encyclopedia of database systems (pp. 532–538). https://doi.org/10.1007/978-0-387-39940-9_565

113. Ritchie, M. D., Hahn, L. W., Roodi, N., Bailey, L. R., Dupont, W. D., Parl, F. F., & Moore, J. H. (2001). Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer. The American Journal of Human Genetics, 69(1), 138–147. https://doi.org/10.1086/321276

114. Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., & Kim, D. (2015). Methods of integrating data to uncover genotype-phenotype interactions. Nature Reviews Genetics. https://doi.org/10.1038/nrg3868

115. Rosenberg, N. A., Huang, L., Jewett, E. M., Szpiech, Z. A., Jankovic, I., & Boehnke, M. (2010). Genome-wide association studies in diverse populations. Nature Reviews Genetics, 11(5), 356–366. https://doi.org/10.1038/nrg2760

116. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., … Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision, 115(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

117. Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. Bioinformatics. https://doi.org/10.1093/bioinformatics/btm344

118. Saez-Rodriguez, J., Costello, J. C., Friend, S. H., Kellen, M. R., Mangravite, L., Meyer, P., … Stolovitzky, G. (2016). Crowdsourcing biomedical research: Leveraging communities as innovation engines. Nature Reviews Genetics. https://doi.org/10.1038/nrg.2016.69

119. Sandve, G. K., Nekrutenko, A., Taylor, J., & Hovig, E. (2013). Ten Simple Rules for Reproducible Computational Research. PLoS Computational Biology, 9(10). https://doi.org/10.1371/journal.pcbi.1003285

120. Schapire, R. E. (1990). The Strength of Weak Learnability. Machine Learning, 5(2), 197–227. https://doi.org/10.1023/A:1022648800760

121. Scikit-learn developers (2017). Choosing the right estimator. Retrieved December 19, 2017, from http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

122. Sebastiani, G., Po, A., Miele, E., Ventriglia, G., Ceccarelli, E., Bugliani, M., … Dotta, F. (2015). MicroRNA-124a is hyperexpressed in type 2 diabetic human pancreatic islets and negatively regulates insulin secretion. Acta Diabetologica, 52(3), 523–530. https://doi.org/10.1007/s00592-014-0675-y

123. Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. Journal of Statistical Planning and Inference, 90(2), 227–244. https://doi.org/10.1016/S0378-3758(00)00115-4

124. Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian Optimization of Machine Learning Algorithms. Adv. Neural Inf. Process. Syst. 25, 1–9. https://doi.org/2012arXiv1206.2944S

125. Sonnenburg, S., Braun, M. L., Ong, C. S., Bengio, S., Bottou, L., Holmes, G., … Williamson, R. (2007). The Need for Open Source Software in Machine Learning. J. Mach. Learn. Res., 8, 2443–2466. https://doi.org/citeulike-article-id:11849756

126. Suigyama, M., Nakajima, S., Kashima, H., Buenau, P., Kawanabe, M., Sugiyama, M., … Kawanabe, M. (2007). Direct Importance Estimation with Model Selection and Its Application to Covariate Shift Adaptation. In NIPS (pp. 1–8). https://doi.org/10.1007/s10463-008-0197-x

127. Thornton, C., Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2013). Auto-WEKA. Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’13, 847. https://doi.org/10.1145/2487575.2487629

128. Tothill, R. W., Tinker, A. V., George, J., Brown, R., Fox, S. B., Lade, S., … Bowtell, D. D. L. (2008). Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. Clinical Cancer Research, 14(16), 5198–5208. https://doi.org/10.1158/1078-0432.CCR-08-0196

129. Turner, S., Armstrong, L. L., Bradford, Y., Carlson, C. S., Crawford, D. C., Crenshaw, A. T., … Ritchie, M. D. (2011). Quality Control Procedures for Genome-Wide Association Studies. In Current Protocols in Human Genetics. https://doi.org/10.1002/0471142905.hg0119s68

130. Tycko, B. (2010). Mapping Allele-Specific DNA Methylation: A New Tool for Maximizing Information from GWAS. American Journal of Human Genetics, 86(2), 109–112. https://doi.org/10.1016/j.ajhg.2010.01.021

131. Urbanowicz, R. J., Kiralis, J., Sinnott-Armstrong, N. A., Heberling, T., Fisher, J. M., & Moore, J. H. (2012). GAMETES: A fast, direct algorithm for generating pure, strict, epistatic models with random architectures. BioData Mining, 5(1). https://doi.org/10.1186/1756-0381-5-16

132. Van Hulse, J., Khoshgoftaar, T. M., & Napolitano, A. (2007). Experimental perspectives on learning from imbalanced data. In Proceedings of the 24th international conference on Machine learning - ICML '07 (pp. 935–942). https://doi.org/10.1145/1273496.1273614

133. Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics, 7. https://doi.org/10.1186/1471-2105-7-91

134. Velez, D. R., White, B. C., Motsinger, A. A., Bush, W. S., Ritchie, M. D., Williams, S. M., & Moore, J. H. (2007). A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. Genetic Epidemiology, 31(4), 306–315. https://doi.org/10.1002/gepi.20211

135. Vellido, A., Martín-Guerrero, J. D., & Lisboa, P. J. (2012). Making machine learning models interpretable. ESANN, 12.

136. Verma, S. S., Cooke Bailey, J. N., Lucas, A., Bradford, Y., Linneman, J. G., Hauser, M. A., … Pericak-Vance, M. (2016). Epistatic Gene-Based Interaction Analyses for Glaucoma in eMERGE and NEIGHBOR Consortium. PLoS Genetics, 12(9). https://doi.org/10.1371/journal.pgen.1006186

137. Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. American Journal of Human Genetics. https://doi.org/10.1016/j.ajhg.2011.11.029

138. Wetterstrand, K. A. (2016). DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program. Retrieved December 19, 2017, from www.genome.gov/sequencingcostsdata

139. Wolpert, D. H. (1996). The Lack of {{\textbackslash}textitA} Priori Distinctions Between Learning Algorithms. Neural Computation, 8(7), 1341–1390. https://doi.org/10.1162/neco.1996.8.7.1391

140. Xu, Z., & Taylor, J. A. (2009). SNPinfo: Integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. Nucleic Acids Research, 37(SUPPL. 2). https://doi.org/10.1093/nar/gkp290

141. Yang, J., Ferreira, T., Morris, A. P., Medland, S. E., Madden, P. A. F., Heath, A. C., … Visscher, P. M. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nature Genetics, 44(4), 369–375. https://doi.org/10.1038/ng.2213

142. Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-Silva, D., … Flicek, P. (2016). Ensembl 2016. Nucleic Acids Research, 44(D1), D710–D716. https://doi.org/10.1093/nar/gkv1157

143. Yeung, K. Y., Bumgarner, R. E., & Raftery, A. E. (2005). Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. Bioinformatics (Oxford, England), 21(10), 2394–2402. https://doi.org/10.1093/bioinformatics/bti319

144. Zhang, B., Kirov, S., & Snoddy, J. (2005). WebGestalt: An integrated system for exploring gene sets in various biological contexts. Nucleic Acids Research, 33(SUPPL. 2). https://doi.org/10.1093/nar/gki475

145. Zhao, X., Yang, Y., Sun, B. F., Shi, Y., Yang, X., Xiao, W., … Yang, Y. G. (2014). FTO-dependent demethylation of N6-methyladenosine regulates mRNA splicing and is required for adipogenesis. Cell Research, 24(12), 1403–1419. https://doi.org/10.1038/cr.2014.151

146. Zhu, X. (2007). Semi-Supervised Learning Literature Survey. Sciences-New York, 1–59. https://doi.org/10.1.1.146.2352

147. Zipp, F., Ivinson, A. J., Haines, J. L., Sawcer, S., Dejager, P., Hauser, S. L., & Oksenberg, J. R. (2013). Network-based multiple sclerosis pathway analysis with GWAS data from 15,000 cases and 30,000 controls. American Journal of Human Genetics, 92(6), 854–865. https://doi.org/10.1016/j.ajhg.2013.04.019

148. Zuk, O., Hechter, E., Sunyaev, S. R., & Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. Proceedings of the National Academy of Sciences, 109(4), 1193–1198. https://doi.org/10.1073/pnas.1119675109