Special Issue

Social Science Computer Review
2021, Vol. 39(5) 768-801

DOI: 10.1177/0894439319888445
journals.sagepub.com/home/ssc
SAGE

Assessing Replicability of Machine Learning Results: An Introduction to Methods on Predictive Accuracy in Social Sciences

Ranjith Vijayakumar[1] and Mike W.-L. Cheung[1]

Abstract

Machine learning methods have become very popular in diverse fields due to their focus on predictive accuracy, but little work has been conducted on how to assess the replicability of their findings. We introduce and adapt replication methods advocated in psychology to the aims and procedural needs of machine learning research. In Study 1, we illustrate these methods with the use of an empirical data set, assessing the replication success of a predictive accuracy measure, namely, R^2 on the cross-validated and test sets of the samples. We introduce three replication aims. First, tests of *inconsistency* examine whether single replications have successfully rejected the original study. Rejection will be supported if the 95% confidence interval (CI) of R^2 difference estimates between replication and original does not contain zero. Second, tests of *consistency* help support claims of successful replication. We can decide apriori on a region of equivalence, where population values of the difference estimates are considered equivalent for substantive reasons. The 90% CI of a different estimate lying fully within this region supports replication. Third, we show how to combine replications to construct meta-analytic intervals for better precision of predictive accuracy measures. In Study 2, R^2 is reduced from the original in a subset of replication studies to examine the ability of the replication procedures to distinguish true replications from nonreplications. We find that when combining studies sampled from same population to form meta-analytic intervals, random-effects methods perform best for cross-validated measures while fixed-effects methods work best for test measures. Among machine learning methods, regression was comparable to many complex methods, while support vector machine performed most reliably across a variety of scenarios. Social scientists who use machine learning to model empirical data can use these methods to enhance the reliability of their findings.

[1] National University of Singapore, Singapore

Corresponding Author:
Ranjith Vijayakumar, Department of Psychology, Faculty of Arts & Social Sciences, National University of Singapore, Block AS4, Level 2, 9 Arts Link, Singapore 117570, Singapore.
Email: r.v@u.nus.edu

Keywords
machine learning, model comparison, predictive accuracy, psychological research, replicability

This article is part of the SSCR special issue on "Big Data in the Behavioral and Social Sciences", guest edited by Michael Bosnjak (Leibniz Institute for Psychology Information. Trier, Germany).

Machine learning methods, a diverse set of mathematical algorithms that focus on predicting patterns in the given data, are increasingly popular in many areas of science and policy, such as crime detection, prevention and reform (Berk, 2012; Bhattacharyya, Jha, Tharakunnel, & Westland, 2011; Ngai, Hu, Wong, Chen, & Sun, 2011; Zeng, Ustun, & Rudin, 2016), medical diagnosis and treatment (Bayati et al., 2014; Kukar, Kononenko, Grošelj, Kralj, & Fettich, 1999; Lavrač, 1999; Walsh, Ribeiro, & Franklin, 2017), financial forecasting (Dingli & Fournier, 2017; Lazzeri, 2018), and urban planning (Glaeser, Kominers, Luca, & Naik, 2015). Increasingly, social scientists are encouraged to use machine learning methods to supplement or replace traditional statistical methods. This comes at a time of increased scrutiny of unreplicated findings and questionable research practices in the social sciences (Gelman & Loken, 2014; Spellman, 2015).

This is mainly due to a traditional focus on the practice of statistical significance and unaccounted uncertainty in the process of design and analysis—termed the "researcher degrees of freedom" (Simmons, Nelson, & Simonsohn, 2011). It is felt that machine learning procedures can help mitigate this by forcing the researcher to explicitly take such practices into account. This is done by separating model estimation, model selection by comparing models, and estimation of the accuracy of the selected model—events that are never explicitly separated in traditional hypothesis testing (Yarkoni & Westfall, 2017). Machine learning accomplishes this by separating the data set into three—a *training* set to estimate the optimal parameters of each model under consideration, a *validation* set to compare and select among the competing machine learning models, and a final *test* set to get the estimate of the accuracy of the selected optimal model. Basing selection of the final model on predictions on a separate set is hoped to mitigate the problems of spurious research findings that bedevil current psychological research.

Another reason for the popularity of machine learning methods is the increased consensus that the phenomena of interest are usually parts of complex interacting systems (Forster, 2002; Varian, 2004). Most of the machine learning methods seek to capture nonlinear patterns in the data *without prior specification* by the researcher, unlike traditional regression methods. This ability, when coupled with the use of separate data sets for estimation and comparison to prevent overfitting to the sample, makes machine learning methods a convenient tool for complex data structures (Shmueli, 2010).

However, machine learning methods may not guarantee rescue from the problems concerning replicability in the social sciences. First, no method is superior to all others in every situation (Wolpert, 1996, 2001); much depends on the nature of the data, the search algorithms, and particular measures of accuracy used. Sometimes, traditional methods such as logistic regression are comparable or better than more complex methods in prediction (e.g., Dreiseitl & Ohno-Machado, 2002; Held, Cape, & Nathan, 2016; Steyerberg et al., 2014). Second, machine learning methods usually depend on larger sample size for superior performance compared to traditional methods (e.g., Van der Ploeg, Austin, & Steyerberg, 2014).

Perhaps most importantly, predictive accuracy may not by itself lead to better replicability in models. In previous research (Vijayakumar & Cheung, 2018), we found that machine learning methods that have better predictive accuracy do not always give rise to replicable models. In the presence of small predictor effects in empirical and simulated data, variables selected as important by machine learning models were not replicated more than those chosen by traditional regression methods, even when machine learning was superior to regression in prediction. So a major concern

for using machine learning in the social sciences is whether these methods give rise to replicable results and how to measure replicability in machine learning models.

The issue of replicability has surfaced in machine learning fields as well. Hutson (2018) notes that there is a "replication crisis" in artificial intelligence similar to the ones affecting psychology and other social sciences, mainly due to an inability to reproduce results in the presence of different training data, insufficient details about the algorithm code. Perhaps of more concern, even differences in random numbers used to kick off the training, and differences in hyperparameters that decide the learning rate, produce substantively different findings. Similar problems have been highlighted regarding machine learning performance in time series forecasting by Keogh and Kasetty (2003) and Makridakis, Spiliotis, and Assimakopoulos (2018). Drummond (2006) highlights problems in machine learning caused by reliance on benchmark data sets that may not be transferable to real-life data, comparisons using different hyperparameter tunings, overemphasis on null hypothesis testing to decide performance, and use of scalar performance measures that might hide the dependence of performance on the particular contexts of testing.

Given these concerns, the question remains as to how machine learning methods can be best used in social science research for robust findings. We will now briefly describe the steps taken by psychologists to examine replicability in their studies. After that, we introduce our adaptation of those replication methods for machine learning.

Replication in Psychology

Psychologists have been wrestling with replicability concerns for quite a while (Anderson & Maxwell, 2016; Hedges & Schauer, 2018). Many of the studies that have investigated replication used p values and significance thresholds of the replication study's effects to determine whether the replication was successful (Camerer et al., 2018). Methodologists have pointed out that this is not ideal. Maxwell, Lau, and Howard (2015) point out several of these flaws: The published effects sizes are a biased sample tending to overestimate the true effect; the sampling variability in the original study's effect size is not taken into account; moreover, failure to reject a null finding should not be taken as confirming a null finding.

Various solutions have been proposed (Hedges & Schauer, 2018; Lakens, 2017; Taylor & Muller, 1996; Yuan & Maxwell, 2005): Examining these proposals, Anderson and Maxwell (2016) detailed a series of possible goals of replication and proposed methods to realize these aims. We focus on three goals. First, to check if the replication effect size is *inconsistent* with the original effect size, the authors advocate constructing a confidence interval (CI) around the difference between the original and replication effect sizes. If the CI of this difference does not include zero, then the replication is inconsistent with the original study. Second, even if the CI contains zero, we cannot accept the null hypothesis; to test for *consistency* with the original, we need a different approach. Construct a *region of equivalence* around zero based on substantive considerations, if the CI of difference of effect sizes falls completely within this equivalence zone, one can conclude that the replication is clearly consistent with the original. A third aim combines replication and original effect sizes to get a population estimate of the effect size with an estimate of uncertainty using fixed or random meta-analytical approaches.

Replication of Machine Learning Results

Now, we look at how the procedures mentioned above can be used to estimate replicability of machine learning methods. Unlike traditional methods that look at measures of error or variance explained within the sampled data, machine learning methods use procedures like k-fold cross-validation (CV) to ensure that the model performance measures are not contaminated by overfitting. Here, the training sample is divided into k folds; then, one of the folds is kept apart for validation,

and the rest are combined to form the training set on which the models are fit. Then, the trained model is fit to the validation fold; this is repeated k times with alternating folds. Thus, we can select the best method by comparing methods on their CV performance; this avoids model overfitting. Note that CV estimates are not the optimal generalizable estimates; a separate test set is usually used to obtain the final population-level performance of the machine learning methods. In our study, we will compare replication with the original study using both CV and test measures.

So, machine learning uses performance measures to select models that can explain the most variance and thus lead to least generalization error in future. The hope is that the use of validation or test sets ensures greater reliability. This may not always be assured (e.g., Vijayakumar & Cheung, 2018); variables selected by models with the superior prediction may not always lead to better replication.

Researchers replicating machine learning findings in social sciences may have different motivations from the goals in artificial intelligence. Replications in artificial intelligence may focus on ensuring that identical codes lead to identical results or that tuning hyperparameters from the same range reproduces the estimated model and results (Drummond, 2006). However, researchers in the social sciences may be more interested in the conceptual reliability of the findings. If a researcher explains 40% of the variance in an outcome using a particular set of predictors, her main focus is about the relationship between variables. While she would wish to ensure that hyperparameters are tuned optimally, she would be more concerned about the robustness of the explained variance estimate across samples. This has substantive significance: Researchers could use the explained variance to establish the importance of their model for the policy or for persuading funding agencies of the practicality of their results.

Goals of This Article

We seek to adapt methods used in psychological research to deal with replication uncertainty in machine learning. We have two goals. In Study 1, we introduce three aims of replication, describe the appropriate replication procedures, and illustrate their adoption in machine learning with an empirical data set (from the European Social Survey [ESS, 2018], henceforth ESS data, detailed further below). We take samples from the same data set to form the original "study" and "replications," using these samples to then illustrate the replication procedures. In Study 2, we examine the diagnostic performance of these replication procedures by comparing their ability to distinguish replications from nonreplications. We manipulate the ESS data to ensure that we have a healthy mix of samples from populations with same and different effect sizes to create replications and nonreplicating samples and then run the replication procedures on these manipulated data to measure the diagnostic performance of the methods.

Study 1: Introduction and Illustration of Proposed Replication Methods

Following Anderson and Maxwell (2016), this study focuses on three aims of replication that are distinct but not mutually exclusive. The first two goals involve comparing the replication estimate with the original study. First, one could focus on whether the replication estimate falls within the margin of sampling error of the original study's estimate. If not, then the replication is considered *inconsistent* with the original study. (Note, however, that since one cannot accept a null hypothesis, a replication falling within the margin of error should not be judged consistent.)

Second, one could see if the replication can be *consistent* with the claim of a particular magnitude of predictive accuracy; the CIs of replication should fall entirely within a particular prior-decided interval based on research or policy interest. The researcher can determine a priori the interval of interest within which she wants the estimate to fall: This is called the *region of*

equivalence for the estimate. These two goals are different, when testing for inconsistency if a replication's point estimate lies within the margin of error of the original study, it does not automatically pass the second test of consistency. If the estimate has too wide a CI such that values can vary substantively and fall outside this region of equivalence, then the consistency of the original study's claim is suspect.

The third goal is not to compare single studies but to quantify the precision of the estimate in the population; here, we use the procedure of constructing meta-analytic CIs by combining studies. When comparing single replications with the original study to examine the consistency of replication, the strategy is to construct a CI for differences between estimates of the two studies. When combining replications and the original study for better precision, the focus is on constructing meta-analytic CIs. Our goals can focus on performance in both CV and test accuracy measures.

For all three aims, we need to quantify the uncertainty of the sample estimate in the presence of either multiple data points on that estimate (as in CV fold estimates of R^2) or single estimates in a study sample (as in test R^2). We detail these procedures in the Methods section. Then in the Results section, we illustrate how to use these procedures at appropriate stages to get the study standard error (*SE*) estimate, which can then be used to fulfill all three aims of replication.

We will use an empirical data set to illustrate the use of these replication procedures to compare different machine learning algorithms. The methods we compare are the following: (i) additive linear regression model, (ii) LASSO (*Least Absolute Shrinkage and Selection Operator*), (iii) Elastic Net, (iv) Multivariate Adaptive Regression Splines (MARS), (v) Random Forest, (vi) Support Vector Machines (SVM), and (vii) Neural Network (Hastie, Tibshirani & Friedman, 2016). Next, we detail the data sets and methods before moving onto results.

Method

Data

To illustrate the methods proposed below and to compare machine learning methods in their replicability, we use a subset of the available data sets from the ESS (2018). ESS is a cross-national survey conducted every 2 years across Europe. We use their survey data sets from Germany collected 2 years apart, from years 2002 (labeled "Round 1") to 2016 ("Round 8"). We combine all these years into a single data set ($N = 17{,}427$) with the assumption that the combined data are the population of interest. (Our motivation for this is detailed in the next section.)

For our illustration, we took 36 continuous variables as predictors; these include the following variables: The 21-item Human Value Scale modified by Schwartz from his earlier Human Values Scale (Schwartz, 1994, 2012); number of years of education; indices of subjective well-being such as happy, social activity, the time for social activities and company, and the level of intimate company; general trust toward others; level of expectation of fairness from others; and indices of satisfaction with nation and civic life such as government, democracy, economy, education and health, and their political alignment in the left–right dimension. The variable "happy" was chosen to be the outcome of interest, which other variables are used to predict. We chose these variables as they span important psychological and social dimensions which mirror the kind of variables social scientists and psychologists have to deal with. We were not interested in any specific theory but use these variables from this popular survey to illustrate the replication procedures of machine learning methods and to examine their performance.

Study Design of Replication Simulations

It might seem that the ideal way to show replications using a longitudinal data set is to select a random sample from say Round 1 as the original study and to use samples from the other rounds as

replications. This mirrors how replications are usually done in empirical research where the replication study takes place sometime after the original study. While an attractive use of longitudinal data in replication, this is not conducive for our purpose of illustrating replication procedures, where we need to ensure control over the status of true replications and nonreplications. If a succeeding round in a longitudinal study does not replicate, it could mean either that the method does not replicate or that the "effect" was not a stable pattern across time in a country. Replications depend on the underlying population being the same, and we wish to avoid uncertainty in what a nonreplication might mean for the method and the measured "effect."

Hence, we combine all the rounds, taking random samples of equal size from all the rounds to make a single population. Now, we create studies of equal size by randomly sampling from this population (henceforth ESS data). We then manipulate sampling from this population to conduct our two sets of simulations: (i) simulating effect size replicating samples to illustrate replication procedures in Study 1, which we detail below; and (ii) simulating effect size replicating and nonreplicating samples to compare replicability of machine learning methods in Study 2.

Simulating replications for illustration of replication procedures. We select eight random samples from the ESS data, each having a sample size of 2,000. The first sample is treated as the original study, and the following seven are considered replications. Since all study samples are taken from same population, they are true replications of the first sample denoted as the "original study." For each study, we separate the data into training and test sets, split randomly in a 2:1 ratio. We train the machine learning models on a training set using 10-fold CV and obtain predictive accuracy measures both on the CV folds as well as the test set. The main advantage of using as studies random samples from the same data set is that we know the replications are the same as the original study. The findings will thus illustrate how good the methods are in demonstrating replicability.

Machine Learning Methods

We use the open-source R statistical platform (R Core Team, 2019) for our statistical analyses. We used the following packages in R for analysis: *nnet* (Venables & Ripley, 2002) for neural network, *kernlab* (Karatzoglou, Smola, Hornik, & Zeileis, 2004) for radial SVM, *randomForest* (Liaw & Wiener, 2002), *earth* (Milborrow, 2007) for MARS, *glmnet* (Friedman et al., 2010) for LASSO and elastic net; and the native *lm* function for regression. The *caret* package (Kuhn, 2008) was used for preprocessing the predictors and model training and CV. For constructing meta-analytic intervals, we use the *meta* package (Schwarzer, 2007). The R codes used for our replication procedures are available in the Supplemental Appendix S2 (http://dx.doi.org/10.23668/psycharchives.2637).

Optimization of hyperparameters. Machine learning methods have method-specific parameters (called *hyperparameters*) that are user-determined and not optimized during the training of the model. We determine the optimal values for hyperparameters by *tuning*; we use a random search (Bergstra & Bengio, 2012) for tuning, except where grid search is appropriate for the method in question. For specific tuning parameters of each method, please refer to the Supplement, Appendix S1 (http://dx.doi.org/10.23668/psycharchives.2637).

Performance Measures

Among the commonly used measures of global predictive accuracy such as Root Mean Square Error (RMSE), Mean Square Error (MSE), and R^2, we use R^2 to illustrate our methods. We choose R^2 because statistical methods to quantify uncertainty in estimates of sample R^2 as well as differences between R^2s of different samples have already been developed, and there are well-established

procedures (Chan, 2008) that researchers can use for their replicability goals. While MSE can also be used by these procedures, R^2 might be more attractive for social science researchers as it is routinely used as a measure of the explanatory power of the models in the social sciences.

Methods to Quantify Uncertainty in R^2

In checking replication of R^2 for both CV and test subsets, we need to quantify uncertainty in R^2 and then use the derived *SE* to construct the CIs for the R^2 difference estimate between the original study and replication. Chan (2008) describes several ways to measure this, including two that we adopt for replicating machine learning methods: (i) bootstrap method and (ii) Olkin and Finn's method. Another way to quantify uncertainty is by adapting meta-analytic procedures.

Bootstrap method. We describe the bootstrap method here, but for our illustrative purpose, we will use Olkin and Finn's method detailed in the next section. For bootstrap, we randomly sample with replacement from the "study" samples to generate bootstrap samples of each study, train all machine learning methods on each bootstrapped sample, and obtain R^2s. In each iteration, the bootstrapped R^2 of the original study is subtracted from that of bootstrapped replication to get a distribution of bootstrapped differences. From this distribution, we construct CIs by the *empirical bootstrap method*. We calculate the actual difference between original and replication study R^2s, get the difference between this value (actual difference) and the bootstrap difference of each iteration, and then get the 95th percentile of this distribution to construct an interval around the original estimate (for various bootstrap methods and rationale for this method, see Davison & Hinkley, 1997; Rice, 2013).

Olkin and Finn's method. We will mainly use Olkin and Finn's (1995) asymptotically valid method in this study. To calculate *SE* of R^2 of a single sample, their formula is as follows:

$$SE = \sqrt{\left(R^2\left(1 - R^2\right)^2 4/N\right)} \tag{1}$$

where N stands for the number of data points contributing to the R^2.

For calculating CIs around R^2 differences for tests of inconsistency and consistency (described in the Results), we will use Olkin and Finn's simple asymptotic (SA) method (Zou, 2007). Following Chan (2008, p. 567), we construct a 95% CI:

$$\left(d - z_{(1-\alpha/2)}SE_d,\ d + z_{(1-\alpha/2)}SE_d\right) \tag{2}$$

Here $d = {R_1}^2 - {R_2}^2$, the difference between sample R^2 estimates; $z_{(1-\alpha/2)}$ is the $100(1-\alpha/2)$th percentile point of the standard normal distribution, and SE_d is the asymptotic standard error of d, calculated from the sum of variance of the independent sample R^2s, which are derived using Equation 1 above.

Note that apart from the bootstrap and SA methods, there is another method: modified asymptotic (MA) method by Zou (2007). The SA method has been shown to perform inadequately in small samples and unequal sample sizes (Algina & Keselman, 1999), while the MA method requires an iterative procedure instantiated in select statistical packages. Here, we use the SA method for ease of depiction but encourage researchers to consider the bootstrap method.

Meta-analytic methods. Units in our data are hierarchical (such as studies sampled from a population or CV folds within a study) where the lower level units are randomly sampled from a higher level. At the lower level, we can form *SE*s by bootstrap or Olkin and Finn's methods described above; at the

higher level, we form *SE*s by using meta-analytic procedures to combine units at the lower level, to provide the appropriate CIs.

Both fixed- and random-effects meta-analysis methods can be used; fixed-effects meta-analysis generalizes the estimates only to the exact studies involved, and thus their CIs are inaccurate in situations of effect size heterogeneity, while random-effects meta-analytic CIs assume that the studies are a random sample from a population of studies (Borenstein, Hedges, Higgins, & Rothstein, 2009; Cheung, 2015). We will illustrate both fixed- and random-effects meta-analytic approaches using the ESS data. We use the *meta* package to conduct meta-analysis in R.

In psychology and the social sciences, random effects meta-analysis is preferred for the purpose of combining studies as it is not obvious that the studies with differences in sampling units and design features are estimating the same effect (Borenstein et al., 2009). Similarly, since machine learning studies may differ in details of sampling units as well as algorithm-related measures such as hyperparameters, we suggest that the researchers consider using random-effects meta-analysis when combining estimates of different studies. In the Results section, we will detail specific stages in replication where meta-analysis can be used.

Results and Discussion

The specific procedures depend on the nature of the data. All procedures described further below depend on first deriving *SE* estimates of R^2 for each study using the methods described above. For CV folds with their fold R^2s, we first derive *SE*s of each single fold of a study using Olkin and Finn's methods and then use meta-analysis on fold estimates to form study-level *SE*s. For test data, we do not have folds, only a single test sample estimate of R^2. Here, we can derive *SE*s for the study sample using bootstrap or Olkin and Finn's methods. Once we obtain study-level *SE*s for CV or test data, these are then used to test single replications or construct population estimates using meta-analysis. We will first compare single studies and then combine studies to derive population estimates.

Single Replications

To compare single studies, we need to quantify the uncertainty in their R^2 estimates. This may depend on the information we have about the original study's data and the specific nature of CV as opposed to testing data.

CV R^2: Illustration of procedures. A researcher conducting a study with machine learning might know the original study's CV measures of accuracy. Each CV fold can give an estimate of R^2 so that a k-fold CV has k R^2 estimates. The variation of these fold R^2s does not take into account sampling error within each fold. We can use Olkin and Finn's equation 1 to obtain the *SE* of each fold's R^2. These individual folds can be considered as independent samples of the same study "population." So we can construct meta-analytic CIs around each study's CV R^2 estimate.

We illustrate this using the ESS data (see Figure 1). Figure 1 shows both the fixed- and random-effects meta-analytic intervals constructed using 10-fold CV R^2 for all eight studies. We first derive *SE* for each fold R^2 using Olkin and Finn's method equation (1) and then construct meta-analytic CIs for the CV R^2 of each study, using folds as samples and each study as the population.

The choice of fixed- or random-effects model is important. On the one hand, since all folds of a study are obtained through random sampling, a fixed-effects framework should be ideal. On the other hand, if some methods give more variability in their R^2 measures across folds, these methods may be less reliable in their CV R^2 estimates. Random-effects meta-analysis across the CV folds enables us to capture this heterogeneity in R^2 estimates.

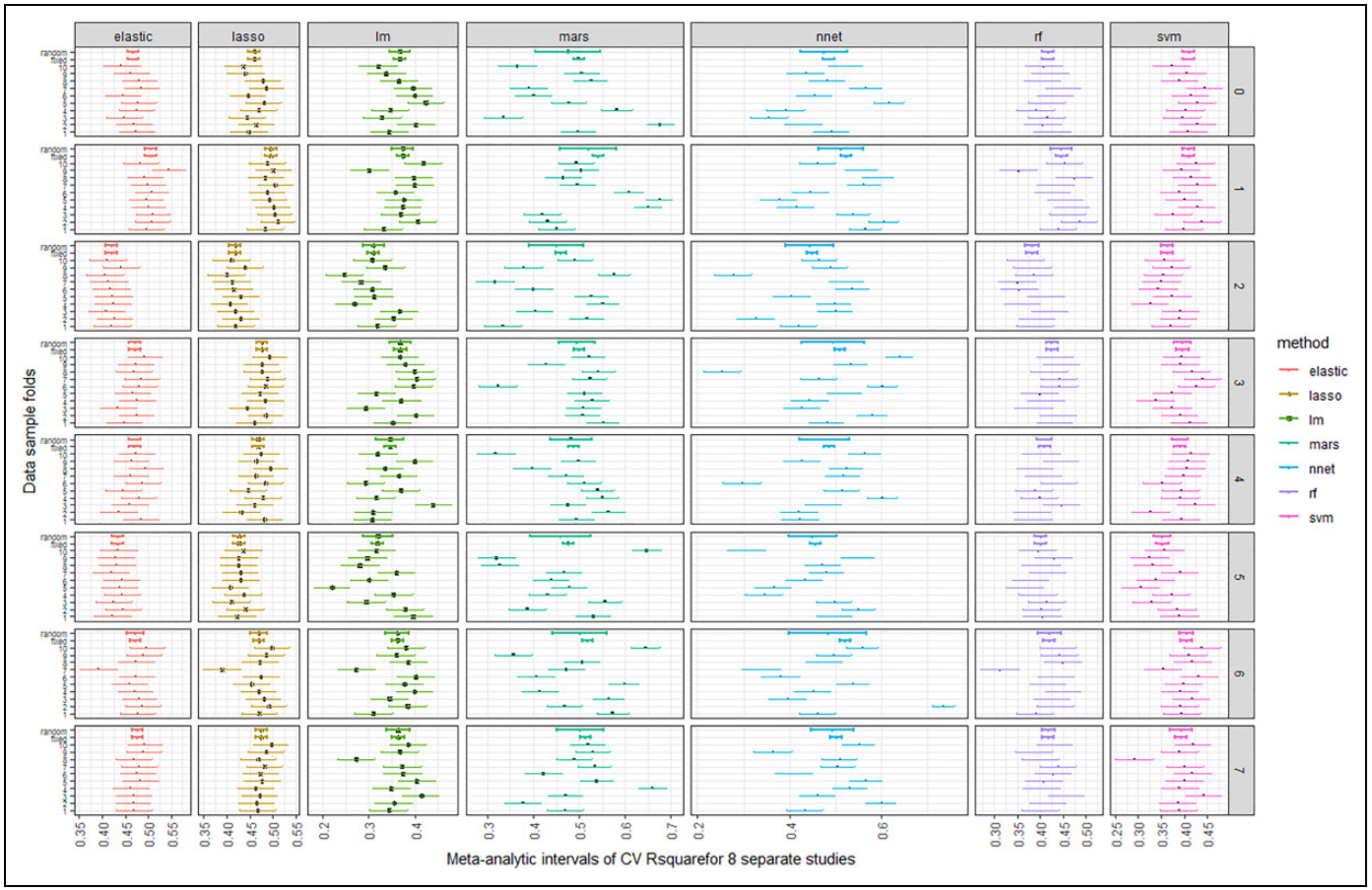

Figure 1. The meta-analysis-derived 10-fold cross-validation (CV) R2 estimates and standard error (SE) of each study depicted; the SEs of the R2 of individual folds were derived using Olkin and Finn's method. The R2 estimates of different methods are depicted separately. Note that these meta-analytic summary estimates will then be used as the mean and SE of CV R2 estimate for further analysis of consistency of replications. The terms "lm" and "rf" refer to regression and random forest, respectively.

Comparing R^2 point estimates, we see that both SVM and regression methods such as LASSO and elastic show little variation between folds. However, MARS and neural nets show great variation across folds of a study. This information is captured by random-effects meta-analysis but not the fixed-effects. Figure 2, which depicts only the study fixed- and random-effects intervals from Figure 1, shows that fixed- and random CIs differ in magnitude for methods such as MARS and neural networks and less for regression, but this is far less for SVM, LASSO, and elastic net.

When inferring replicability, a naive approach would be to judge replications based on the overlap between the CIs of individual studies. For instance, the random-effects CIs for regression (see Figure 2 which shows just the study-level CIs from Figure 1) may tempt us to conclude that all studies except Study 2 replicate the original CV R^2 estimate because their 95% CIs overlap with the original study CI. However, making such inferences based on graphical examination of 95% CIs lead to incorrect conclusions (Cumming & Finch, 2005), as estimates with overlapping CIs may be significantly different from each other.

Hence, we need to calculate the difference between the sample estimates of R^2 and then construct CIs around these different estimates. This allows us to both test for *inconsistency* of replications by using 95% CI of difference estimates and test for *consistency* of replication using 90% CI of difference estimates and the concept of *region of equivalence*.

Test for inconsistency. A statistically defensible method is to examine the difference between the R^2 estimates of two studies and construct a 95% CI around this difference estimate. If the 95% CI contains zero, we cannot reject the replication as a failure. However, if zero is outside the interval, then according to Anderson and Maxwell's (2016) criterion, one should consider the replication to be *inconsistent* with the original study. Note that this is an asymmetric decision: When CI excludes zero, we can conclude that the replication is inconsistent with the original; however, failure to reject the null cannot lead us to infer that the replication is consistent with the original.

In Figure 3, we show the result of constructing 95% CIs of the difference between R^2s of original study and each of the replications; the CIs of R^2 difference estimates in Figure 3A used individual study *SE*s constructed with random-effects meta-analysis of CV folds, while Figure 3B difference CIs used study *SE*s derived from fixed-effects meta-analysis.

We estimate the 95% CIs of the different estimates between R^2s of the original and replication studies; the CIs used *SE*s of individual studies that were constructed using either random-effects (Figure 3A) or fixed-effects (Figure 3B) meta-analysis of CV folds.

The random-effects method ensures that we can estimate the variability between methods in their estimation of CV R^2 differences, taking into account the between-folds variability that differ across methods (Figure 3A). Study 3 shows huge variability for complex methods such as SVM, neural networks, MARS, and random forest but less variability for regression and its regularized variants. Some other studies (Study 4) show high variability in R^2 estimates across methods. Yet other studies (Study 5) show high variability across most methods, but methods such as SVM and random forest seem immune. In our illustration, all studies are random samples from a population, and these variations reflect random chance effects. In such situations, fixed-effects meta-analysis might lead to spurious rejections of the null hypothesis due to the narrower interval. For instance, in our study, fixed-effects meta-analysis leads to most methods rejecting null hypothesis more than random-effects. Among the methods, random forest replications contain zero in their 95% CIs most often; neural nets lead to spurious rejections more often.

Test for consistency (Test of statistical equivalence). We saw above that the use of CIs around difference estimates could help us decide if a study is inconsistent with the original study. A researcher can go further and ask whether a replication is entirely *consistent* with the original study. Anderson and Maxwell (2016) suggest deciding on a *region of equivalence* around the population value of effect size, where the researcher would consider the effect to be equivalent for their

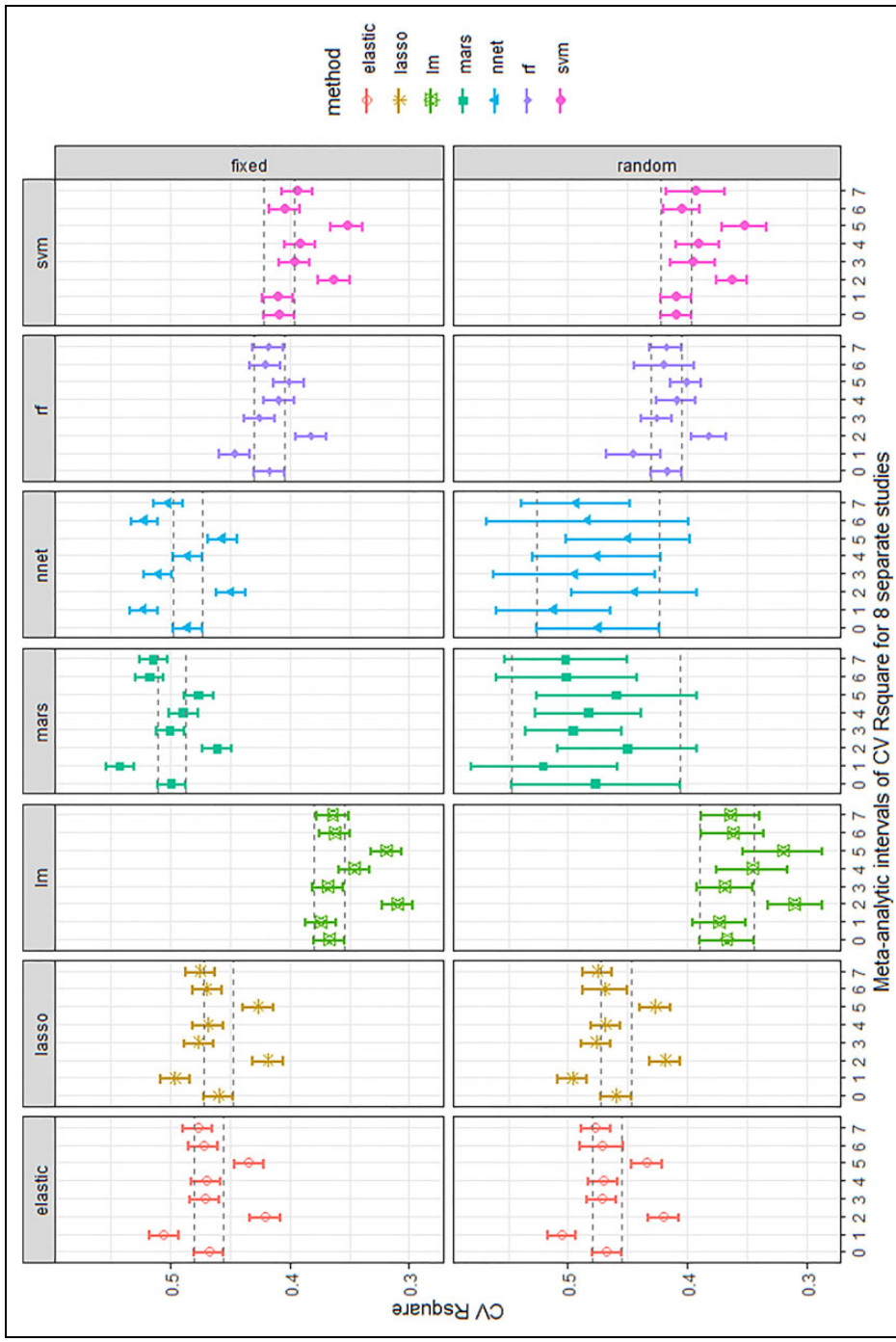

Figure 2. The fixed- and random-effects meta-analysis-derived 10-fold cross-validation R2 estimates of eight studies, labeled Study 0 to Study 7. The terms "lm" and "rf" refer to regression and random forest, respectively. The horizontal dashed lines indicate the 95% confidence interval of the original study (Study 0) for comparison with the seven replications.

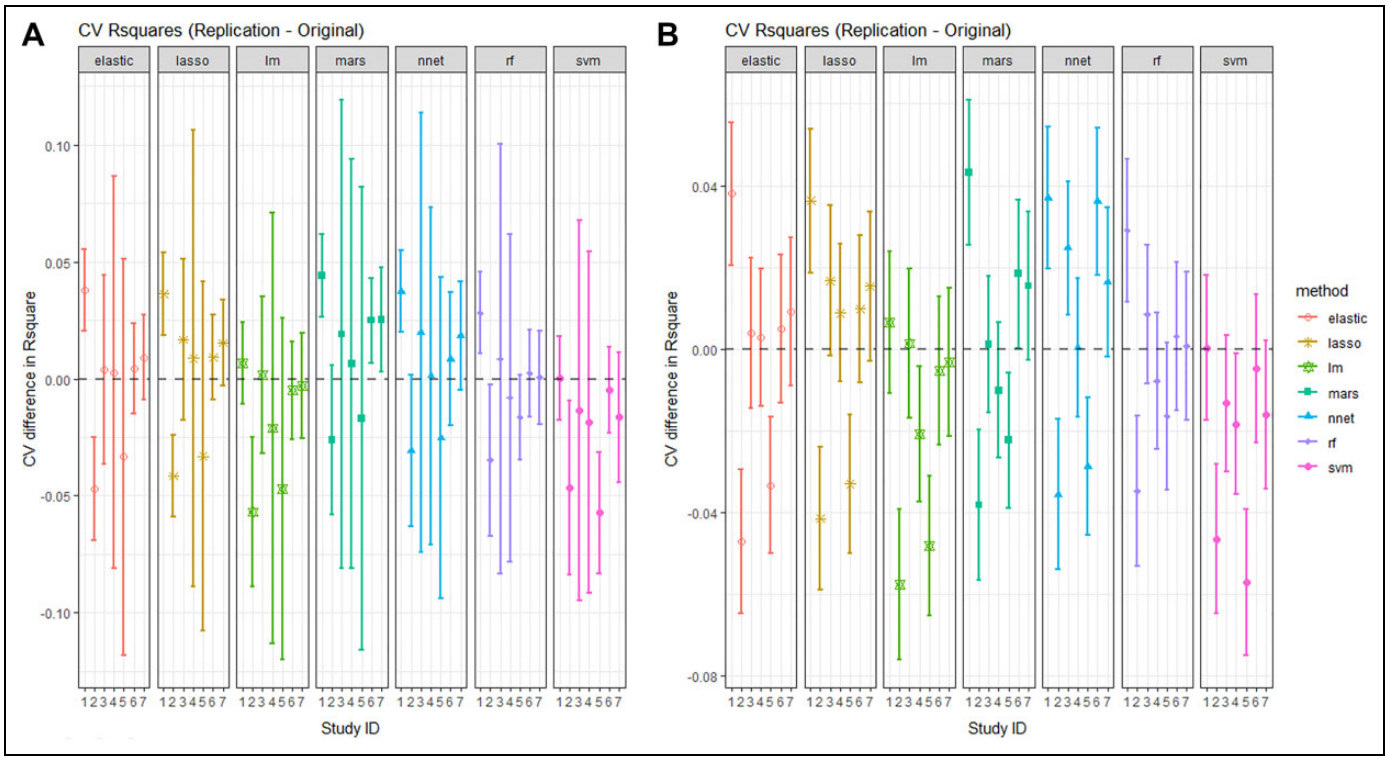

Figure 3. Test of inconsistency for R2 difference betweeen replication and original cross-validation (CV) estimates. The terms "lm" and "rf" refer to regression and random forest, respectively. The error bars depict the difference estimates and their 95% confidence intervals; if they do not contain zero, the study is judged inconsistent. The individual study CV estimates were derived by (A) a random-effects meta-analysis of the R2 estimate across the folds and (B) fixed-effects meta-analysis of the R2 estimate across the folds.

substantive purpose. Then, if the CI of the different estimates falls entirely within that region of equivalence, the replication is considered consistent with the original. The underlying idea is that failure to reject the null is not the same as inferring equivalence of the effect sizes. Note that Anderson and Maxwell (2016) advise to use a 90% CI and check whether it falls within the region of equivalence; the reason is that this corresponds to 2 one-tailed tests, each at $\alpha = .05$ (Walker & Nowacki, 2011). So these tests for consistency are also abbreviated as TOST (*two one-sided tests*). We use these terms interchangeably here.

Remember that when we tested for inconsistency by examining whether 95% CIs captured zero, intervals were wide enough to arouse skepticism about the ability of such methods to consistently replicate findings. To ask whether a replication is completely consistent with a study result, we now construct regions of equivalence around zero. Suppose a researcher has substantive or policy reasons to consider an R^2 difference of up to 5% to be irrelevant, she can consider an interval of width .05 around zero to be the *region of equivalence*. Constructing a 90% CI allows us to test the statistical equivalence (following Anderson & Maxwell, 2016), if the entire CI of the difference falls within this region, we can consider the replication to be consistent with the original.

The results show that using fixed-effects meta-analysis-derived study *SE*s (Figure 4B) to construct the R^2 difference score CIs allows most studies to show a high level of consistency across methods: Most of the 90% CIs lie entirely within the region of equivalence. When we use random-effects meta-analysis for study *SE*s (Figure 4A), most methods show poor consistency across replications: Due to the high variability in their estimates, their CIs rarely fall completely within the region of equivalence. For instance, for all methods except regression and SVM, the point estimates of R^2 difference of all studies fall within the region of equivalence (*TOST width*), but the variability in their 90% CIs ensured that most of these studies cannot be considered wholly consistent with the original study.

So when we examine true replications in this data, random-effects meta-analysis of CV fold R^2 estimates lead to better tests of inconsistency but fixed-effects are better for tests of consistency. The diagnostic performance of these replication procedures will be systematically examined in Study 2.

Test R^2: Illustration of procedures. CV measures of accuracy are poor indicators of accuracy as they are biased (Cawley & Talbot, 2010; Varma & Simon, 2006); test measures of predictive accuracy are thus a more reliable indicator of model performance. When a replication researcher has access only to reported results but not the original sample's test data, the results usually include the test R^2 value, which serves as a point estimate of the generalization error. The replication study has its own test subset and an accompanying test R^2 value.

The test set R^2 estimates for our ESS replications (Figure 5) showed comparable estimates for all methods, except two: random forest, which had lower estimates than most other methods comparing corresponding studies; and neural networks, which showed more disparity for test R^2 across the seven replications.

To see if these results are reliable in other samples of the same population, we need to obtain a CI around the difference scores of study R^2s, as in the earlier CV analysis. However, a test set yields only one R^2 estimate. So we can use Olkin and Finn's method (equation 2) or bootstrap to get the R^2 *SE*, which can then be used to estimate difference *SE* estimates. The *SE*s can be used to construct 95% CI to check for the tests of inconsistency (i.e., whether difference 95% CIs do not contain zero) and tests of consistency (the 90% CI for the TOST test for equivalence). We will use Olkin and Finn's methods for computational ease, but in empirical research, we encourage researchers to use bootstrap methods (Chan, 2008).

Tests for inconsistency and consistency. The results for tests for inconsistency and consistency with *SE*s derived using Olkin and Finn's equation 2 are shown in Figure 6A and 6B, respectively. Most

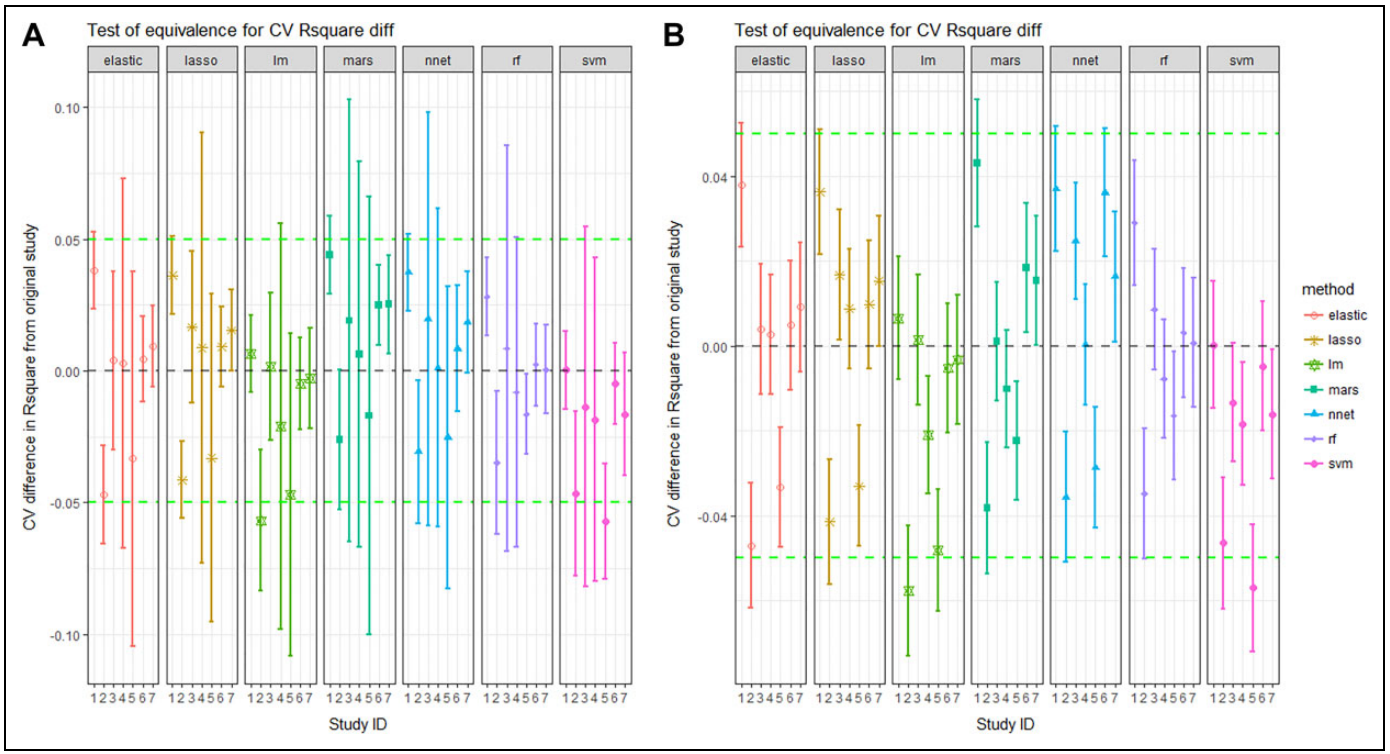

Figure 4. Test of consistency (statisitcal equivalence) for meta-analysis-derived difference estimates betweeen replication and original 10-fold cross-validation (CV) R2 measures. The terms "lm" and "rf" refer to regression and random forest, respectively. The error bars indicate the R2 difference estimates and their 90% confidence intervals. The green dotted line depicts an illustrative region of equivalence (an interval of .05 around zero) that contains population values considered equivalent for a particular research area. The individual study CV estimates were derived by (A) random-effects meta-analysis of the R2 estimate across the 10 folds and (B) fixed-effects meta-analysis of the R2 estimate across the 10 folds.

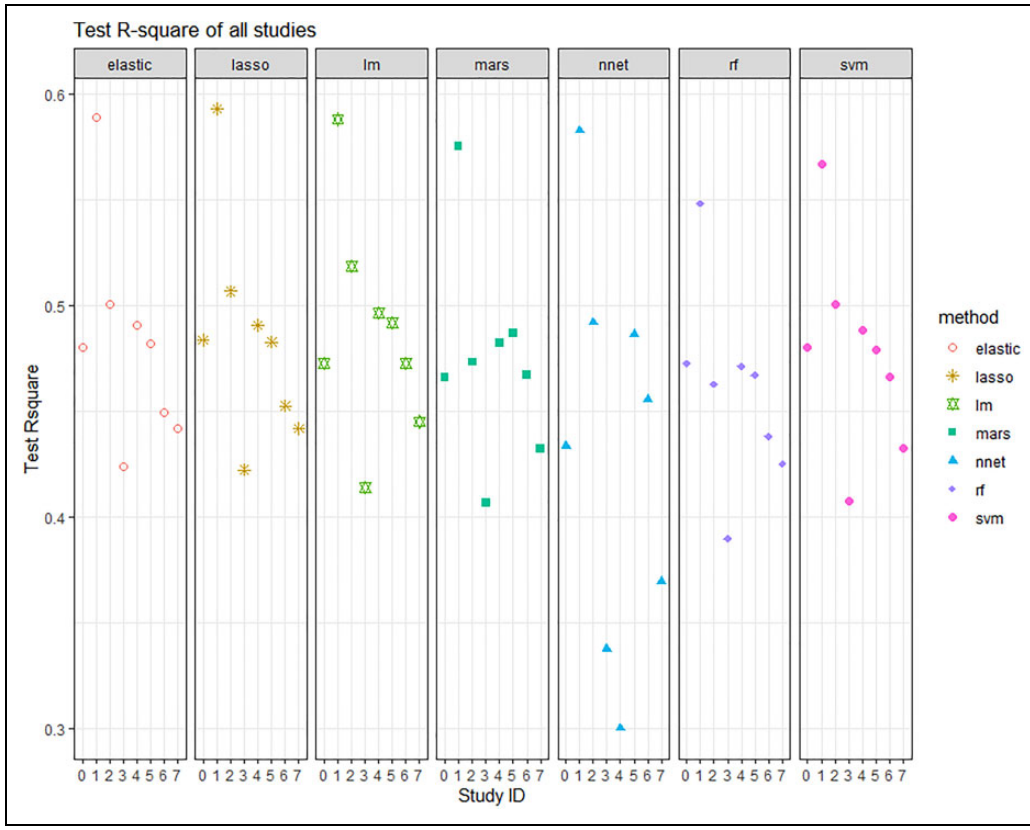

Figure 5. Test set R2 values for the eight studies in the European Social Survey study. The terms "lm" and "rf" refer to regression and random forest, respectively.

methods had most of their replications' 95% CI capturing zero and thus would not be incorrectly rejected as inconsistent. However, all of them also failed tests of consistency. Neural nets performed poorly on both consistency and inconsistency due to their wide discrepancy in test R^2 values across studies.

Combining Studies for Population R^2 *Estimates*

Apart from comparing single studies to the original, another possible goal of replication is to obtain a better understanding of population effect size (Anderson & Maxwell, 2016); this can be done by combining the original and replication studies as a "small" meta-analysis.

CV population estimate: Illustration of procedures. Here, the researcher must choose between fixed- and random-effects meta-analytic intervals on two successive steps: first, when she combines the CV fold R^2 estimates to get each study's CV R^2 estimate and second, when she combines individual studies to get a population estimate of the CV R^2 estimate. Regardless of her choice of meta-analytic interval for the first step, once she is faced with the task of combining studies, her choice depends on whether the studies can be assumed to come from the same population. If the researcher can consider all studies as estimating the same population, then she can use a fixed-effects meta-analysis.

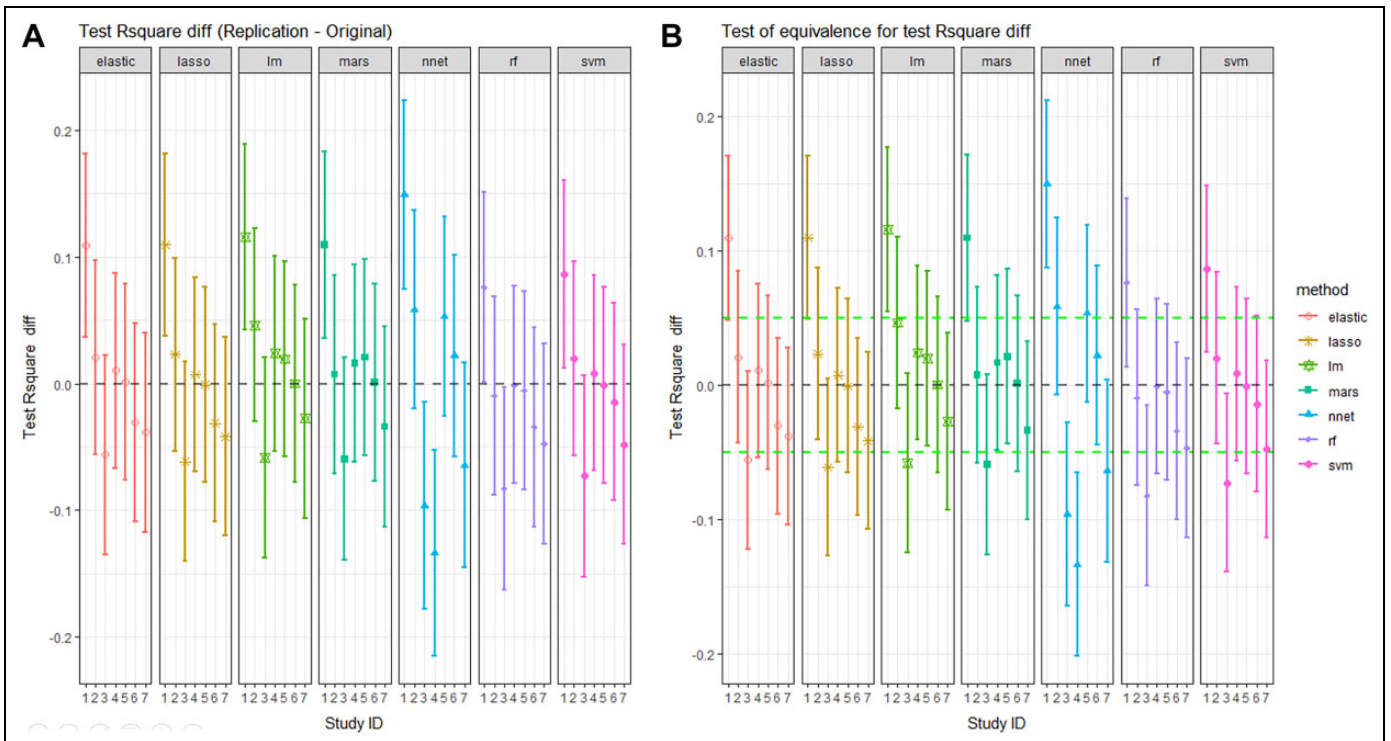

Figure 6. Test results of R2 difference scores between replications and the original study using Olkin and Finn's method. The terms "lm" and "rf" refer to regression and random forest, respectively. (A) Test of inconsistency: The error bars show the 95% confidence intervals (CIs); if zero is not contained in them, the replication is rejected as inconsistent with the original study. (B) Test of consistency on difference scores between original and replications; error bars show 90% CIs. Replication CIs falling entirely within this interval are consistent with original study.

Otherwise, she is trying to estimate not a single effect but the average of a distribution of different true effects and should use random-effects meta-analysis.

We depict the results of meta-analytic choices made at each of the two stages in Figure 7. We see that even in our data where single "studies" are random samples from the same population, fixed-effects meta-analysis done on studies capture fewer study estimates in their CI compared to random-effects meta-analysis regardless of the particular meta-analytic method used in the previous step to combine fold estimates to form study *SE*s.

Test population estimate: Illustration of procedures. We can use either Olkin and Finn's equation 1 or the bootstrap to derive *SE* estimates of each study and then construct fixed- and random-effects meta-analytic intervals to get population estimates. For illustration, we use Olkin and Finn's method to derive study-level *SE*s.

Meta-analytic results (Figure 8) show that random-effects meta-analytic intervals of neural nets are wider than that of regression-based methods; the others have comparable intervals. Regression has meta-analytic test R^2 estimates same or higher than complex methods such as SVM, MARS, and random forest, while neural nets have the lowest R^2 but the widest intervals. We cannot extrapolate from this to other data sets, but since the samples are randomly sampled from the same population, we can assert that machine learning methods are not appreciably better at replication than regression for this population, even though they lead to higher values of R^2.

Overall Summary

In conclusion, while most test replications captured zero in their difference CIs for all methods except neural nets, the lack of precision in their estimates make the replication not consistent with the original study, when using TOST. Replications in CV data seem to give better results across studies and methods; here, the choice between fixed- and random-effects meta-analysis to derive study-level *SE*s proved crucial. Fixed-effects meta-analytic-derived *SE*s lead to CIs that are narrow and may spuriously reject the null hypothesis, as happened in our study where replications are actually random samples from the same population. On the flip side, when testing for equivalence, fixed-effects meta-analysis-derived study *SE*s are narrower and give fewer spurious rejections. There is thus a trade-off between the two meta-analytic methods. Since CV folds come from the same study sample, it makes conceptual sense to use a fixed-effects meta-analysis to derive the study *SE*s from individual folds.

Comparing methods, we find that regression performs similarly to other methods in replication, with random forests performing the best. MARS and neural nets lead to wide variation in CV fold estimates that are reflected by inconsistent replications for CV R^2; neural nets also show wide variation in test estimates leading to poor performance in replication.

In Study 1, we adapted replication procedures used in psychology along with meta-analytic methods to illustrate assessing the replicability of machine learning methods. We used samples randomly selected from a single population to stand as studies; so these are indeed true replications. To be effective, the replication procedures should help us distinguish between true replications and nonreplications; next, we use a small study manipulation to the ESS data to explore this.

Study 2: Examining Diagnostic Performance of the Proposed Methods

A procedure that measures replicability must help us in not just capturing true replications but also in separating replications from nonreplications. We seek in our methods two diagnostic qualities: It should have a high score of true positives (i.e., the fraction of true replications accurately sensed: its *sensitivity*) and a low score of false positives (i.e., fraction of nonreplications that it misjudges to be

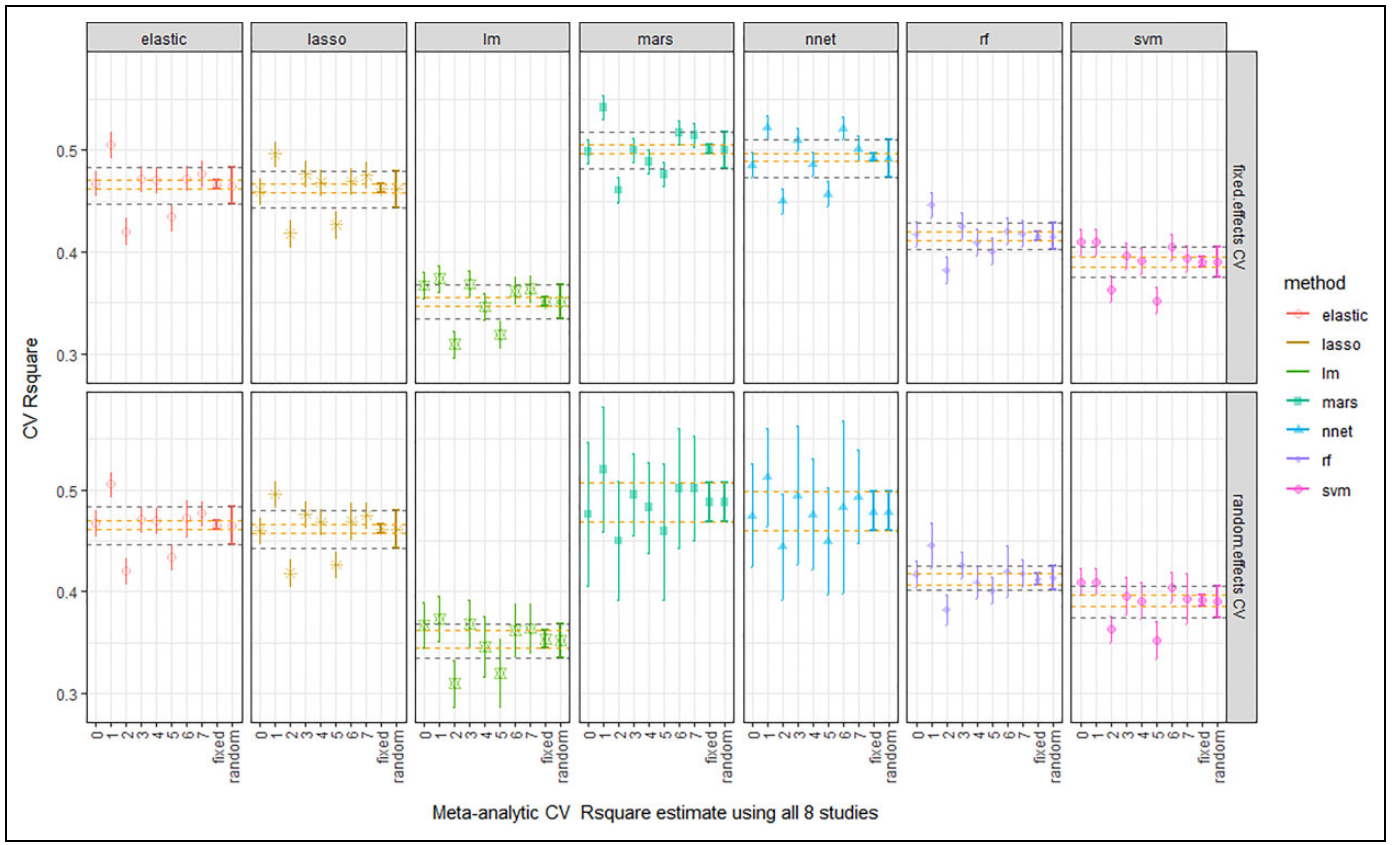

Figure 7. Fixed- and random-effects meta-analytic population cross-validation (CV) R2 estimate based on eight studies across machine learning methods. The terms "lm" and "rf" refer to regression and random forest, respectively. All error bars depict 95% confidence intervals (CIs). First, within each study, CV fold R2 estimates were combined to form study-level standard error (SE) estimates using either fixed-effects meta-analysis (top row) or random-effects meta-analysis (bottom row). Then, the studies were combined to form population level fixed- and random-effects meta-analytic intervals. These population-level meta-analytic CIs are shown side by side along with individual study SE-derived CIs. The orange dashed lines indicate random-effects meta-analytic population CI; the black dashed lines show fixed-effects CI.

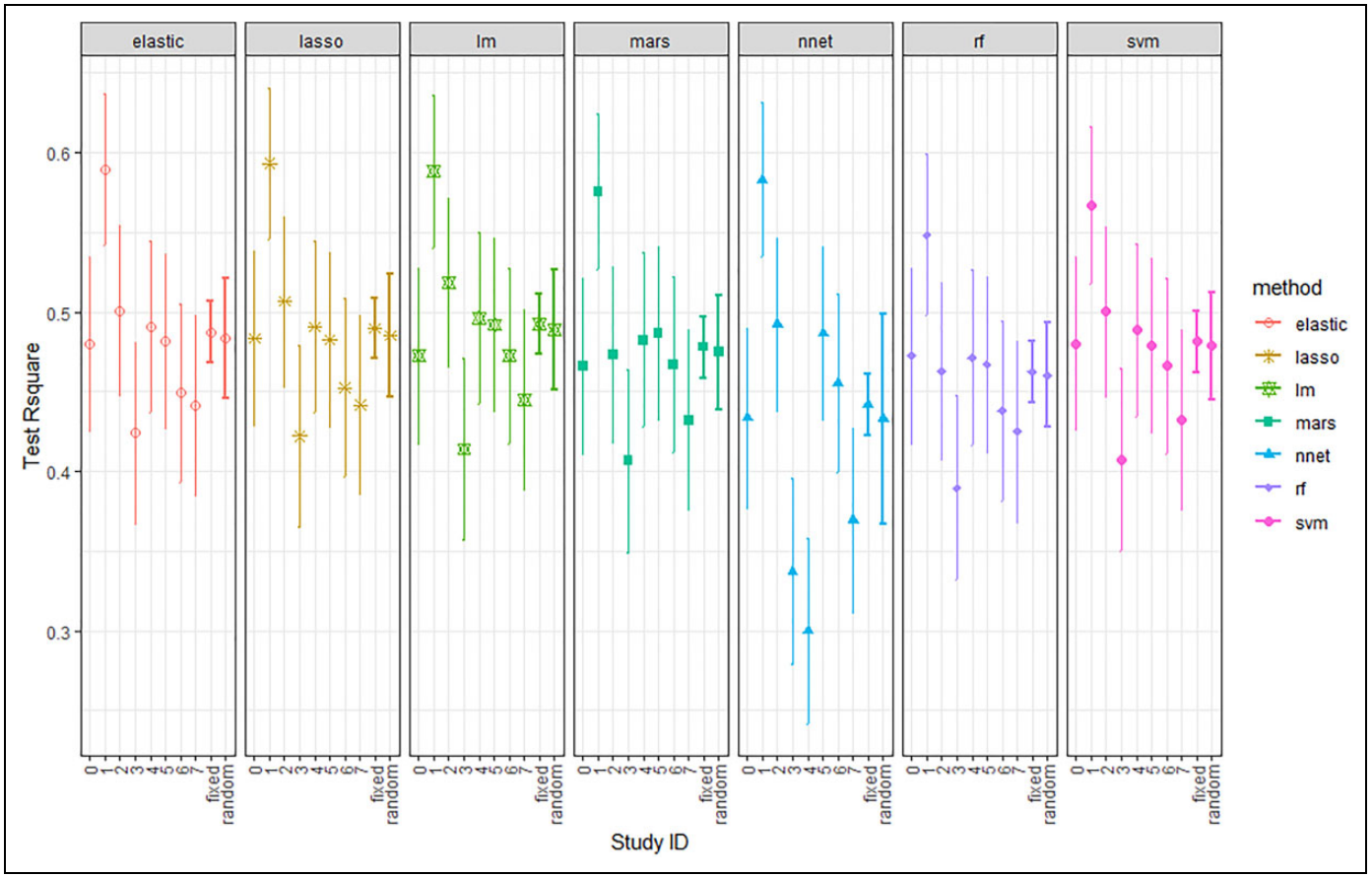

Figure 8. Fixed- and random-effects meta-analytic confidence intervals (CIs) for test R2 estimates of the eight European Social Survey study samples. Error bars depict 95% CIs, based on bootstrap-derived standard errors of individual studies. The terms "lm" and "rf" refer to regression and random forest, respectively.

true; this is often called *specificity*, adjudged as the frequency of correct rejections). In Study 2, we have two relating aims. First, we examine the diagnostic performance of fixed- versus random-effects meta-analysis we use at various steps of replication (introduced in Study 1). Second, we wish to explore how to assess machine learning methods' replicability across studies. This would help to better contextualize our suggestions regarding replicability of machine learning methods. For these purposes, we will use manipulated samples that show known deviation from the original sample; we then see how the different procedures and machine learning methods perform in capturing these deviations using the tests we already described for CV and test R^2 estimates in Study 1.

Method

Data and Simulation Design

The ESS data used in study 1 serve as the template data. To recap, the final data set contained eight random samples denoted as Studies 0–7, with study 0 designated as the original study. We will create scenarios where a subset of these eight studies will come from populations that have added error and thus reduced R^2s. We manipulate two factors: the amount of R^2 change between studies and the number of studies with changed R^2. The changed R^2 scenarios are R^2 changes of .05, .1, .15, .2, .3, and .4. This will help us in assessing replicability of single replications. We also manipulate the number of studies with this changed R^2: either four studies of the eight (the samples labeled Studies 4–7) or two studies (the samples labeled Studies 6 and 7). This will help us assess the impact on population-level meta-analytic estimates for the various methods. We will detail the steps taken to generate the desired R^2 below.

To change the R^2, first calculate the error term to add to the data based on the following equation regarding the variance components in the population: $R^2 = \sigma^2_{\text{model}}/\sigma^2_{\text{total}}$

where σ^2 denotes variance partitioned into the fitted model and remaining error:

$$\sigma^2_{\text{total}} = \sigma^2_{\text{model}} + \sigma^2_{\text{error}}$$

Now when we reduce the R^2 to a new value, R^2_{new}, this would be the result of adding to the old error term an added error (error_{change}) with a certain variance which we denote σ^2_{change}:

$$R^2_{\text{new}} = \sigma^2_{\text{model}}/\ \{\sigma^2_{\text{model}} + \sigma^2_{\text{error-old}} + \sigma^2_{\text{change}}\}.$$

We can thus compute the σ^2_{change} to be added to the error term, based on the desired change in R^2. An error term with this variance σ^2_{change} is then added to the outcome variable in the study subset:

$Y_{\text{new}} = Y_{\text{old}} +\ \text{error}_{\text{change}},$ where $\text{error}_{\text{change}} \sim N(0, \sigma^2_{\text{change}})$, and Y is the outcome variable in the study sample. This generates a subset of studies sampled from a population with the specified change in R^2.

Machine Learning Methods

The same as in Study 1.

Procedures

We use the replication procedures introduced in Study 1 on both CV and test R^2 estimates. Results and Discussion follow.

Results and Discussion

We examine the performance of machine learning methods in assessing replication by focusing on their sensitivity to the change in R^2 in a subset of studies. We give the results on testing single CV and test replications and then the results of population-level meta-analysis combining all studies. For diagnostics, since we compare multiple choices of replication procedures and manipulated factors in each test, we will summarize the sensitivity performance over the range of R^2 changes by focusing on the proportion of true versus nonreplications captured by these procedures.

We will first compare single studies and then combine studies to derive population estimates.

Single Studies

When comparing single studies, we use the scenario where four of eight studies have reduced R^2.

CV R^2: diagnostics. Our aim is to compare the diagnostic performance of CIs formed by fixed- or random-effects meta-analytic methods across folds to construct study-level R^2 *SE*. We determine this by measuring the difference in proportions of replications and non-replications captured by these tests.

Tests of inconsistency. Figure 9A shows the difference between replications and nonreplications in the proportion of inconsistent studies. When fixed-effect meta-analysis is used to combine CV fold R^2s, the study CIs always excluded zero for nonreplications, so they were correctly rejected as inconsistent; however, two thirds of replications were also rejected as inconsistent; this occurred for all machine learning methods, except regression and SVM where only 25% of replications were rejected. Using a random-effect meta-analysis in forming study *SE*s resulted in the gap between nonreplication and replications widening, with replications now judged inconsistent only one third of the time, except for random forest, LASSO, and elastic which continued to incorrectly reject two thirds of replications. This suggests that for this data set and small set of replications, random-effects meta-analysis to combine CV fold *SE*s leads to better diagnostic performance. Comparing machine learning methods, we find that LASSO, elastic net, and random forest are least reliable as both nonreplications and replications are judged inconsistent frequently. Regression and SVM are best performers across both fixed- and random-effects methods, and MARS and neural nets perform well for random-effects.

Tests of consistency. To test for consistency, we compare fixed- and random-effects meta-analysis constructed study-level CIs and use values of .05, .075, .1, .15 for the TOST interval within which a study is declared to be consistent with the original estimate. If we subtract the proportion of nonreplications judged as consistent from the proportion of consistent replications, we can get a diagnostic measure of consistency. Figure 9B shows that for additive models such as regression, LASSO, and elastic net, random-effects performed almost as well as and sometimes better than fixed-effects meta-analysis in distinguishing replications from non-replications, especially when the region of equivalence for R^2 was more than .05; this effect occurred irrespective of the R^2 change in replication from the original. For complex methods like SVM, random forest, and neural networks, fixed-effect *SE*s were better at diagnosing consistency if purported interval for equivalent R^2 is up to .075; but once TOST width increased to .1, random-effects intervals performed as well as fixed-effects except for MARS which worked better with fixed-effecst for most of the R^2 and TOST ranges.

The exact recommendation of meta-analytic procedure for particular machine learning methods thus depends on the precision of R^2 in the field; for predictor effects that allow for equivalence up to .05 in R^2, complex methods—SVM, MARS, neural nets, and random forests—along with regression

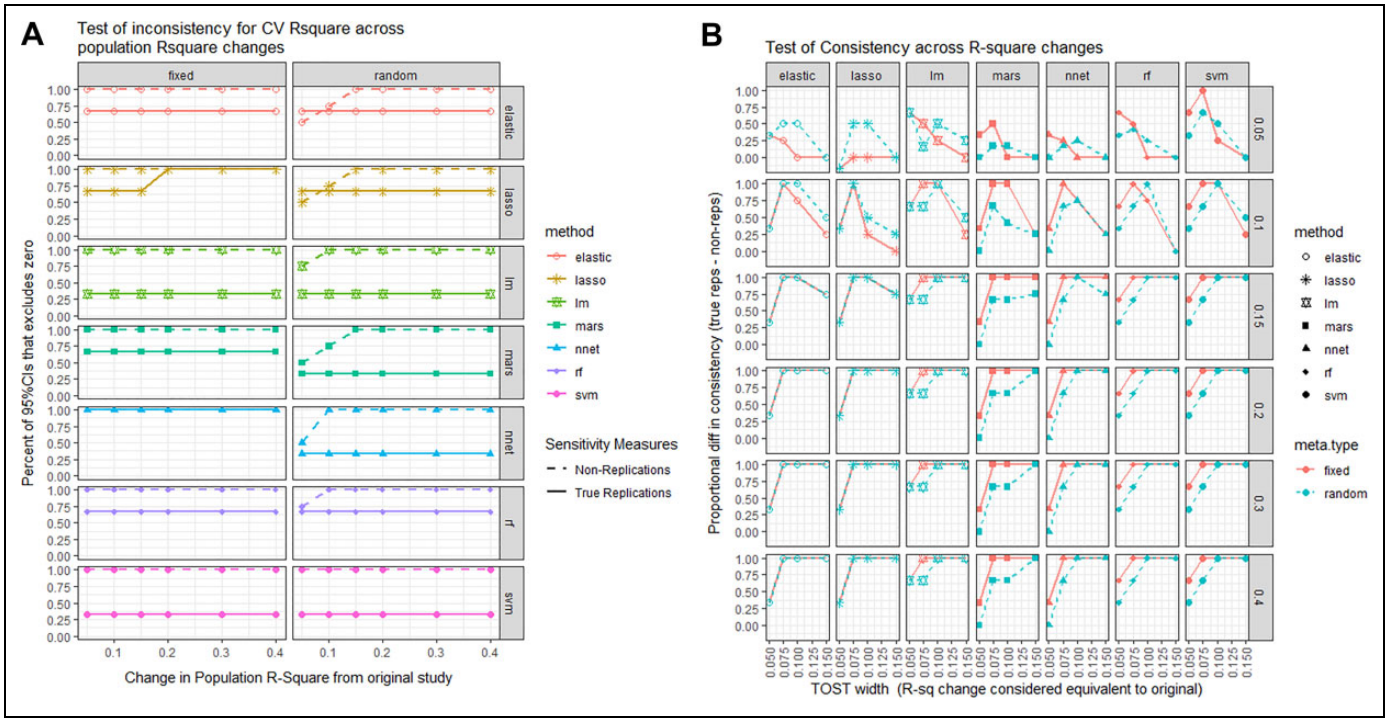

Figure 9. Cross-validation R2 measures. Comparison of replications and nonreplications using tests of inconsistency and consistency across R2 changes of .05, .1, .2, .3, and .4 from the original study. The terms "lm" and "rf" refer to regression and random forest, respectively. (A) Test of inconsistency: The proportion of studies with their 95% confidence intervals (CIs) in their R2 difference scores (Replication study – Original) that do not capture zero; such studies would be considered inconsistent with original. The lines show proportion of true replications and non-replications judged to be inconsistent with original. (B) Tests of consistency: Comparing performance of study CIs formed by fixed- versus random-effects meta-analytic methods across purported regions of equivalence with R2 width of .05, .075, .1, and .15. The lines represent difference between proportion of true replication and nonreplications and are separately shown for fixed- and random-effects meta-analytic standard errors.

perform better using fixed-effects meta-analysis than random-effects, while regularized regression-based methods (LASSO and elastic net) perform better using random-effects meta-analysis. For fields where R^2s are less precise (TOST width greater than .1), most methods except MARS perform better with random-effects meta-analysis. Regarding algorithms, regression performs comparably to more complex machine learning methods, with only SVM using fixed-effects meta-analysis outperforming regression, and that too only for lower ranges of TOST and R^2.

Test set R^2: Diagnostics. We use the study design for diagnostics as mentioned earlier, with an altered subset of studies with reduced R^2 ranging from .05 to .4. We then estimate study *SE*s across these scenarios using Olkin and Finn's equation 2 and then get the CI of difference scores (replication − original) for use in tests.

Tests of inconsistency and consistency. Figure 10A shows the results of tests of inconsistency for difference scores of the test data, and Figure 10B depicts results for test of consistency.

All machine learning methods accurately judge nonreplications as inconsistent especially when the R^2 change is .1 or more, except neural nets, which needed R^2 changes of .15 to reject all nonreplications. Regarding replications, results vary among methods: Neural networks and random forests perform poorly incorrectly rejecting two third of replications, while both regression, Elastic and LASSO perform well, similar to SVM and MARS. Overall, we can surmise that all methods perform equally well in judging inconsistency, with only neural networks and random forests performing worse.

We test consistency across TOST widths ranging from .05 to .15. When the TOST width is .05, all methods fail to distinguish replications from nonreplications. As the TOST width increases to .075, MARS starts to distinguish one third of replications as consistent, while other methods still fail. For a majority of replications to be judged accurately as consistent, TOST width needs to be .15 or more in this data set. Here, regression and most other methods perform well, with random forest and SVM performing the best. Neural nets perform poorly across all TOST values.

Combining Studies to Form Population R^2 Estimates

We explore two issues here. The first issue is examined for both CV and test set R^2s: the effect on population CIs of including nonreplications by reducing R^2s in either two or four studies of the total eight. We prefer methods that give more weight to the R^2 found in more studies and are less influenced by a few outlying estimates. At the same time, as studies increase with different R^2s, we would want the meta-analytic intervals to capture this imprecision. Since it is difficult to picture all variables here, we index this by the proportion of true replications and nonreplications whose 95% CIs overlap at least partially with the population level meta-analytic interval. If a study's CI is completely outside the population meta-analytic interval, we judge that study to be not captured by the meta-analytic CI. The second issue specifically concerns choices of meta-analytic procedures in CV R^2 analysis detailed in the next section.

CV population estimate: Diagnostics. Here, the choice of fixed- or random-effects meta-analytic methods occurs at two steps: (i) the combination of CV folds into a study-level *SE* and (ii) the combination of study *SE*s into population-level *SE*. We focus on how these choices affect the classification of replications and nonreplications. Results are shown in Figure 11.

When two studies have reduced R^2 (Figure 11A and B), fixed-effects meta-analysis at the population level to combine study estimates perform poorly in capturing either replications or nonreplications; the only exception is MARS and neural nets which perform better at low to moderate R^2 changes, provided random-effects was used to combine CV folds within each study.

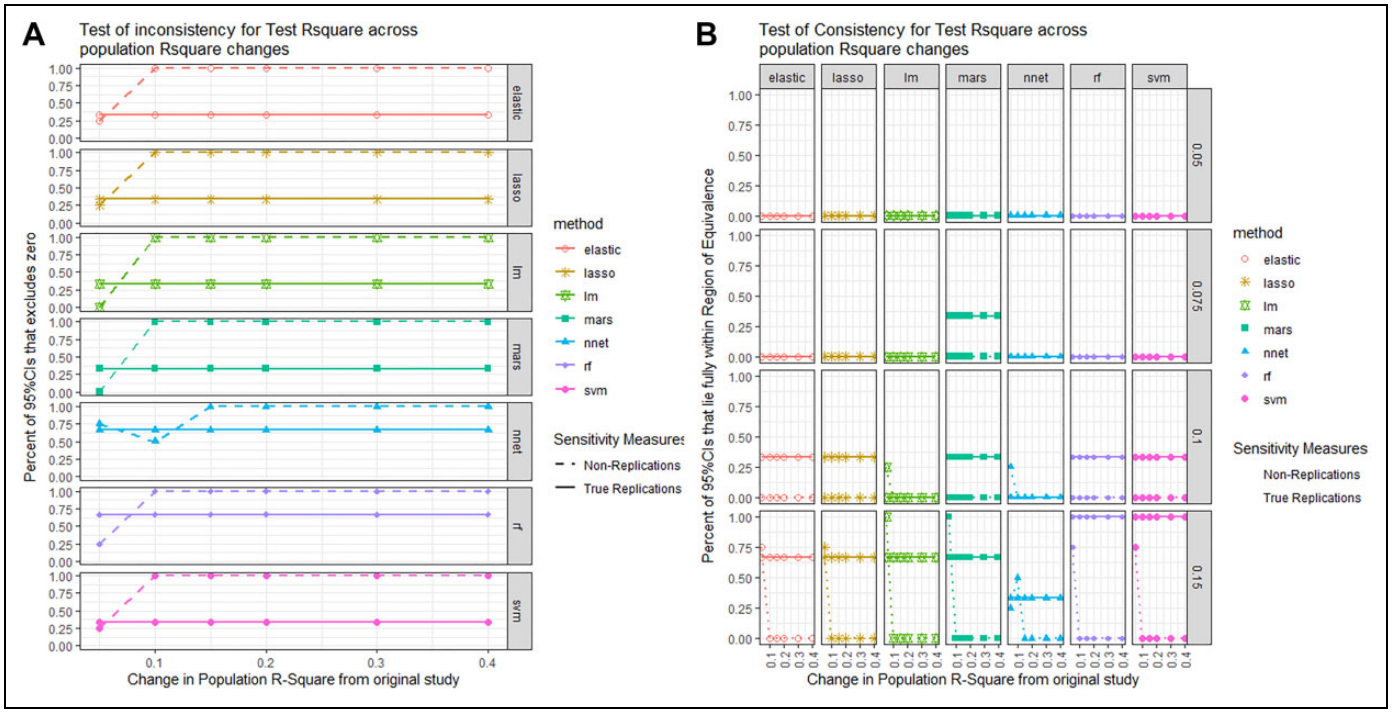

Figure 10. Test R2 measures: Comparison of machine learning methods in distinguishing replications and nonreplications using tests of inconsistency and consistency across R2 changes of .05, .1, .2, .3, and .4 from the original study. The terms "lm" and "rf" refer to regression and random forest, respectively. (A) The proportion of inconsistent studies: studies with their 95% confidence intervals in their R2 difference scores (Replication − Original) that do not capture zero. (B) Comparing methods in tests of consistency across purported regions of equivalence with R2 width of .05, .075, .1, and .15.

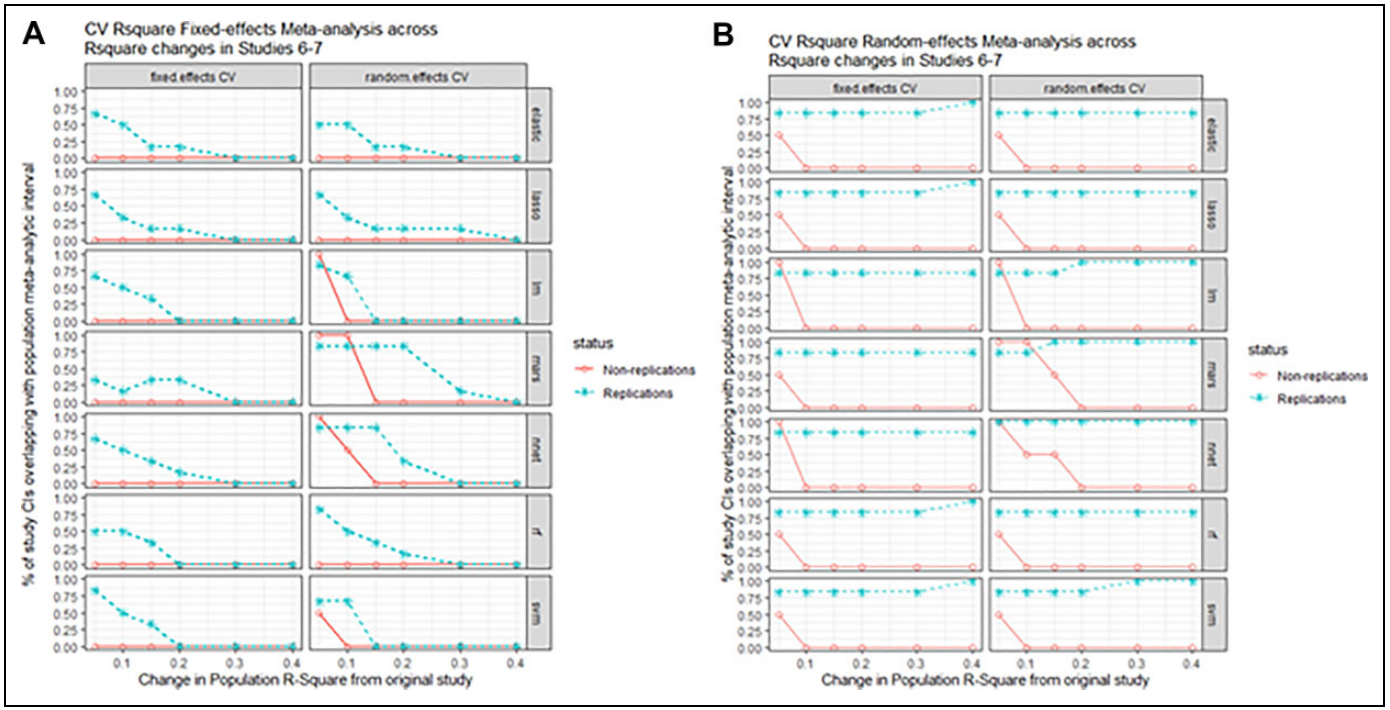

Figure 11. Population-level meta-analysis for cross-validation (CV) R2, depicted using four graphs. The rows show results when R2 changes occurred for two studies (top row: A and B) and four studies (bottom row: C and D). The x-axis shows the actual population R2 changes in the nonreplication studies from the original study. The lines depict the proportion of studies whose 95% confidence intervals (CIs) are at least partially overlapping with the CI of the population R2 estimate. In each graph, separate columns (fixed-effects CV and random-effects CV) denote the meta-analytic method used to combine folds to get each study's standard error (SE) in Step 1. The two figure columns depict the meta-analytic method used to combine study SEs to form population estimates—fixed-effect (left column: A and C) and random-effect (right column: B and D).

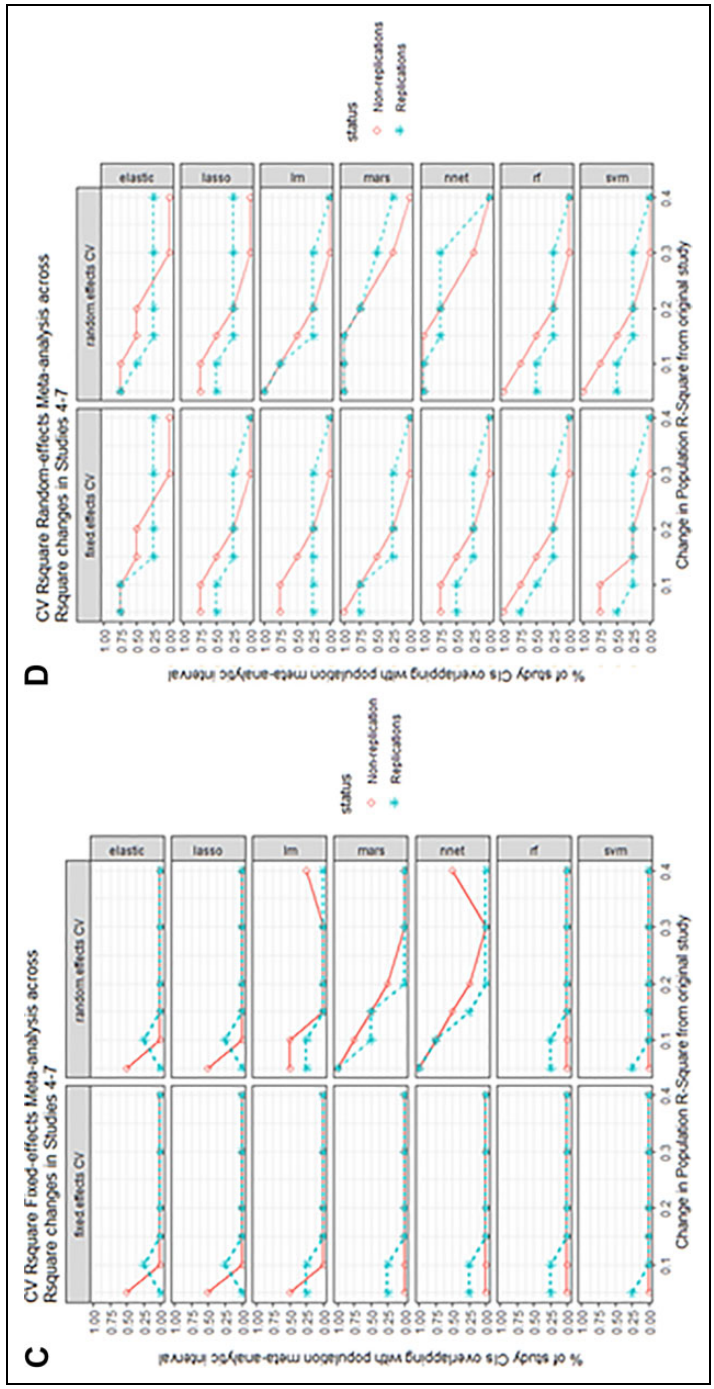

Figure 11. (continued).

Random-effects meta-analysis done at the population level fare best in distinguishing replications from nonreplicating outliers, especially when fixed-effects meta-analysis is used within each study to combine CV folds.

As R^2 reduces in four of eight studies (Figure 11C and 11D), we want the meta-analytic method to capture this imprecision by including more study estimates. Again, fixed-effects meta-analysis to combine studies performs poorly across almost all methods with neural network and MARS performing better than others, while random-effects meta-analysis performs well by capturing most studies even as R^2 reduces by a large amount. For MARS and neural nets, combining CV folds in a study by using random-effects meta-analysis works as well or better than fixed-effects as R^2 discrepancy increases; for other methods, both ways of combining CV folds perform similarly. Note that MARS and neural nets also showed wide variation across their CV folds in study 1.

Hence, we can tentatively surmise that at the population level, it is almost always preferable to combine studies using random-effects meta-analysis. At the study level, recommendations depend on expected precision of the method or data: when R^2s in a field are more precise, it is better to use fixed-effects to combine the CV folds; but if the studies have widely different R^2s or if the machine learning method is known to give highly discrepant CV fold R^2, it may be better to use random-effects meta-analysis to combine the CV fold R^2 as well.

Test population estimate: Diagnostics. Results are shown in Figure 12.

In test data, we find a pattern different from CV data. First, the meta-analytic intervals seem to be influenced by nonreplications even with only two studies undergoing R^2 changes. With fixed-effects meta-analysis, when two studies have reduced R^2, the meta-analytic intervals capture true replications better than nonreplications. When more studies have R^2 change, the fixed-effects methods lead to estimates and CIs covering few of the study CIs, which is expected given the smaller width of fixed-effects. Random-effects methods are unduly swayed by R^2 changes even when only two studies have reduced R^2. Overall regression seems to perform almost as well as more complex methods, with all methods giving a more acceptable performance using fixed-effects meta-analysis in our simulation where studies estimate same population R^2.

Overall Summary

In summary, comparing CV R^2 for single studies, random-effects meta-analysis is better at distinguishing between replications and nonreplications on tests of inconsistency and performs better at tests of consistency if R^2 precision in the relevant field assessed by region of equivalence is less (in our illustrative study, for TOST intervals greater than .075). For precise R^2 equivalence regions, fixed-effects meta-analysis performs better for most methods. Overall, it seems a reasonable approach to use random-effects meta-analysis to combine CV folds in each study before making comparisons.

Among machine learning methods, for tests of inconsistency, regression seems to perform favorably or at par with complex methods such as MARS. Among machine learning methods, SVM closely followed by regression performs best in tests of consistency and inconsistency over a wide range of R^2 and TOST values. Random forests give uneven results for CV and test data: For both CV and test data, it performs well for some ranges of R^2 changes in tests of consistency, but poorly on inconsistency. Neural nets perform well on CV tests of inconsistency and consistency, but badly for test data. Overall, SVM performs best on both CV and test data.

For the population-level meta-analysis, random-effects meta-analysis is preferred for combining studies in in CV data, but fixed-effects are better for test data. This may have to do with the difference between CV and test data. The CV folds of each study are resamples from that study alone and not other studies; this would mean that different studies estimate different "true"

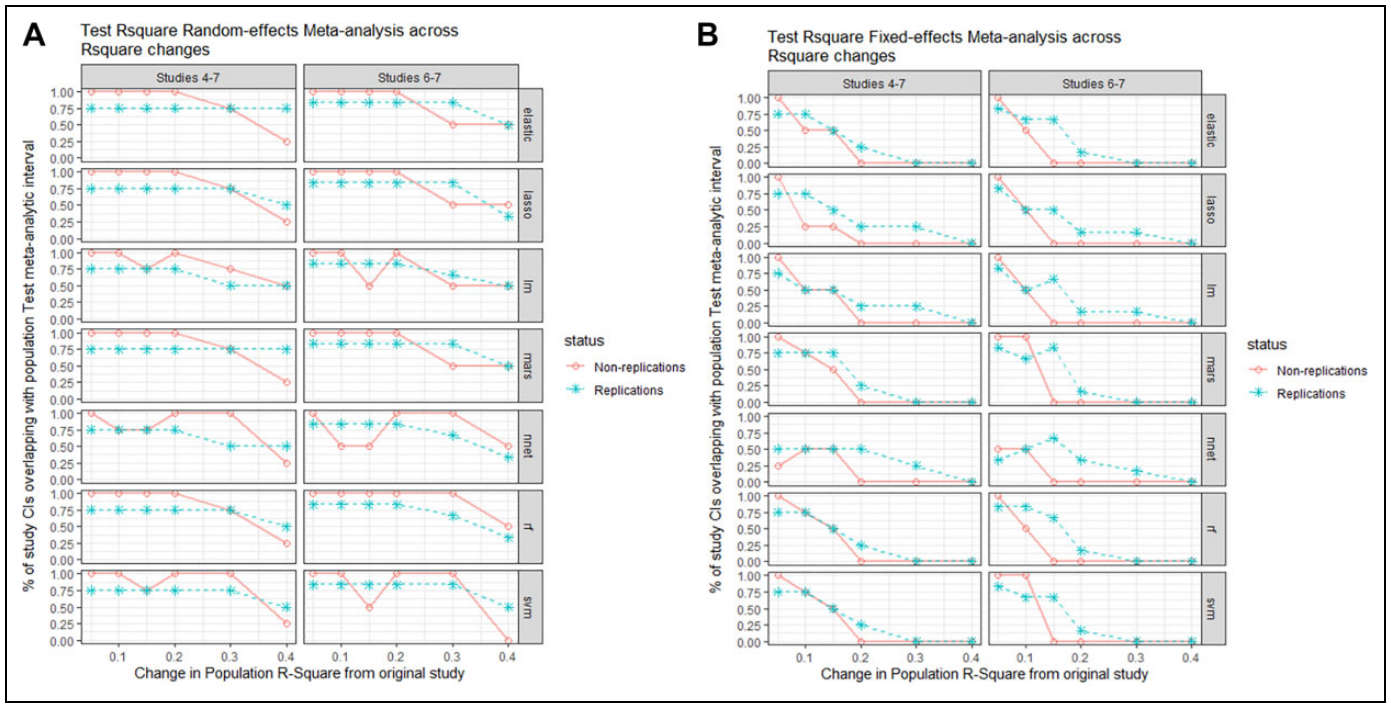

Figure 12. Population-level meta-analysis for test R2 when R2 changes occur in two or four studies. The x-axis shows the actual population R2 changes in the nonreplication studies from the original study. The lines depict the proportion of studies whose 95% confidence intervals (CIs) are at least partially overlapping with the CI of the population R2 estimate. (A) Random-effects meta-analysis used to combine studies. (B) Fixed-effects meta-analysis used to combine studies.

population values of CV R^2 (even if studies themselves are sampled from the same population). Hence, random-effects meta-analysis might work better for CV data, unlike test data in our simulation. As expected, when we estimate effects from studies of same population as in test data, we find fixed-effects working better than random-effects.

This follows theoretical expectations regarding fixed- and random-effects meta-analysis, and the success of random-effects for CV data and fixed effects for test data across methods suggests these procedures useful for testing machine learning replication. This also suggests that when actual studies are used which usually estimate different true effect sizes, researchers might be better advised to use random-effects for both CV and test data.

In short, random-effects meta-analysis performs better diagnostically for CV data, while fixed-effects worked for test data for studies sampled from same population. In general, random-effects meta-analysis is recommended for use in combining both folds and studies when you have multiple estimates, if you have access to the data at hand, bootstrapping to estimate study *SE*s may be advised. We used a data set where the original "study" had an R^2 of .5 across methods. Here, regression performed well diagnostically more or less at par with more complex machine learning methods, with SVM performing the best. Simulations with known nonlinear relations may be used to further study if these findings persist across various data scenarios.

General Discussion and Future Directions

In this article, we described certain methods used in psychological research and showed how we could adapt them to the specific concerns of machine learning research in social sciences. Researchers in social sciences might use machine learning for its predictive power; they would then like to ensure that the predictive accuracy for a specific machine learning method is reliable across samples from the same population. This would assure them of the usefulness of the set of predictors for either diagnostic purpose or for modeling specific effects of interest. This could also be useful for more pragmatic reasons for selecting a predictor set as substantively useful in policy research.

This article focused on three specific procedures that replication researchers can use in a variety of circumstances. First, the test of inconsistency, by constructing 95% CIs around difference estimates between replications and the original study, if the interval does not include zero, the replication is considered inconsistent with the original study. In the second procedure, test of consistency, a region of equivalence is constructed around zero where the difference in estimates is considered irrelevant for substantive reasons. If the 90% CIs of different estimates are fully within this interval, then they are held to be consistent with the original study. Finally, we use meta-analytic intervals to calculate uncertainty in R^2 estimates in estimating cross-validated and test accuracy measures across replications.

In Study 1, we illustrated their use in samples randomly drawn from an empirical data set; this allowed us to showcase how specific steps can be taken to achieve the three aims above. In Study 2, we also conducted diagnostic procedures to check which meta-analytic procedure is better at distinguishing true replications from non-replications. The results allow us to give tentative suggestions about appropriate methods suitable for testing replications. This also gives researchers one possible template they can build on when deciding how to examine the suitability of particular replication procedures and machine learning methods.

We consider the methods proposed to be a starting point for a discussion on this topic and acknowledge that it has certain limitations. We have focused on R^2 and not on other measures such as RMSE or Mean Absolute Error (MAE), which are also commonly used in machine learning. We chose R^2 for a couple of reasons. It is a popular measure of model performance in social science research, and it gives an intuitive understanding of the importance of the model for explaining the variation in the outcome. Moreover, researchers, such as Olkin and Finn (1995), Zou (2007), and

Chan (2008) among others, have provided different ways of obtaining the uncertainty estimates of R^2. Lastly, our previous research (Vijayakumar & Cheung, 2018) suggests that R^2 measures might be important for the success of replication attempts on predictor selection. So we focused on R^2 for this study; RMSE and MAE can also be approached in this fashion.

We do not cover all scenarios facing a replication researcher in machine learning, but these procedures can be adapted to serve there as well. For instance, our analysis assumed that the replication researcher has access to information about the individual CV fold R^2s in each study. This may not always be possible; reported results may include just the CV R^2 estimate of the original study. Similarly, the researchers may have access only to the point estimates of the original study's test R^2 or may have access to the original study sample but not the exact test subset. In such situations, the researcher can still try to estimate replicability of the results, albeit with less certainty about the result. One method is to use Olkin and Finn's method which requires just information about R^2 values and sample Ns to calculate the difference between R^2 values. A better method might be to bootstrap the replication CV or test sample at hand and calculate the difference between each bootstrap and the original reported R^2 value. We then use the bootstrapped CIs for tests on CV or test data.

Another limitation is that we have focused here exclusively on global measures such as R^2, which summarize the performance. However, researchers, especially in the social sciences, are concerned about the replicability of specific predictors and the nature of their effects. Summary measures cannot help us answer such questions. We also do not know if replicability of summary measures have any tangible connection with replicability of predictor effects; in principle, we could have widely varying predictor effects across samples giving rise to the same R^2 in the presence of many predictors. Researchers may have to analyze the replicability of predictors separately.

In prior research (Vijayakumar & Cheung, 2018), we focused on the replicability of predictor selection. There we found that in simulations with higher test R^2, relative superiority in predictive performance (MSE) was accompanied by increased replicability of selected predictors. However, in the presence of noise and small effects, improved predictive accuracy relative to other methods does not lead to increased replicability of predictors. Future research may try to further assess the link between replication of global measures and specific predictor selection.

Note that while random-effects meta-analysis is generally recommended as it leads to more generalizable results (Hedges & Vevea, 1998), fixed-effects models will have more power than random-effects for a given effect size, but at an increased cost of type-I error for heterogenous effect sizes (Cohn & Becker, 2003). Hence, the choice between random- and fixed-effects depends on several factors such as the number of studies, the nature of the conceptualized effect in the relevant field, tests for heterogeneity, and so on. (Borenstein et al., 2009). Our study has a small number of replications, and our suggestions regarding meta-analytic methods are tentative. More comprehensive simulation studies with power analysis should give more robust inferences.

While machine learning methods are already used for prediction and diagnosis, there are concerns regarding its replicability. We feel that methods used by psychologists can be used in specific procedures to assess the replicability of specific machine learning findings. As we detailed before, concerns about replicability in machine learning have focused on better reproducibility of the codes and specific algorithm details. This article's focus on quantifying uncertainty in the replication of predictive measures is a different but relevant aspect that can help users of machine learning establish and persuade others about the relevance and robustness of their models in real-life data sets. We hope this article is a start in this regard.

Data Availability

The data come from European Social Survey (ESS, 2018) and can be publicly downloaded.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Mike W.-L. Cheung was supported by the Academic Research Fund Tier 1 (FY2017-FRC1-008) from the Ministry of Education, Singapore.

Software Information

All data analysis was done using the open-source R statistical platform (R Core Team, 2019). We have detailed the R packages used for each method in the Methods section of the manuscript (see Machine learning methods subsection). The R codes for this analysis are publicly available online at PsychArchives (http://dx.doi.org/10.23668/psycharchives.2637).

References

Algina, J., & Keselman, H. J. (1999). Comparing squared multiple correlation coefficients: Examination of a confidence interval and a test of significance. *Psychological Methods*, *4*, 76–83.

Anderson, S. F., & Maxwell, S. E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods*, *21*, 1–12. doi:10.1037/met0000051

Bayati, M., Braverman, M., Gillam, M., Mack, K. M., Ruiz, G., Smith, M. S., & Horvitz, E. (2014). Data-driven decisions for reducing readmissions for heart failure: General methodology and case study. *PLoS One*, *9*, e109264. doi:10.1371/journal.pone.0109264

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*, 281–305.

Berk, R. (2012). *Criminal justice forecasts of risk a machine learning approach*. New York, NY: Springer.

Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems*, *50*, 602–613. doi:10.1016/j.dss.2010.08.008

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Hoboken, NJ: John Wiley.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T., Huber, J., Johannesson, M., … Wu, H. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, *2*, 637–644. doi:10.1038/s41562-018-0399-z

Cawley, G. C., & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, *11*, 2079–2107.

Chan, W. (2008). Bootstrap standard error and confidence intervals for the difference between two squared multiple correlation coefficients. *Educational and Psychological Measurement*, *69*, 566–584. doi:10.1177/0013164408324466

Cheung, M. W.-L. (2015). *Meta-analysis: A structural equation modeling approach*. Chichester, England: John Wiley.

Cohn, L. D., & Becker, B. J. (2003). How meta-analysis increases statistical power. *Psychological Methods*, *8*, 243–253.

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *The American Psychologist*, *60*, 170–180. doi:10.1037/0003-066X.60.2.170

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge, MA: Cambridge University Press.

Dingli, A., & Fournier, K. S. (2017). Financial time series forecasting—A deep learning approach. *International Journal of Machine Learning and Computing*, *7*, 118–122. doi:10.18178/ijmlc.2017.7.5.632

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics*, *35*, 352–359.

Drummond, C. (2006). *Machine learning as an experimental science (Revisited)*. Retrieved from https://www.aaai.org/Papers/Workshops/2006/WS-06-06/WS06-06-002.pdf.

European Social Survey Cumulative File, ESS 1-8. (2018). *Data file edition 1.0. NSD—Norwegian Centre for Research Data, Norway—Data Archive and distributor of ESS data for ESS ERIC*. doi:10.21338/NSD-ESS-CUMULATIVE

Forster, M. R. (2002). Predictive accuracy as an achievable goal of science. *Philosophy of Science*, *69*, S124–S134.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*, 1–22.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*, 460. doi:10.1511/2014.111.460.

Glaeser, E., Kominers, S. D., Luca, M., & Naik, N. (2015). *Big data and big cities: The promises and limitations of improved measures of urban life* (Working Paper 16-065). Cambridge, MA: NBER. doi:10.3386/w21778

Hastie, T., Tibshirani, R., & Friedman, J. (2016). *Elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York, NY: Springer.

Hedges, L. V., & Schauer, J. M. (2018). Statistical analyses for studying replication: Meta-analytic perspectives. *Psychological Methods*. doi:10.1037/met0000189

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*, 486–504. doi:10.1037/1082-989X.3.4.486

Held, E., Cape, J., & Nathan, T. (2016). Comparing machine learning and logistic regression methods for predicting hypertension using a combination of gene expression and next-generation sequencing data. *BMC Proceedings*, *10*, 141–145.

Hutson, M. (2018). Missing data hinder replication of artificial intelligence studies. *Science*. doi:10.1126/science.aat3298. Retrieved from https://www.sciencemag.org/news/2018/02/missing-data-hinder-replication-artificial-intelligence-studies

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab—An S4 package for Kernel methods in R. *Journal of Statistical Software*, *11*, 1–20.

Keogh, E., & Kasetty, S. (2003). On the need for time series data mining benchmarks: A survey and empirical demonstration. *Data Mining and Knowledge Discovery*, *7*, 349–371.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, *28*, 1–26. doi:10.18637/jss.v028.i05

Kukar, M., Kononenko, I., Grošelj, C., Kralj, K., & Fettich, J. (1999). Analysing and improving the diagnosis of ischaemic heart disease with machine learning. *Artificial Intelligence in Medicine*, *16*, 25–50. doi:10.1016/S0933-3657(98)00063-3

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*, 355–362. doi:10.1177/1948550617697177

Lavrač, N. (1999). Selected techniques for data mining in medicine. *Artificial Intelligence in Medicine*, *16*, 3–23. doi:10.1016/S0933-3657(98)00062 -1

Lazzeri, F. (2018, September 21). *Neural networks for forecasting financial and economic time series*. Retrieved from https://medium.com/microsoftazure/neural-networks-for-forecasting-financial-and-economic-time-series-6aca370ff412

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*, 18–22.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PLoS One*, *13*, e0194889. doi:10.1371/journal.pone.0194889

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*, 487–498. doi:10.1037/a0039400

Milborrow, S. (2007). *earth: Multivariate adaptive regression spline models* (R package Version 2.0-2). Retrieved from http://CRAN.R-project.org/package=earth.

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, *50*, 559–569.

Olkin, I., & Finn, J. D. (1995). Correlations redux. *Psychological Bulletin*, *118*, 155–164.

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rice, J. (2013). *Mathematical statistics and data analysis* (3rd ed.). Belmont, CA: Cengage Learning.

Schwartz, S. H. (1994). Are there universal aspects in the content and structure of values? *Journal of Social Issues*, *50*, 19–45. doi:10.1111/j.1540- 4560.1994.tb01196.x

Schwartz, S. H. (2012). An overview of the Schwartz theory of basic values. *Online Readings in Psychology and Culture*, *2*, 11. doi:10.9707/2307-0919.1116

Schwarzer, G. (2007). meta: An R package for meta-analysis. *R News*, *7*, 40–45.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*, 289–310. doi:10.1214/10-STS330

Simmons, J., Nelson, L., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. doi:10.1177/0956797611417632

Spellman, B. A. (2015). A short (personal) future history of revolution 2.0. *Perspectives on Psychological Science*, *10*, 886–899. doi:10.1177/1745691615609918

Steyerberg, E. W., Ploeg, T., & Calster, B. (2014). Risk prediction with machine learning and regression methods. *Biometrical Journal*, *56*, 601–606.

Taylor, D. J., & Muller, K. E. (1996). Bias in linear model power and sample size calculation due to estimating noncentrality. *Communications in Statistics Theory and Methods*, *25*, 1–13.

Van der Ploeg, T., Austin, P. C., & Steyerberg, E. W. (2014). Modern modelling techniques are data hungry: A simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology*, *14*, 137.

Varian, H. (2004). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, *28*, 3–28.

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

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York, NY: Springer.

Vijayakumar, R., & Cheung, M. W. (2018). Replicability of machine learning models in the social sciences. *Zeitschrift Für Psychologie*, *226*, 259–273. doi:10.1027/2151-2604/a000344

Walker, E., & Nowacki, A. S. (2011). Understanding equivalence and noninferiority testing. *Journal of General Internal Medicine*, *26*, 192–196. doi:10.1007/s11606-010-1513-8

Walsh, C. G., Ribeiro, J. D., & Franklin, J. C. (2017). Predicting risk of suicide attempts over time through machine learning. *Clinical Psychological Science*, *5*, 457–469. doi:10.1177/2167702617691560

Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation*, *8*, 1341–1390.

Wolpert, D. H. (2001). *The supervised learning no-free-lunch theorems*. Proceedings of the 6th Online World Conference on Soft Computing in Industrial Applications.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*, 1100–1122.

Yuan, K.-H., & Maxwell, S. E. (2005). On the post hoc power in testing mean differences. *Journal of Educational and Behavioral Statistics*, *30*, 141–167. doi:10.3102/10769986030002141

Zeng, J., Ustun, B., & Rudin, C. (2016). Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *180*, 689–722. doi:10.1111/rssa.12227

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods*, *12*, 399–413.

Author Biographies

Ranjith Vijayakumar is a PhD candidate at the Department of Psychology of National University of Singapore.

Mike W.-L. Cheung is an Associate Professor at the Department of Psychology of National University of Singapore.