Training replicable predictors in multiple studies

Prasad Patil[a,b] and Giovanni Parmigiani[a,b,1]

[a]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA 02215; and [b]Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA 02115

Edited by Victoria Stodden, University of Illinois at Urbana–Champaign, Champaign, Il, and accepted by Editorial Board Member Susan T. Fiske December 17, 2017 (received for review June 22, 2017)

This article considers replicability of the performance of predictors across studies. We suggest a general approach to investigating this issue, based on ensembles of prediction models trained on different studies. We quantify how the common practice of training on a single study accounts in part for the observed challenges in replicability of prediction performance. We also investigate whether ensembles of predictors trained on multiple studies can be combined, using unique criteria, to design robust ensemble learners trained upfront to incorporate replicability into different contexts and populations.

ensemble learning | replicability | cross-study validation | machine learning | validation

Science is facing an important debate about both reproducibility and replicability of results (1, 2). The dialogue on scientific replicability has focused predominantly on whether study results, such as inferences about hypotheses, are confirmed in a repeated study (3, 4). In this article, we shift the focus to the equally important but less commonly examined issue of replicability of the performance of predictors. Prediction algorithms developed using statistical and machine-learning techniques have been extraordinarily successful and are routinely evaluated in a variety of studies. Yet they have been largely rooted in methodologies and theories based on a single training dataset and often rely on cross-validation to approximate out-of-study performance. It is well established that cross-validation provides optimistic assessments of prediction ability compared with a cross-study assessment (5–9). Recent work in genomics also suggests that statistical learning strategies that optimize within-study cross-validation metrics are not necessarily the same as those that optimize cross-study validation metrics (8). Importantly, the model complexity of predictors whose performance is replicable can be different (often far lower) from that of predictors successful at cross-validation (10).

Our work is motivated by precision medicine, where the use of machine learning remains both essential and controversial. Tests such as Mammaprint (11) and OncotypeDX (12), which are Food and Drug Administration approved for clinical use, originated from such predictors. One of the most important milestones in translating a genomic prediction algorithm from bench to bedside is establishing the generalizability of its performance beyond the originating study (13). In what settings and for which patient populations will the algorithms perform as expected? Typically, answers come from training and validating on either one or a small number of patient datasets collected via different study protocols. Variation in measurement and data collection technologies, study populations, and sampling designs, as well as technological artifacts, inappropriate training strategies, and execution errors can have substantial impact on the replicability of predictions (5, 14, 15).

A general approach to investigating this issue is training on multiple studies and examining ensembles of prediction models, each trained on a different study. With multistudy ensembles we can quantify how the heterogeneity observed across studies contributes to the observed challenges in replicability of prediction performance and critically evaluate the limitations of within-study evaluations, such as cross-valuation, in training replicable predictors (8).

It is increasingly common to have available multiple datasets which measure the same outcome and many of the same covariates (16, 17). It is also important that these be simultaneously and systematically considered. Here we begin to investigate whether ensembles of prediction models trained on multiple studies can be used to design robust prediction algorithms that are trained upfront to incorporate replicability to different contexts and populations. Options for training predictors in the multistudy setting include merging all datasets together (thus ignoring heterogeneity) and directly modeling heterogeneity (e.g., via meta-analysis). As an alternative, we examine the use of weighted ensembling. We propose a range of weighting strategies that reward cross-study performance among predictors in the ensemble. We explore strategies motivated by decision theoretic criteria applied to hypothetical future studies, as well as heuristic strategies.

Although ensemble learning is a common paradigm (18–20), the role of multiple studies and their heterogeneity in optimal ensembling strategies has not been explored. Here, we outline a general approach and investigate the characteristics of several implementations in simulations and in a comprehensive collection of datasets including gene expression and survival in patients with ovarian cancer. We ask whether and in what scenarios training ensembles with heterogeneous datasets can improve the generalizability and out-of-sample performance of the resulting predictors and increase replicability.

Results

Cross-Study Learners. The general architecture of a cross-study learner (CSL) is in Fig. 1. CSLs are uniquely specified by three choices: (*i*) a study subsetting strategy, (*ii*) one or more single-study learners (SSLs), and (*iii*) a combination approach. Subsetting is concerned with both inclusion and exclusion criteria for candidate studies and the definition of groups of studies with similar distributions of predictors and outcomes. An SSL can be any algorithm that produces a prediction model using a single study. Combination approaches use multiple prediction models to deliver a single prediction rule applicable to external validation studies. Here we investigate unique combination approaches designed to enhance replicability. This approach bypasses the potentially complex task of explicitly modeling

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Reproducibility of Research: Issues and Proposed Remedies," held March 8–10, 2017, at the National Academy of Sciences in Washington, DC. The complete program and video recordings of most presentations are available on the NAS website at www.nasonline.org/Reproducibility.

Author contributions: P.P. and G.P. designed research; P.P. and G.P. performed research; P.P. analyzed data; and P.P. and G.P. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. V.S. is a guest editor invited by the Editorial Board.

Data deposition: Code and instructions to reproduce analyses are available on GitHub at https://github.com/prpatil/csml_rep.

[1]To whom correspondence should be addressed. Email: gp@jimmy.harvard.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1708283115/-/DCSupplemental.

Published online March 12, 2018.

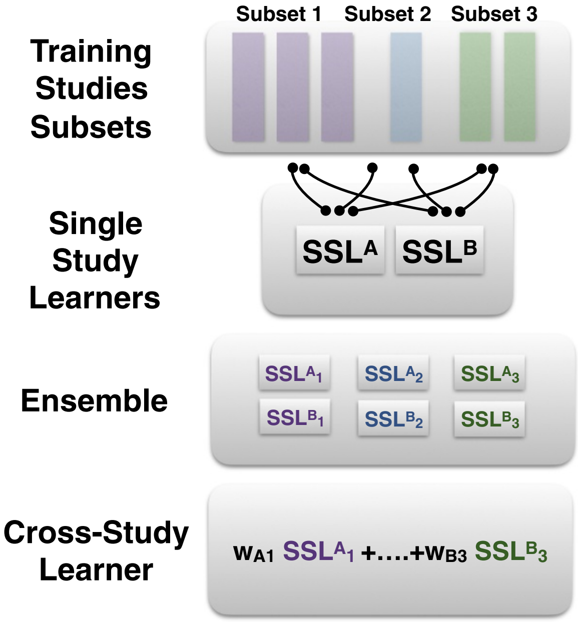

Fig. 1. The architecture of a CSL, illustrated with six studies divided into three subsets, two SSLs, and general weights.

variation in the relationship between Y and $\mathbf{X}$ across datasets. Another important property of this architecture is that it can support cross-study learning when the training data cannot be merged, because of either access restrictions or size limitations.

For a very simple illustration of Fig. 1, we could choose CART (21) and linear regression as the two SSLs, form three subsets of studies, train both CART and regression on each subset, and combine the predictions by simple averaging. The resulting ensemble algorithm is a CSL.

More generally, consider L learners (SSLs), K training, and V validation datasets, within which we have a comparable outcome Y_k of size n_k and matched sets of predictors $\mathbf{X}_k$ of dimension $n_k \times m$, $k = 1, \ldots, K+V$. Our goal is to make predictions in the V validation datasets, using the K datasets for training. Let the prediction function trained on dataset k using learner l be $\hat{Y}_k^l(\mathbf{x})$, where $\mathbf{x}$ is an m-dimensional point in the predictor space. We also use $\hat{Y}_k^l(\mathbf{X})$ to denote predictions made at n points represented by the $n \times m$ matrix $\mathbf{X}$. A linear CSL with weights w_{lk} has the form

$$\hat{Y}(\mathbf{x}) = \sum_{l=1}^{L} \sum_{k=1}^{K} w_{lk}\, \hat{Y}_k^l(\mathbf{x}). \qquad [1]$$

If $K = 1$, we fall back to standard ensemble learning, while we can have CSLs with $L = 1$.

We envision two general strategies for the combination step. A heuristic approach is to define weights that incorporate cross-study performance. For example, we conjecture that cross-study validation offers a better estimate of the generalizability of a predictor and that weighting by measures of training-set cross-study performance will yield better performance across a test set of studies. A more formal approach is to pose the combination problem from a decision theoretic angle and use utility functions that reflect expected performance in future studies. A CSL is then evaluated on the validation datasets using a decision theoretic criterion:

$$\mathcal{U}(\mathrm{CSL}) = \sum_{k=K+1}^{V} \lambda_k\, U(\hat{Y}(\mathbf{X}_k), Y_k). \qquad [2]$$

U is an expected utility function quantifying the quality of predictions (22), such as negative mean squared error, and λ_ks allow for different validation studies to be weighted differently.

The definition of what constitutes a study depends on the context and can be chosen to address questions about generalizability to desired target populations. In Fig. 1, six studies are available, but they are clustered into $K = 3$ subgroups, perhaps because the differences across these groups are important scientifically, and capturing this variation helps with future out-of-training-samples use. Other choices are available to control prediction properties. Subgroups may overlap; for example, we could consider the power set of a given collection of independent studies. Alternatively, we could use resampling to generate a set of artificial studies for each of the observed ones and build a CSL on the full collection. In Fig. 1 we could generate 100 studies for each of the 6 initial ones and build a CSL with $K = 600$.

Simulations. We next describe results of simulations that begin to explore the many options opened by the architecture of Fig. 1 and to evaluate the performance of CSLs. Following Friedman (23), we vary the outcome generation function and use differing error distributions. Similarly to the `SimulatorZ` Bioconductor package (24), we use predictor profiles resampled from real data, in our case `CuratedOvarianData` so that the predictors have a realistic distribution. We then generate the outcomes given the predictors, using known linear relations. We are interested in modeling variation across studies, so we perturb the coefficients across datasets within a uniform window. Varying the size of this window across sets of simulations allows for easy comparisons of performance across different degrees of study heterogeneity. We report results based on a small set of features for prediction here and consider larger sets in *SI Appendix*, section 2.2.

We consider seven choices for the SSL box in Fig. 1: (*i*) Lasso, (*ii*) CART, (*iii*) Neural Network, (*iv*) Mas-o-Menos, (*v*) Random Forests, (*vi*) model-based boosting, and (*vii*) the union of 1 through 4 (see *Materials and Methods* for implementation details). These learners are different, widely used, and straightforward to apply out of the box. While they are far from exhausting the set of useful possibilities, they can provide an informative initial exploration.

For each of the SSLs above, we form a CSL using five alternative choices for weights: simple average of predictions from each SSL ("Avg"), average weighted by study sample size ("n-Avg"), average weighted by cross-study performance ("CS-Avg"), stacked regression ("Reg-s"), and averages of study-specific regression weights ("Reg-a"). The last three reward replicability. The regression weights are motivated by optimality under squared error loss. In this section they include intercept terms and do not normalize the weights. Further details are in *Materials and Methods*. As a baseline comparator, we merge all studies and learn a single SSL (the merged learner).

Fig. 2 summarizes the results. The most important comparison is between the merged learner and the rest, who are all CSLs. At small perturbation levels, the merged learner performs comparably to CSLs and wins only when the SSL is the neural net. As study heterogeneity increases, the merged learner's performance advantage deteriorates. CSLs can generally provide robust learning without losing performance even when study heterogeneity is limited. On the other hand, the margin of improvement from adopting a CSL at higher heterogeneity can be substantial.

Comparing CSL weighting strategies, fixed weights (Avg and n-Avg) do well, but are always inferior to one of the weighting schemes that reward replicability. More specific comparisons depend on the SSL, suggesting that the choice of weighting needs to be tailored to the SSL or SSLs. As a general trend, Reg-s weights do well at low perturbations, while Reg-a and CS-Avg

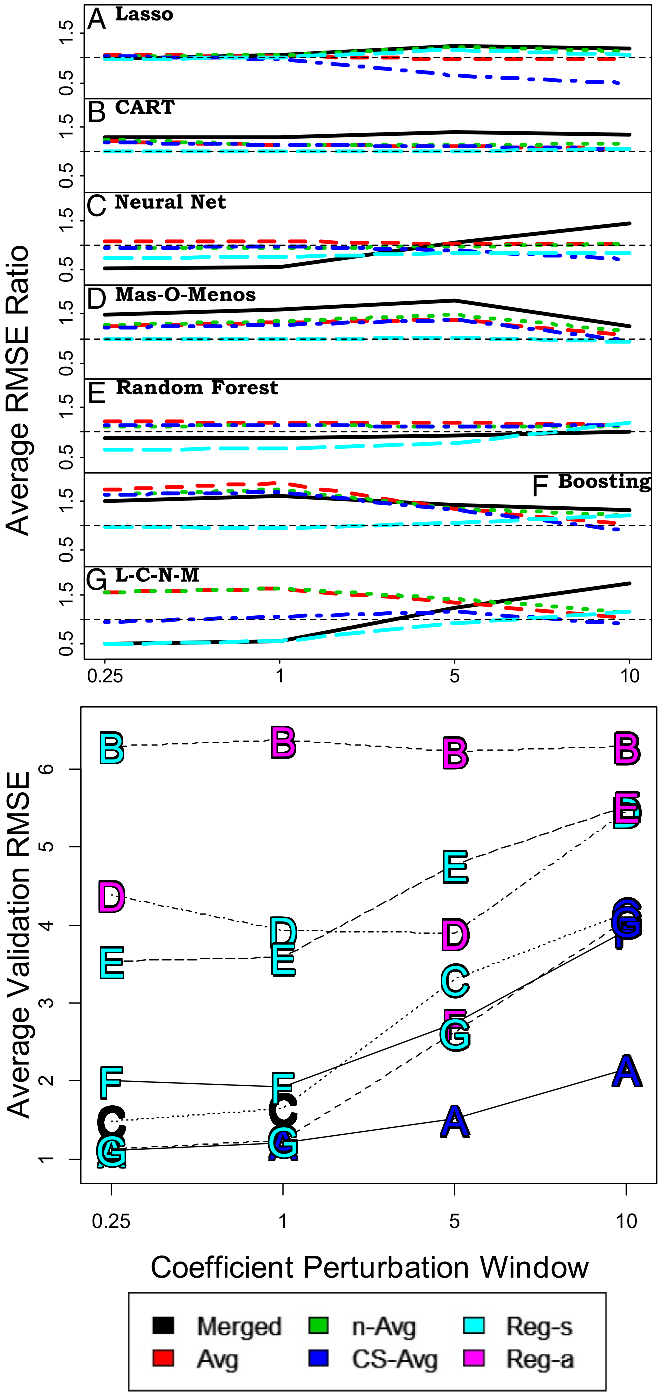

Fig. 2. Ratios of validation rms errors (rmses) to the rmse of the Reg-a weighting strategy, averaged over 100 simulation iterations, as we vary the coefficient perturbation window. *Top* seven panels correspond to different choices of SSL; the colors correspond to different weighting schemes. *Bottom* displays average validation rmse of the best-performing scheme (indicated with color) for each SSL (indicated by letter) at each perturbation window.

weights do well at high perturbation. When Lasso is the SSL, at low perturbation, all CSLs perform similarly. As perturbation increases, CS-Avg weights do substantially better than the rest, while Reg-a weights perform similarly to Avg, and Reg-s weights are only slightly better than the merged learner. These observations, however, do not extend to other panels in Fig. 2. For example, when four SSLs are used jointly, CS-Avg weights are only slightly better than Reg-a and Reg-s weights do quite well. Comparing CART to Random Forest, which performs some ensembling internally, we see that the merged learner performs better in the Random Forest SSL case than in CART, while in both cases gains are relatively insensitive to the level of perturbation.

While the seven panels in Fig. 2, *Top* show relative differences, Fig. 2, *Bottom* summarizes the absolute comparison, for reference. The Lasso is based on a model that is relatively similar to the data-generating mechanism. Additionally, shrinking of coefficients may prove beneficial as study heterogeneity increases. Accordingly, CSLs using Lasso perform best overall.

Normalized and no-intercept versions of regression weights performed similarly to the Reg-a and Reg-s shown in Fig. 2. We present this information and additional variations on this simulation in *SI Appendix*, section 2.1. In these variations, we generate the outcome under different error structures and change the coefficient perturbation windows of the validation datasets.

Ovarian Cancer Prognosis. We next consider an application to predicting survival of patients with ovarian cancer, using gene expression profiles taken at time of diagnosis as predictive features. CuratedOvarianData (25) is a comprehensive collection of datasets relevant to our specific question, generated by a systematic literature review—an important step toward avoiding biases arising in hand-picked collections of datasets (9). Riester et al. (26) used CuratedOvarianData to build a predictor using a meta-analytic method. Their predictor outperformed a comprehensive (9) collection of existing predictors. Here we compare their predictor to a CSL. For comparability, our CSL uses a single-study approach consistent with their method as the SSL. We repeatedly ($R = 250$) separate 15 ovarian cancer datasets into groups of $K = 12$ training and $V = 3$ validation datasets. On the training sets, we then train a Riester predictor, a CSL, a merged learner, and an SSL on the largest study, The Cancer Genome Atlas (TCGA) study. See *Materials and Methods* for details on each step.

Riester uses a multistudy training method that meta-analyzes regression coefficients and produces a score that acts as an increment to the hazard of death in a proportional hazard model. To measure performance we use the hazard ratio associated with a change of one unit in the score vector, as evaluated in the validation datasets. This is a practical and clinically interpretable measure of discrimination.

Fig. 3 provides a comparison between several approaches. The regression weighting strategies, motivated by optimality

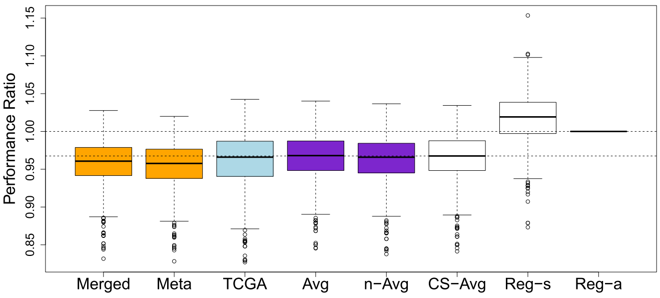

Fig. 3. Differential discrimination of alternative classifiers. For each classifier we compute the hazard ratio associated with a change of one unit in the score vector, as evaluated in the validation datasets. The vertical scale is the ratio of this performance measure to that of the Reg-a CSL. Colors indicate classes of learning strategies: White is weighted CSLs with weights addressing cross-study prediction, purple is CSLs with fixed weights, orange is merging and meta-analysis, and blue is a SSL trained on the TCGA dataset. Horizontal lines are at $y = 1$ and at median performance of CS-Avg.

criteria and explicitly aiming at improving replicability, perform best. Stacking specifically shows a clear advantage. Ensembling with cross-study weights and fixed weights is next. The latter are examples of forecast combination approaches that do not attempt to account for interstudy performance. Merging and meta-analysis have the worst overall performance, suggesting that there is enough interstudy heterogeneity that combining the datasets or their study-specific parameter estimates is not advantageous. The TCGA study contributes more than half of the samples. When used in isolation, it can still contribute predictors whose performance is comparable to that of some of the resembling approaches. This illustrates that trade-offs between multistudy and single-study learning need to be weighed carefully in specific applications.

This example illustrates that multistudy learning can outperform traditional approaches. Importantly, the collection of studies is comprehensive and thus unbiased with respect to performance; our reference comparators are best in class, so we can confidently state that the CSLs are the best so far; the SSL we used is the same as in the original Riester et al. (26) paper, so all of the gains are directly attributable to multistudy ensembling. In addition to the results presented here, in *SI Appendix*, section 1 we provide an examination of multistudy ensembling strategies when signatures are pretrained and when the task is classification. These results reiterate that CSLs have competitively generalizable performance compared with other approaches.

Discussion

A potentially useful approach to increasing replicability of prediction performance is to use statistical concepts that acknowledge cross-study variation and explicitly aim to produce results which are likely to be replicable across studies. Thus, we explore the implications of training on multiple studies and then ensembling prediction models, each trained on one of the studies. We outline a general class of learners termed CSLs, implemented through three components: (*i*) one or more single-study learners, (*ii*) a study subsetting strategy, and (*iii*) a combination approach. We show how to quantify the loss of generalizability associated with single–training-set practices compared with multistudy practices (Fig. 3), providing insight into determinants of replicability. We also hope to contribute to a more systematic utilization of multiple studies in machine learning.

When sufficiently harmonized data from multiple studies are available, weighted combinations of learners trained on multiple training sets are feasible and may increase the chance of replicable performance. Different learners respond to cross-study heterogeneity in different ways (Fig. 2). Each of several options we presented can work well, depending upon the extent of interstudy heterogeneity, the choice of learning algorithm, and other data attributes. In our specific application to ovarian cancer prognosis, we considered a comprehensive collection of studies, unbiased with respect to performance; we illustrate that multistudy learning can outperform best-in-class comparators. We chose our SSL so we can attribute gains unequivocally to multistudy ensembling. We conclude that this approach warrants close attention and further work.

As more data become publicly available, researchers are presented with many opportunities for working with multiple datasets simultaneously to train better predictors. These datasets may have been collected in different circumstances, with different procedures, and on different populations. The features from different studies are also unlikely to be exactly the same. This heterogeneity typically implies systematic differences and biases that challenge the replicability of the performance of predictors. Merging the data, perhaps after some preprocessing to improve harmonization, and/or adding study-specific features to the set of predictors are viable options. Some of the study-to-study variation can also be reduced by selecting predictors that are more likely to be comparably measured across studies (27, 28), constructing shared features using unsupervised methods (29, 30), and considering comparable ascertainment mechanisms for subjects or samples. The covariate shift literature (31–33) addresses the problem of differing marginal distributions of predictors through case reweighting and relearning to make algorithms more applicable to a target population of known distribution.

When differences across studies are small, or when these harmonization steps are very successful, it is natural to combine all training studies, to exploit the power of larger training sample sizes. More often, variation in the relationship between predictors and outcome across studies can be complex and hard to remove. Observed and unobserved attributes of the data collection process may affect this relationship even though the marginal distribution of the covariates appears similar across studies. This limits the efficacy of the methods described in the previous paragraph. Cross-study learning provides a straightforward and intuitive way forward and also offers a direct means of incorporating replicability into the learning process. As heterogeneity increases, our simulations (Fig. 2) indicate a "transition point" in the heterogeneity scale where variation in the relation between predictors and outcomes across studies becomes large enough to make CSL preferable to merged learning.

Our application embeds a significant data harmonization effort (25) and our simulations address only heterogeneity in the regression coefficients. Data heterogeneity and harmonization difficulties remain critical challenges for multistudy learning as well. More work is needed to characterize scenarios where heterogeneity is too extreme to even attempt multistudy learning. Transfer learning takes knowledge learned for completing one task and tries to use it to better learn how to complete a different task (34–36). Connections may be worth exploring when outcomes differ across studies.

Much of the scientific debate on replicability and reproducibility focuses on whether the results, or conclusions, of a study are found again when a sufficiently similar study is conducted. Additional progress is needed in making these terms rigorous and building conceptual frameworks for measuring reproducibility and replicability (1). When we consider predictions, however, we have direct and well-established means of evaluating replicability of performance, assuming we can adequately address data harmonization. A contribution of our approach is to not directly require detailed understanding and modeling of the contributors to study-to-study heterogeneity and to difficulties in replicability. Instead we account for these within the context of a well-defined language of evaluating predictions via the cross-study performance of each SSL.

We hope our work will encourage a systematic exploration of cross-study replicability of prediction performance. Multistudy learning approaches are a promising direction to pursue in the quest for practical remedies.

Materials and Methods

Regression Weights. We first provide general motivation and justification for choosing the weights w_{lk} in Eq. **1**. Ideally, we would seek optimal properties consistent with the decision theoretic criterion in Eq. **2**. As a formal theory is not yet developed, we provide approximations. Consider minimizing the least-squares distance between $\hat{Y}(\mathbf{x})$ and $E\{Y|\mathbf{x}\}$ in a hypothetical $(K+1)$st study, with respect to the weights. To derive an analytic solution, we start with predictors that are discrete, with a finite set of values. Let $\hat{\mathbf{Y}}_k^l$ be the vector of predictions generated by prediction function $\hat{Y}_k^l(\mathbf{x})$, with each vector element in $\hat{\mathbf{Y}}_k^l$ corresponding to a specific point $\mathbf{x}$ in the predictors' space. Let $E^x\{Y\}$ be the corresponding vector of true conditional means in study $K+1$ and let $\tilde{\mathbf{Y}} = [\hat{\mathbf{Y}}_1^1, \ldots, \hat{\mathbf{Y}}_1^L, \hat{\mathbf{Y}}_2^1, \ldots, \hat{\mathbf{Y}}_2^L, \ldots, \hat{\mathbf{Y}}_K^1, \ldots, \hat{\mathbf{Y}}_K^L]$. Conditional on the unknown $E^x\{Y\}$, and assuming $\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}}$ is invertible, the optimal unconstrained weights can be straightforwardly shown to be the

KL-dimensional vector of coefficients $(\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}})^{-1}\tilde{\mathbf{Y}}'E^x\{Y\}$ from the regression of $E^x\{Y\}$ on the $\hat{\mathbf{Y}}^l_k$s.

In practice one can estimate $E^x\{Y\}$ by the vector $\sum_k \bar{n}_k\bar{Y}_k$, where the elements of vector $\bar{Y}_k$ are the study-specific mean responses evaluated at each point $\mathbf{x}$, and $\bar{n}_k = n_k/\sum_{k'} n_{k'}$. This leads to weights $\mathbf{w}_{\text{Reg}} = \sum_k \bar{n}_k(\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}})^{-1}\tilde{\mathbf{Y}}'\bar{Y}_k$. When some of the predictors are continuous, points in the space of predictors are typically observed only once and do not overlap across studies. Then $E^x\{Y\}$ can no longer be directly estimated by averaging across studies. We can continue to use the previous result as a guide if we define study-specific $\tilde{\mathbf{Y}}_k$ as an $n_k \times KL$ matrix formed similarly to $\tilde{\mathbf{Y}}$ above, but each one restricted to include predictions generated in correspondence with points observed in study k only. These can be used to form

$$\mathbf{w}_{\text{Reg-a}} = \sum_{k=1}^{K} \bar{n}_k(\tilde{\mathbf{Y}}'_k\tilde{\mathbf{Y}}_k)^{-1}\tilde{\mathbf{Y}}'_k Y_k. \qquad [3]$$

Each term $(\tilde{\mathbf{Y}}'_k\tilde{\mathbf{Y}}_k)^{-1}\tilde{\mathbf{Y}}'_k Y_k$ is a KL-dimensional vector of weights. For $(K-1)L$ of these weights (the cross-study contributions) the response Y_k has not been previously used in generating the regressors in the corresponding columns of $\tilde{\mathbf{Y}}_k$, while for the remaining L (the within-study contributions) it has.

Alternatively, $E^x\{Y\}$ can be approximated directly by the totality of the observed Ys in the training sets. This motivates a "stacked weights" approximation of the optimum (37). We first form the vector of the observed responses $Y = [Y'_1, \ldots, Y'_K]'$, of dimension $N = \sum n_k$, and the $N \times KL$ matrix of study-specific SSL predictions $\mathbf{T} = [\tilde{\mathbf{Y}}'_1, \ldots, \tilde{\mathbf{Y}}'_K]'$. We then define the stacked weights to be

$$\mathbf{w}_{\text{Reg-s}} = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'Y. \qquad [4]$$

We can also use nonnegative least squares to estimate the coefficients for both regression-based weighting approaches. A nonnegative constraint on weights is recommended by ref. 38 in the single-study resampling setting. These approaches generalize work on combinations of forecasters (39, 40).

Replicability Weights. An alternative strategy for forming weights is to directly engineer them to reward cross-study replicability. Cross-study performance within the K training studies informs about the generalizability of a predictor, so weighting $\hat{Y}^l_k$ by measures of cross-study performance may yield better performance across a future set of test studies. We propose an approach based on refs. 9 and 41. For a pair of studies k and k' in the training set, let the vector $\hat{Y}^l_{kk'}$ include the predictions obtained using model $\hat{Y}^l_{k'}(\mathbf{x})$, trained in study k', on the prediction profiles in study k. We can evaluate cross-study performances via a utility function $U(Y_k, \hat{Y}^l_{kk'})$. From these we can construct, for each l, a $K \times K$ matrix $\mathbf{Z}$, with entries $Z^l_{kk'} = -U(Y_k, Y^l_{kk'})$. The rows of this matrix represent how each $\hat{Y}^l_{k'}$ performs across all of the training datasets. Giving higher weights to prediction functions (SSL) that show evidence of replicability within the training studies should help improve the CSL replicability in the presence of study heterogeneity. The V validation studies are never used in this CSL. For each l, we construct weights by first summarizing row k of $\mathbf{Z}$ by the average of the off-diagonal elements, or $z_{lk} = \frac{1}{K-1}\sum_{i\neq k} Z^l_{ik}$, so that the resubstitution error Z^l_{kk} is excluded. Other row summarizations could also give useful results. We then compute

$$w^{\text{CS}}_{lk} = \frac{|z_{lk} - \max(z_{11}, \ldots, z_{LK})|}{\sum_i |z_{lk} - \max(z_{11}, \ldots, z_{LK})|}. \qquad [5]$$

This assigns a weight of zero to the worst-performing SSL, shifts losses for all other SSLs accordingly, and normalizes weights.

Weights Used in Comparisons. Motivated by these concepts, we compare six combination approaches:

Merged: Combine all datasets into one large dataset and train a single prediction function, as in single-study learning.

Avg: Simple average: $w_{lk} = 1/LK$.

n-Avg: Sample-size–weighted average: $w_{lk} = n_k/L\sum_{k'} n_{k'}$. This allows us to explore whether larger datasets will on average produce a more reliable or more stable CSL.

CS-Avg: Replicability weights (5). In the simulations we form Zs with elements $Z^l_{jk} = \text{MSE}^l_{jk} = \frac{1}{n_j}\sum(Y_j - \hat{Y}^l_{jk})^2$ and summarize rows by the root of the average off-diagonal mean-square error $z_{lk} = (\frac{1}{K-1}\sum_{i\neq k} Z^l_{ik})^{1/2}$. In the cancer application, we form Zs with the discrimination measure discussed in *Results* and proceed similarly thereafter.

Reg-a: Averages of study-specific coefficients similar to $\mathbf{w}_{\text{Reg-a}}$, but computed via nonnegative least squares (simulations) as given by the nnls R package (42) or Cox regression (ovarian example). In *SI Appendix*, section 2.1 we also examine weights normalized to 1 and with no intercept.

Reg-s: Stacked regression weights similar to $\mathbf{w}_{\text{Reg-s}}$, but computed via nonnegative least squares (simulations) or Cox regression (ovarian example). In addition we examine the same variations as for Reg-a.

Simulated Data. We generate the outcome vectors Y_k conditional on the observed predictors. Within each iteration ($R = 100$ across learners and scenarios), we randomly separated datasets into $K = 12$ training and $V = 3$ validation sets. We then reduced each dataset to a random subset consisting of the same 20 genes, randomly chosen in each iteration. Up to 20 genes in this subset are used to create a simple data-generating model; this is not a feature selection step. We show in *SI Appendix*, section 2.2 that this choice does not affect the conclusions of our simulation by recreating Fig. 2 with 40- and 100-gene subsets.

We generate Y using linear models. At each iteration we randomly select a subset of at least two of the available genes. We generate "pivot" β coefficients uniformly from $[-5, -0.5] \cup [0.5, 5]$, bounded away from zero so that each selected gene would contribute to the outcome. To simulate $P(Y_k|\mathbf{X_k}) \neq P(Y_j|\mathbf{X_j})$, we then perturbed the coefficients across datasets. We generated each β_{ik} for each dataset uniformly from the window $[\beta_i - \eta, \beta_i + \eta]$. η was taken from the set $\{0.25, 1, 5, 10\}$. These perturbation windows were chosen so that heterogeneity we observed in `CuratedOvarianData` (25) would be well within the range used and to explore both small and large heterogeneity. We chose to use a location shift instead of a scale change so that only some covariates would be greatly affected by the dataset-to-dataset variation.

We split the 12 training datasets evenly into low- and high-perturbation groups. The low-perturbation window we used was 0.25 units, which was kept constant as we varied the window for the high-perturbation group through the set of η values listed above. We evaluated performance in cases where the validation set resembled the low-perturbation group. Additional simulation scenarios are described and reported in *SI Appendix*, section 2.3.

Learners in Simulation Analysis. We used six different learning algorithms to generate prediction functions and cross-study learners with $L = 1$: Lasso [`glmnet` (43)], CART [`rpart` (44)], Random Forest [`ranger` (45)], Neural Network [`nnet` (46)], Mas-o-Menos [custom function, following Zhao et. al. (10)], and boosting [`mboost` (47)]. As the point is to investigate CSL, rather than comparing SSLs, we used default settings when possible. For CART, we included a pruning step. For `nnet` all models were fitted with `size` = 10. We also consider an $L = 4$ implementation including Lasso, CART, Neural Network, and Mas-o-Menos, four learners that do not internally implement resampling-based ensembling.

Ovarian Cancer Data. `CuratedOvarianData` (25) provides a manually curated collection of data for gene expression meta-analysis of patients with ovarian cancer, as well as software for reproducible preparation of harmonized datasets. Here we use all 15 studies in `CuratedOvarianData` providing survival information without missing data in the features. The sample sizes of each dataset varied from 42 to 510 subjects, and the 14 datasets had 2,909 gene features in common. The gene expression features of each dataset are normalized.

Learner in Ovarian Cancer Analysis. We develop predictors that produce risk scores for each patient following ref. 26. The SSL first chooses the top 200 genes via univariate Cox regression. If g_{ik} is the coefficient estimate for the ith gene in study k, the risk score for a new individual with gene expression values $\mathbf{x}$ is $\sum_i g_{ik}\mathbf{x}_i$. The meta-analysis predictor (26) combines these coefficients using the `rma()` function from the `metafor` package (48). We used the `DL` option to include a random effect to account for heterogeneity.

Performance of each approach was evaluated by fitting a proportional hazard model in the validation datasets, with survival as the censored outcome, and the score as predictor, and computing the hazard ratio associated with a change of one unit in the score vector. This measures discrimination. Cox coefficients were also used as the performance metric in z_{lk} for the cross-study weighted average. We compared the six combination approaches described above to each other as well as the meta-analytic signature and the signature produced by the TCGA dataset, which exhibited the best overall SSL performance. The TCGA signature was not applied to the TCGA dataset if it appeared in the validation datasets.

Reproducibility. Code and instructions to reproduce analyses are at https://github.com/prpatil/csml_rep.

ACKNOWLEDGMENTS. We thank Lorenzo Trippa and the reviewers for thoughtful comments and wording suggestions on the manuscript. P.P. was supported by National Cancer Institute (NCI) Training Grant 2T32CA009337-36. G.P. was supported by NCI Core Grant 4P30CA006516-51.

1. Committee on Applied and Theoretical Statistics, Board on Mathematical Sciences and Their Applications, Division on Engineering and Physical Sciences, National Academies of Sciences, Engineering, and Medicine (2016) *Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results, Summary of a Workshop*, ed Schwalbe M (National Academies Press, Washington, DC).
2. Kenett RS, Shmueli G (2015) Clarifying the terminology that describes scientific reproducibility. *Nat Methods* 12:699–699.
3. Open Source Collaboration, et al. (2015) Estimating the reproducibility of psychological science. *Science* 349:aac4716.
4. Heller R, Bogomolov M, Benjamini Y (2014) Deciding whether follow-up studies have replicated findings in a preliminary large-scale omics study. *Proc Natl Acad Sci USA* 111:16262–16267.
5. Simon R, Radmacher MD, Dobbin K, McShane LM (2003) Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J Natl Cancer Inst* 95:14–18.
6. Ma S, et al. (2014) Measuring the effect of inter-study variability on estimating prediction error. *PLoS One* 9:e110840.
7. Chang LB, Geman D (2015) Tracking cross-validated estimates of prediction error as studies accumulate. *J Am Stat Assoc* 110:1239–1247.
8. Bernau C, et al. (2014) Cross-study validation for the assessment of prediction algorithms. *Bioinformatics* 30:i105–i112.
9. Waldron L, et al. (2014) Comparative meta-analysis of prognostic gene signatures for late-stage ovarian cancer. *J Natl Cancer Inst* 106:dju049.
10. Zhao SD, Parmigiani G, Huttenhower C, Waldron L (2014) Más-o-Menos: A simple sign averaging method for discrimination in genomic data analysis. *Bioinformatics* 30:3062–3069.
11. Van't Veer LJ, et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415:530–536.
12. Paik S, et al. (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New Engl J Med* 351:2817–2826.
13. Marchionni L, et al. (2008) Systematic review: Gene expression profiling assays in early-stage breast cancer. *Ann Intern Med* 148:358–369.
14. Patil P, Bachant-Winner PO, Haibe-Kains B, Leek JT (2015) Test set bias affects reproducibility of gene signatures. *Bioinformatics* 31:2318–2323.
15. Haibe-Kains B, et al. (2012) A three-gene model to robustly identify breast cancer molecular subtypes. *J Natl Cancer Inst* 104:311–325.
16. Kannan L, et al. (2016) Public data and open source tools for multi-assay genomic investigation of disease. *Brief Bioinform* 17:603–615.
17. Klein R, et al. (2014) Data from investigating variation in replicability: A "many labs" replication project. *J Open Psychol Data* 2:e4.
18. Raftery A, Madigan D, Hoeting J (1997) Bayesian model averaging for linear regression models. *J Am Stat Assoc* 92:179–191.
19. Rokach L (2010) Ensemble-based classifiers. *Artif Intelligence Rev* 33:1–39.
20. Costello JC, et al. (2012) Wisdom of crowds for robust gene network inference. *Nat Methods* 9:796–804.
21. Breiman L, Friedman J, Stone CJ, Olshen RA (1984) *Classification and Regression Trees* (Chapman and Hall, New York).
22. Parmigiani G, Inoue LYT (2009) *Decision Theory: Principles and Approaches* (John Wiley & Sons, Chichester, UK).
23. Friedman JH (2001) Greedy function approximation: A gradient boosting machine. *Ann Stat* 29:1189–1232.
24. Zhang Y, Bernau C, Waldron L (2017) SimulatorZ: Simulator for collections of independent genomic data sets, version 1.12.0. Available at https://www.bioconductor.org/packages/release/bioc/html/simulatorZ.html. Accessed January 15, 2017.
25. Ganzfried BF, et al. (2013) CuratedOvarianData: Clinically annotated data for the ovarian cancer transcriptome. *Database* 2013:bat013.
26. Riester M, et al. (2014) Risk prediction for late-stage ovarian cancer by meta-analysis of 1525 patient samples. *J Natl Cancer Inst* 106:dju048.
27. Parmigiani G, Garrett-Mayer ES, Anbazhagan R, Gabrielson E (2004) A cross-study comparison of gene expression studies for the molecular classification of lung cancer. *Clin Cancer Res* 10:2922–2927.
28. Garrett-Mayer E, Parmigiani G, Zhong X, Cope L, Gabrielson E (2007) Cross-study validation and combined analysis of gene expression microarray data. *Biostatistics* 9:333–354.
29. Meng C, et al. (2016) Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief Bioinform* 17:628–641.
30. De Vito R, Bellio R, Trippa L, Parmigiani G (2016) Multi-study factor analysis. arXiv:1611.06350.
31. Shimodaira H (2000) Improving predictive inference under covariate shift by weighting the log-likelihood function. *J Stat Plann Inference* 90:227–244.
32. Sugiyama M, Krauledat M, Mãžller KR (2007) Covariate shift adaptation by importance weighted cross validation. *J Mach Learn Res* 8:985–1005.
33. Sugiyama M, et al. (2008) Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics* 60:699–746.
34. Pan SJ, Kwok JT, Yang Q (2008) Transfer learning via dimensionality reduction. *AAAI* 8:677–682.
35. Pan SJ, Yang Q (2010) A survey on transfer learning. *IEEE Trans Knowledge Data Eng* 22:1345–1359.
36. Dai W, Yang Q, Xue G-R, Yu Y (2007) Boosting for transfer learning. Proceedings of the 24th International Conference on Machine Learning (ICML '07), ed. Ghahramani Z (ACM, New York), pp 193–200.
37. Hashem S (1997) Optimal linear combinations of neural networks. *Neural Networks* 10:599–614.
38. Breiman L (1996) Stacked regressions. *Machine Learn* 24:49–64.
39. Clemen RT (1989) Combining forecasts: A review and annotated bibliography. *Int J Forecast* 5:559–583.
40. Wallis KF (2011) Combining forecasts–forty years later. *Appl Financial Econ* 21:33–41.
41. Trippa L, Waldron L, Huttenhower C, Parmigiani G (2015) Bayesian nonparametric cross-study validation of prediction methods. *Ann Appl Stat* 9:402–428.
42. Mullen KM, van Stokkum IHM (2012) *nnls: The Lawson-Hanson Algorithm for Non-Negative Least Squares (NNLS)*, R Package Version 1.4. Available at https://cran.r-project.org/web/packages/nnls/index.html. Accessed January 16, 2017.
43. Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 33:1.
44. Therneau TM, et al. (2010) rpart: Recursive Partitioning, R Package Version 3. Available at https://cran.r-project.org/web/packages/rpart/index.html. Accessed January 16, 2017.
45. Wright MN, Ziegler A (2017) Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *J Stat Softw* 77:1–17.
46. Venables WN, Ripley BD (2002) *Modern Applied Statistics with S* (Springer, New York), 4th Ed.
47. Hothorn T, Buehlmann P, Kneib T, Schmid M, Hofner B (2017) *mboost: Model-Based Boosting*, R Package Version 2.8-1. Available at https://cran.r-project.org/web/packages/mboost/index.html. Accessed January 16, 2017.
48. Viechtbauer W (2010) Conducting meta-analyses in R with the metafor package. *J Stat Softw* 36:1–48.