Received <day> <Month>, <year>; | Revised <day> <Month>, <year>; | Accepted <day> <Month>, <year>

DOI: xxx/xxxx

ARTICLE TYPE

Accounting for established predictors with the multi-step elastic net

Elizabeth C. Chase | Philip S. Boonstra

[1]Department of Biostatistics
University of Michigan, Ann Arbor, MI

Correspondence

*Elizabeth Chase, Email: ecchase@umich.edu

Present Address

SPH II, 1415 Washington Hts, Ann Arbor, MI, 48109

Abstract

We present the multi-step elastic net (MSN), which considers penalized regression with variables that can be qualitatively grouped based upon their degree of prior research support: established predictors (e.g. age, known biomarkers) vs. unestablished predictors (e.g. a panel of novel biomarkers or other hypothesized risk factors). The MSN chooses between uniform penalization of all predictors (the standard elastic net) and weaker penalization of the established predictors in a cross-validated framework, and includes the option to impose zero penalty on the established predictors. In simulation studies that reflect the motivating context, we show the comparability or superiority of the MSN over the standard elastic net, the IPF-Lasso, the sparse group lasso, and the group lasso, and investigate the importance of including the zero penalty. We use the MSN to update a prediction model for pediatric ECMO patient mortality.

KEYWORDS:

penalized regression, nested models, grouped data

1 | INTRODUCTION

Since Robert Tibshirani's creation of the lasso (1), dozens of extensions within the penalized regression framework have been developed: different families of penalties (2, 3), penalized regression with grouped data (4, 5, 6), the lasso within a hierarchical structure (7, 8), and many more. Here, we consider a modification of the elastic net that incorporates prior knowledge about potential predictors. When building a prediction model, the candidate predictors often differ in their underlying plausibility.

The relationship between smoking and lung cancer is a prime example. A patient's history of tobacco use is the driving factor behind whether to screen for lung cancer or not (9, 10), and, upon diagnosis, an estimated 80% of lung cancer cases in the United States are attributable to smoking (11). Among eight published multivariable lung cancer risk models we reviewed (12, 13, 14, 15, 16, 17, 18, 19), all included at least one predictor derived from smoking history; age and sex were the only other predictors about whose inclusion there was uniform agreement. The consensus regarding the association between smoking and lung cancer, combined with the biological rationale for such a link, suggests that any new lung cancer prediction model should necessarily adjust for some measure(s) of tobacco use.

To a lesser degree, this phenomenon occurs whenever an existing prediction model is subsequently updated with new, untested candidate predictors. Some examples include (i) adding a polygenic risk score to a prediction model for lead levels in the tibia (20, 21); (ii) adding a panel of pre-treatment cytokine measurements to a standard clinical model for the risk of radiation-induced lung or esophageal toxicity after treatment for lung cancer (22, 23); (iii) adding biomarkers to enhance a model for risk of surgical kidney injury (24). In all of these examples, the predictors in the first model had a measure of credibility supporting their inclusion in the second, updated model which the added predictors had not yet attained. An underlying assumption then is that the original factors should require less statistical justification to remain in the second, updated model. This assumption is often implicitly made; however, formally imposing such an assumption, as described in this paper, may improve performance of the final model, where "improve" means better prediction and, when selection of predictors is desired, higher true-positive/true-negative rates with the minimal sacrifice of sensitivity to these new, untested predictors.

Based on this idea, we propose a modification of the elastic net (25) that more formally accounts for the knowledge that some predictors have already been vetted in previous models for the same outcome, but which does not require quantification of this prior evidence. A classical application of the elastic net subjects all possible predictors to the same degree of penalization, regardless of our prior knowledge about them. Our approach qualitatively categorizes the predictors under consideration into two sets, comprised of those supported by prior research (established) and those that are new and relatively untested (unestablished). Through the introduction of additional tuning parameters, it selects from among equal or increasingly differential amounts of penalization on the two groups using standard cross validation techniques.

Some approaches for incorporating varying credibility between predictors have already been proposed (26, 27). Of these, the most relevant to our work is Boulesteix et al.'s 2017 development of the IPF-Lasso. The IPF-Lasso was created for the "omics" data setting, in which investigators have several categories of predictors with varying levels of credibility (i.e. clinical predictors, genetic data, metabalomics, proteomics, etc.). The user creates up to 5 categories of variables with different levels of penalization and inputs different degrees of penalization for each category. A lasso regression is fit, with cross-validation used to select the best combination of penalty factors. Our approach has some key differences. First, we force variables to be more decisively

divided, as either established or not, and we prespecify the amount of differential penalization that is explored. Second, we adapt the elastic net, rather than the standard lasso, which allows for a smoother blend of shrinkage and selection. Third, we include the possibility of zero penalization on the established predictors, while the IPF-Lasso only considers non-zero penalties. For a more thorough detailing of other solutions to varied penalization, we refer the reader to Boulesteix, et al.(2017).

Also related to this problem are penalized regression methods for grouped data, as in the grouped lasso (6). In these methods, all candidate predictors are grouped (e.g. dummy variables representing a single categorical predictor would comprise a group) and members are jointly included or fully excluded from the final model. Apart from our approach being defined for exactly two groups (established and unestablished predictors), the other distinguishing feature from existing grouped penalization methods is that group membership in our penalty is based only upon whether covariates have already been previously studied, and not upon any inherent statistical or logical relationship. For example, smoking history, family history of Lynch syndrome, and infection with schistosomiasis are well-established predictors of bladder cancer (28), but it would be unduly restrictive to require an updated bladder cancer risk model to contain all or none of these existing predictors. Penalized regression methods with a flexible group structure (e.g. the sparse group lasso (29)) vary the degree of within-group and between-group penalization, but still seek to solve a fundamentally different problem from ours by identifying groups that can be fully excluded from the model.

The structure of the paper is as follows. We will present the primary method, which we call the multi-step elastic net (MSN). We will compare the empirical properties of our two proposed methods to the elastic net, the IPF-LASSO extended to the elastic net setting, the group lasso, and the sparse group lasso using a simulation study. We will then demonstrate the utility of the MSN while building a mortality prediction model for pediatric ECMO patients. We conclude with a discussion.

2 | METHODS

Suppose we have a dataset containing n observations for p predictors, $\boldsymbol{\beta}$. Let $\mathbf{y}=(y_1, ..., y_n)^T$ be the outcome and $\mathbf{X}$ be the $n \times p$ design matrix. Let $\mathbf{y}$ be centered and let $\mathbf{X}$ be standardized. The original elastic net penalty minimizes the criterion:

$$L(\lambda, \alpha, \boldsymbol{\beta}) = \|\mathbf{y} - \boldsymbol{X\beta}\|_2^2 + (1-\alpha)\frac{\lambda}{2}\|\boldsymbol{\beta}\|_2^2 + \alpha\lambda\|\boldsymbol{\beta}\|_1 \quad (1)$$

where λ is a tuning parameter that controls the overall degree of penalization and α is a tuning parameter to control the mixture between ridge (L_2) penalization and lasso (L_1) penalization. Both α and λ are usually tuned through cross-validation, as described in Remark 1 below. Note that $\alpha = 1$ is equivalent to the lasso, while $\alpha = 0$ is equivalent to ridge regression (30). The elastic net is implemented for linear, logistic, and proportional hazards regressions in the glmnet R package (31).

Now, let $\boldsymbol{\beta_1}$ denote the well-established predictors–those with strong prior support in the literature–and let $\boldsymbol{\beta_2}$ denote the unestablished or untested predictors. We propose the criterion:

$$L(\lambda, \alpha, \phi, \boldsymbol{\beta}) = \|\mathbf{y} - \boldsymbol{X\beta}\|_2^2 + (1-\alpha)\frac{\lambda}{2}(\phi\|\boldsymbol{\beta_1}\|_2^2 + \|\boldsymbol{\beta_2}\|_2^2) + \alpha\lambda(\phi\|\boldsymbol{\beta_1}\|_1 + \|\boldsymbol{\beta_2}\|_1) \tag{2}$$

where ϕ is a tuning parameter to control the amount of penalization on the established predictors relative to the unestablished predictors. In general, we would expect ϕ to be less than or equal to 1, as the established predictors should be penalized less than the unestablished predictors. To select ϕ, for a fixed grid of α and λ, separately fit an elastic net for each of the following:

1. $\phi = 0$: no penalization on the established predictors.
2. $\phi = \frac{1}{16}$: established predictors receive $\frac{1}{16}$ of the penalization that unestablished predictors receive.
3. $\phi = \frac{1}{2}$: half-penalization on the established predictors.
4. $\phi = 1$: the standard elastic net, with equal and standard penalization on all predictors.

Select the best of these four models using cross-validation, as described in Remark 1 below. We note that, by allowing for the possibility of equal penalization ($\phi = 1$), this approach should, in large samples, be non-inferior to the classical elastic net.

Remark 1 Although the MSN adds an additional tuning parameter, ϕ, in addition to λ and α, it can still be straightforwardly implemented in the glmnet function, as demonstrated in the provided code. We extend the use of five-fold cross-validation (FFCV) to select the values of each. Specifically, for fitting the standard elastic net, glmnet uses efficient coordinate-descent algorithms over a grid of λ values, at a fixed value of α. In FFCV, the data are partitioned into five 'folds,' and the model at each value of λ is fit to each of the five combinations of four folds. The model is then tested against the remaining held-out fold using some loss function, e.g. deviance. The selected λ is the one that minimizes the held-out loss, averaged over the five combinations. Ideally, multiple such partitions are constructed, and the average over five combinations are, themselves, averaged over multiple partitions, to smooth out results. For selecting α, one then profiles this process across a grid of αs to be tested, using an identical set of partitions. For MSN, we further profiled over the set of four ϕ values. For the example, with a grid of three values of α at 0, 0.1, and 0.2, which is what we used in our numerical studies and example, fitting a MSN model requires profiling over and selecting from $3 \times 4 = 12$ elastic nets. We constructed 25 unique partitions for each elastic net.

For comparison, we also present the penalties for the IPF-Lasso, sparse group lasso (SGL), and group lasso (GLASSO).

IPF-Lasso (27) The IPF-Lasso, applied to the present context, would minimize:

$$L(\lambda, \phi_1, \phi_2, \boldsymbol{\beta}) = \|\mathbf{y} - \boldsymbol{X\beta}\|_2^2 + \lambda(\phi_1\|\boldsymbol{\beta_1}\|_1 + \phi_2\|\boldsymbol{\beta_2}\|_1) \tag{3}$$

Note that the IPF-Lasso only uses L_1 penalization, and that $\phi_1, \phi_2 > 0$. Decision-making about the best combination of ϕ_1, ϕ_2 is left to the discretion of the investigator, although it could be implemented in a cross-validated setting, as we do with the MSN. In order to give a fair comparison between the IPF-Lasso and the MSN in our simulation study, we fixed ϕ_2 at 1 and used cross-validation to select the best of options 2-4 listed above for ϕ_1.

We also extended the IPF-Lasso to the elastic net setting (IPF-EN) in order to discern the importance of the zero penalty option. As executed here, the IPF-EN minimizes the same criterion as in equation 2, but without option 1 (zero penalization); it selects the best of options 2-4.

SGL, GLASSO (29, 6) The SGL, applied to the present context, would minimize:

$$L(\lambda, \alpha, \boldsymbol{\beta}) = \|\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + (1-\alpha)\lambda(\sqrt{p_1}\|\boldsymbol{\beta_1}\|_2^2 + \sqrt{p_2}\|\boldsymbol{\beta_2}\|_2^2) + \alpha\lambda(\|\boldsymbol{\beta_1}\|_1 + \|\boldsymbol{\beta_2}\|_1) \tag{4}$$

where p_1, p_2 are the number of coefficients in the established and unestablished groups, respectively. Here, we set $\alpha = 0.95$, as recommended by Simon et al. (2013). The GLASSO minimizes the same criterion as above, but with α always equal to 0 (6).

3 | SIMULATION STUDY

We evaluated and compared our proposed MSN penalty against four existing approaches (elastic net, IPF-Lasso, SGL, and GLASSO) and one "investigatory" approach (the IPF-EN) in a binary outcome setting using logistic regression. We constructed eighteen scenarios that varied in sample size (n = 200 or n = 1000), the number of covariates (10-20 established predictors; 30-480 unestablished predictors), correct specification of the established predictors, and magnitude of log-odds ratios (ORs). In all cases, the predictors were sampled from a multivariate normal distribution with mean zero, variance 1, and a compound-symmetric correlation structure with value 0.2. Further, in each scenario, the distribution of predictors and the true log-OR values were such that the true model AUC was 0.8 (for more information on AUC construction, see Supplement 1). We simulated 500 replicates for each scenario. The eighteen scenarios are described in table 1 .

We assessed performance using a range of prediction and estimation metrics. For prediction, we evaluated the AUC and Brier score on an independent validation dataset of size 1000, drawn from the same population. These are respectively defined as

TABLE 1 *Simulation study settings*

Scenario	n	$p_{established}$	Magnitude	$p_{unestablished}$	Magnitude
1A	200	10	all 0.26	30	0
1Ap	1000	10	all 0.26	30	0
1B	200	10	all 0.2	30	one 0.6, rest 0
1Bp	1000	10	all 0.2	30	one 0.6, rest 0
1C	200	10	all 0.25	30	five 0.05, rest 0
1Cp	1000	10	all 0.25	30	five 0.05, rest 0
2A	200	10	all 0.26	90	0
2Ap	1000	10	all 0.26	90	0
2B	200	10	all 0.2	90	one 0.6, rest 0
2Bp	1000	10	all 0.2	90	one 0.6, rest 0
2C	200	10	all 0.25	90	five 0.05, rest 0
2Cp	1000	10	all 0.25	90	five 0.05, rest 0
3A	200	20	half 0.26, half 0	480	0
3Ap	1000	20	half 0.26, half 0	480	0
3B	200	20	half 0.2, half 0	480	one 0.6, rest 0
3Bp	1000	20	half 0.2, half 0	480	one 0.6, rest 0
3C	200	20	half 0.25, half 0	480	five 0.05, rest 0
3Cp	1000	20	half 0.25, half 0	480	five 0.05, rest 0

AUC Let $y_1, y_2, ..., y_n$ be the true outcome, and let $\hat{y_1}, \hat{y_2}, ..., \hat{y_n}$ be the predicted outcome. For a randomly selected $y_i = 1$ and $y_j = 0$, the AUC is the probability that

$$\hat{y_i} > \hat{y_j} \tag{5}$$

Brier score The Brier Score is calculated

$$\frac{1}{n}\Sigma_{i=1}^{n}(\hat{y_i} - y_i)^2. \tag{6}$$

To assess estimation, we recorded the root mean squared error (rMSE). Results are presented in figures 1 -4 .

Across all three metrics, the MSN performed well. In the smaller sample scenarios (figures 1 and 3), the MSN and IPF-EN did best in all scenarios for all metrics, followed by the IPF-Lasso. As sample size increased to 1000, the performance of the MSN, IPF-EN, and IPF-Lasso became more equivalent, which is to be expected; however, the SGL and GLASSO still hung behind. For all methods, performance appeared to be worst in the B type of scenarios, in which the unestablished covariates are all zero, except for one extremely large covariate. Misspecification of the established covariates (the 3A, 3B, 3C scenarios) hurt the performance of the MSN/IPF-EN/IPF-Lasso, but they still outperformed the EN, SGL, and GLASSO, suggesting that even in the presence of established covariate misspecification, our method would be a fine choice. In terms of estimation (figures3 and 4), the MSN, IPF-EN, and IPF-Lasso had much more variability in rMSE than the EN, SGL, and GLASSO. This may be

because of the discreteness of the ϕ tuning parameter, which may have added an element of jaggedness in the covariate estimates that the SGL and GLASSO did not have.

On all metrics and all scenarios, the IPF-EN and MSN performed identically: the zero penalization option was never selected for the MSN, so the two methods reduced to the same thing. In our experience, the established covariates have to be remarkably powerful to make zero penalization a contender, and in these scenarios, it seems they were not impressive enough relative to the unestablished predictors to make zero penalization attractive.

4 | DATA EXAMPLE

We apply the MSN to build a prediction model of mortality among pediatric ECMO patients receiving respiratory support via ECMO (extracorporeal membrane oxygenation). In 2016, Barbaro et al. built the Ped-RESCUERS prediction model for short-term mortality risk for children on ECMO, using data from 1,611 patients in the Extracorporeal Life Support Organization (ELSO) registry between 2009 and 2012 (32). They selected nine predictors for the initial Ped-RESCUERS: time from admission to initiation of ECMO, time from intubation to initiation of ECMO, arterial pH, arterial carbon dioxide [PaCO2], mean airway pressure (separately for conventional or high frequency oscillatory ventilation), primary diagnosis (three variables), the presence of malignancy as a comorbidity, and pre-ECMO treatment with milrinone. Altogether, these predictors comprised eleven “established" covariates.

Pre-ECMO biometric measurements of renal, hepatic, neurologic and hematologic dysfunction are not typically recorded in the ELSO registry but may be associated with short-term mortality on ECMO. In 2018, Barbaro et al. updated Ped-RESCUERS with additional data from ten such biometric variables, the “unestablished" covariates, collected across a non-overlapping cohort of 178 ECMO patients. The variables were bilirubin level, alanine aminotransferase [ALT] level, white blood cell count (both too low [leukopenia] and too high [leukocytosis]), low platelet levels [thrombocytopenia], international normalized ratio [INR], vasoactive infusion score [VIS], lactate levels, ratio of arterial oxygen partial pressure to fractional inspired oxygen [PF Ratio], abnormal pupil response, and acute kidney injury. Because the number of potential predictors (11 established plus 11 unestablished) is high relative to sample size (178), it is crucial to incorporate the knowledge that the eleven established covariates have already been used in an existing model for the same outcome.

Because missing data were present, we used MICE to impute 25 datasets, and then fit an elastic net, IPF-Lasso, IPF-EN, MSN, GLASSO, and SGL to each imputed dataset. For computational efficiency, we treated each imputed dataset as a separate replicate for cross-validation to select the tuning parameters. AUC and Brier score were averaged across imputations/replicates, and we

used the mean of the coefficient estimates across the 25 imputations as our coefficient estimate. Estimates of the estimated coefficients are presented in tables 2 and 3 . Predictive performance of each method is presented in table 4 .

TABLE 2 *Established covariate effect estimates, ECMO Study*

Variable	EN	IPF-EN	MSN	IPF-Lasso	SGL	GLASSO
Admitted hours pre-ECMO (log)	0.03	0.03	0.03	0.00	0.00	0.00
Intubated hours pre-ECMO (log)	0.34	0.34	0.34	0.37	0.31	0.42
pH	-0.03	-0.03	-0.03	0.02	-0.00	-0.02
$PaCO_2$	0.05	0.05	0.05	0.01	0.00	0.02
MAP (CMV), cm H_2O	0.01	0.01	0.01	0.00	0.00	0.00
MAP (HFOV), cm H_2O	0.03	0.03	0.03	0.01	0.00	0.00
Malignancy	0.01	0.01	0.01	0.00	0.00	0.00
Asthma diagnosis	-0.12	-0.12	-0.12	-0.02	-0.00	-0.02
Bronchiolitis diagnosis	-0.07	-0.07	-0.07	-0.03	-0.01	-0.05
Pertussis diagnosis	0.23	0.23	0.23	0.23	0.20	0.25
Milrinone	0.00	0.00	0.00	0.00	0.00	0.00

TABLE 3 *Unestablished covariate effect estimates, ECMO Study*

Variable	EN	IPF-EN	MSN	IPF-Lasso	SGL	GLASSO
Bilirubin, mg/dL (log)	0.21	0.21	0.21	0.15	0.16	0.18
ALT, U/L (log)	0.52	0.52	0.52	0.67	0.64	0.70
Leukocytosis (log)	0.09	0.09	0.09	0.05	0.06	0.08
Leukopenia (log)	-0.08	-0.08	-0.08	-0.02	-0.02	-0.05
Thrombocytopenia (log)	0.03	0.03	0.03	0.00	0.00	0.00
INR	0.06	0.06	0.06	0.01	0.01	0.02
VIS (log)	0.02	0.02	0.02	0.00	0.00	0.00
Lactate, mMol/L (log)	0.31	0.31	0.31	0.34	0.31	0.40
PF-ratio (log)	-0.14	-0.14	-0.14	-0.07	-0.05	-0.13
Abnormal pupillary response	-0.01	-0.01	-0.01	0.00	0.00	0.00
pre-ECMO kidney injury	0.01	0.01	0.01	0.00	0.00	0.00

TABLE 4 *Model Performance on ECMO Dataset*

Method	AUC	Brier
EN	0.83	0.14
IPF-EN	0.83	0.14
IPF-Lasso	0.82	0.15
MSN	0.83	0.14
GLASSO	0.82	0.14
SGL	0.81	0.15

For this example, the MSN, IPF-EN, and EN performed identically, suggesting that for these data, applying equal penalization to the established and unestablished covariates was always the best option. All methods were in agreement that hours of intubation pre-ECMO and pertussis diagnosis were strong established predictors; among the unestablished predictors, all methods agreed that bilirubin, ALT, lactate, and PF-ratio were important predictors of mortality. However, the MSN, IPF-EN, and EN also found asthma diagnosis, $PaCO_2$, and INR to be strong predictors, while the IPF-LASSO, SGL, and GLASSO did not. Predictive performance of all six methods was roughly comparable. These results are largely in concordance with a previous analysis of these data performed by Boonstra and Barbaro using historical priors (33).

5 | DISCUSSION

We present an extension of the elastic net, called the Multi-Step Elastic Net (MSN), which is intended for use when a subset of the predictors under consideration has already been evaluated in previous models. Our method leverages this limited information to improve upon the elastic net's predictive and estimating performance. It can easily be implemented using existing R packages and, because it requires relatively little additional user-input beyond the specification of 'established' and 'unestablished,' is fairly automatic to implement. For researchers with prior knowledge about the credibility of their predictors, the MSN provides a simple way to take that knowledge into account and improve model performance.

This work was as much an exploration of other simple approaches for dealing with prior knowledge as it was a presentation of our new method. Here, our findings were surprising. On the whole, the SGL and GLASSO did not perform well in this context, and we would advise against using either of these methods for this particular problem. We were also curious if there would be any differences between the MSN and the IPF-Lasso, and there were. We think the elastic net's smoother shrinkage and inclusion of groups of correlated predictors may make it preferable to the IPF-Lasso in this context. In initial exploration of the MSN, we gave it the full [0, 1] α sequence in 0.1 increments to choose from. However, it was so rare for $\alpha > 0.2$ to be selected that we ended up restricting to [0, 0.2] for computational expediency. The IPF-Lasso's restriction of only $\alpha = 1$ may be limiting its performance.

Two other key differences between our approach and the IPF-Lasso were the inclusion of a zero penalization option and the restriction to only two groupings: established and unestablished. The zero penalization did not seem to be a major difference. As discussed, it was never selected in our simulations or in our application. However, we believe that including the zero penalization option is important–the difference between zero and a small number is vast, and for cases where the established predictors are extremely strong, it could be an important tool. As for the use of only two groupings, we are less tied. We think that having only two groups is simpler, and a more clear-cut decision may make our method more attractive. In addition, using only two groupings

means that our method can rely on just one tuning parameter, ϕ, as opposed to multiple. In penalty exploration, we found that including more than one tuning penalty parameter for two groups was redundant–the way that λ is selected means that the ratio between the two groups is all that matters to varying penalization. (This was also our rationale for not including the option of infinite penalization on the unestablished covariates–we found that $\phi = \frac{1}{16}$ generally worked out to be the same thing.) With more than two groups, though, multiple ϕ parameters would be necessary, adding another layer of tuning and cross-validation complexity. Future work may want to investigate the utility of including more than two credibility groups.

We did not consider a fully Bayesian approach that incorporates information directly via prior distributions on the established predictors' coefficients. One reason against taking such an approach is that the coefficient estimates corresponding to the established predictors in the previously fitted models will generally differ from those of the current, larger model, both numerically and in interpretation, and particularly in non-linear models such as logistic regression. This is not a consequence of small sample size but rather misspecification in nested models (34). Rather, we focus here on the setting in which it is known that a subset of predictors are likely to be associated with the outcome due to having been included in prior models but, for reasons mentioned above, there is considerable uncertainty about the true magnitude of these associations. The MSN provides a less assumptive way to automatically take this prior knowledge into account and still improve performance.

One drawback of this partial approach, though, is that our method may perform poorly when established covariates are mediators for unestablished covariates. We saw a hint of this in the ECMO application. The key difference between our results and those of Boonstra and Barbaro (who used a fully Bayesian approach) was that we found $PaCO_2$ to have a harmful effect on mortality, while Boonstra and Barbaro found it to be mostly zero (33). $PaCO_2$ is a mediator of lactate's effect on acidosis, which increases risk of mortality. Therefore, in truth $PaCO_2$ probably should be zero, but our model was unable to pick up on this causal relationship because it only "knew" that $PaCO_2$ should be underpenalized based on prior research. When established and unestablished covariates have known, complex causal relationships, researchers may wish to turn to a more sophisticated Bayesian analysis. If that is not the case, though, we believe our method can strike an appealing balance between total naiveté about previous research and strong Bayesian assumptions.

Future extensions of this work might offer more options for differential penalization than the four combinations that the MSN considers or additional credibility groups. In addition, it might be interesting to develop ways to work with varying penalties when also dealing with truly grouped or hierarchical data. It may be possible to use the same heuristic that the MSN uses, but with the group lasso or sparse group lasso instead of the elastic net. Future work is needed to assess the empirical performance of that approach.

The MSN provides a simple extension of the elastic net to handle predictors with different degrees of prior support. Use of this method has the potential to improve predictive performance while maintaining variable selection accuracy.

6 | SUPPLEMENT 1: CONSTRUCTING AUC FOR SIMULATION SCENARIOS

For the simulation study, we chose the covariate effect size in order to produce a true model AUC of 0.8. We selected a sample of $n = 5000$ where the predictors were sampled from a multivariate normal distribution with mean zero, variance 1, and a compound-symmetric correlation structure with value 0.2 (the same specifications as for the simulation study's design matrices, but with a larger sample size). Then, we input a guess at an appropriate covariate vector and used it to simulate an outcome vector for the binary logistic regression setting. We calculated AUC using the simulated outcome vector (the truth) and the predicted probabilities (the product of the simulated design vector and the inputted coefficient vector). We repeated this process, updating the coefficient vector as necessary, until the coefficient vector reliably produced an AUC of approximately 0.8.

References

[1] Tibshirani R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B.* 1996;58:267-288.

[2] Zhao P, Rocha G, Yu B. The composite absolute penalties family for grouped and hierarchical variable selection. *Annals of Statistics.* 2009;37(6):3468-3497.

[3] Zhao P, Yu B. Stagewise lasso. *Journal of Machine Learning Research.* 2007;8:2701-2726.

[4] Breheny P. The group exponential lasso for bi-level variable selection. *Biometrics.* 2015;71:731-740.

[5] Jacob L, Obozinski G, Vert JP. Group lasso with overlap and graph lasso. In: :433-440; 2009.

[6] Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B.* 2006;68:49-67.

[7] Bien J, Taylor J, Tibshirani R. A lasso for hierarchical interactions. *Annals of Statistics.* 2013;41(3):1111-1141.

[8] Yuan M, Joseph V Roshan, Zou H. Structured variable selection and estimation. *Annals of Applied Statistics.* 2009;3(4):1738-1757.

[9] Wender R, Fontham ET, Barrera E, et al. American cancer society lung cancer screening guidelines. *CA: A Cancer Journal for Clinicians.* 2013;63:106-117.

[10] Moyer VA. Screening for lung cancer: U.S. preventative services task force recommendation statement. *Annals of Internal Medicine.* 2014;160:330-338.

[11] CDC . *What Are the Risk Factors for Lung Cancer?.* 2017.

[12] Bach PB, Kattan MW, Thornquist MD, et al. Variations in lung cancer risk among smokers. *Journal of the National Cancer Institute.* 2003;95:470-478.

[13] Cassidy A, Myles JP, Tongeren M, et al. The LLP risk model: an individual risk prediction model for lung cancer. *British Journal of Cancer.* 2008;98:270.

[14] Etzel CJ, Kachroo S, Liu M, et al. Development and validation of a lung cancer risk prediction model for African-Americans. *Cancer Prevention Research.* 2008;1:255-65.

[15] Spitz MR, Hong WK, Amos CI, et al. A risk model for prediction of lung cancer. *Journal of the National Cancer Institute.* 2007;99:715-726.

[16] Park S, Nam B-H, Yang H-R, et al. Individualized risk prediction model for lung cancer in Korean men. *PLOS One.* 2013;8:e54823.

[17] Tammemägi MC, Katki HA, Hocking WG, et al. Selection criteria for lung-cancer screening. *New England Journal of Medicine.* 2013;368:728-736.

[18] Hoggart C, Brennan P, Tjonneland A, et al. A risk model for lung cancer incidence. *Cancer Prevention Research.* 2012;5(6):834-846.

[19] Marcus MW, Chen Y, Raji OY, Duffy SW, Field JK. LLPI: Liverpool lung cancer risk prediction model for lung cancer incidence. *Cancer Prevention Research.* 2015;8:570-575.

[20] Park SK, Mukherjee B, Xia X, et al. Bone lead level prediction models and their application to examining the relationship of lead exposure and hypertension in the third National Health and Nutrition Examination Survey (NHANES-III). *Journal of Occupational and Environmental Medicine.* 2009;51:1422.

[21] Cheng W, Taylor JMG, Vokonas PS, Park SK, Mukherjee B. Improving estimation and prediction in linear regression incorporating external information from an established reduced model. *Statistics in Medicine.* 2018;37(9):1515-1530.

[22] Hawkins PG, Boonstra PS, Hobson ST, et al. Radiation induced lung toxicity in non-small-cell lung cancer: understanding the interactions of clinical factors and cytokines with the dose-toxicity relationship. *Radiotherapy and Oncology.* 2017;125:66-72.

[23] Hawkins PG, Boonstra PS, Hobson ST, et al. Prediction of radiation esophagitis in non-small cell lung cancer using clinical factors, dosimetric parameters, and pretreatment cytokine levels. *Translational oncology.* 2018;11:102-108.

[24] Kerr KF, Meisner A, Thiessen-Philbrook H, Coca SG, Parikh CR. Developing risk prediction models for kidney injury and assessing incremental value for novel biomarkers. *Clinical Journal of the American Society of Nephrology.* 2014;9(8):1488-1496.

[25] Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B.* 2005;67:301-320.

[26] Bin R De, Sauerbrei W, Boulesteix AL. Investigating the prediction ability of survival models based on both clinical and omics data: two case studies. *Technical Report, Department of Statistics, University of Munich.* 2014;153.

[27] Boulesteix AL, Bin R De, Jiang X, Fuchs M. IPF-LASSO: Integrative L1-Penalized Regression with Penalty Factors for Prediction Based on Multi-Omics Data. *Computational and Mathematical Methods in Medicine.* 2017;.

[28] ACS . *Bladder Cancer Risk Factors.* 2016.

[29] Simon N, Friedman J, Hastie T, Tibshirani R. A sparse-group lasso. *Journal of Computational and Graphical Statistics.* 2013;22(2):231-245.

[30] Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics.* 1970;12(1):55-67.

[31] Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software.* 2010;33(1):1-22.

[32] Barbaro R, Boonstra P, Paden M, et al. Development and validation of the pediatric risk estimate score for children using extracorporeal respiratory support (PED-RESCUERS). *Intensive Care Medicine.* 2016;42:879-888.

[33] Boonstra PS, Barbaro RP. Incorporating historical models with adaptive Bayesian updates. *Biostatistics.* 2018;.

[34] Robinson LD, Jewell NP. Some surprising results about covriate adjustment in logistic regression models. *International Statistical Review.* 1991;59(2):227-240.

FIGURE 1 *AUC and Brier score of IPF-Lasso, IPF-EN, MSN, SGL, and GLASSO, log-scaled relative to the elastic net, n = 200*

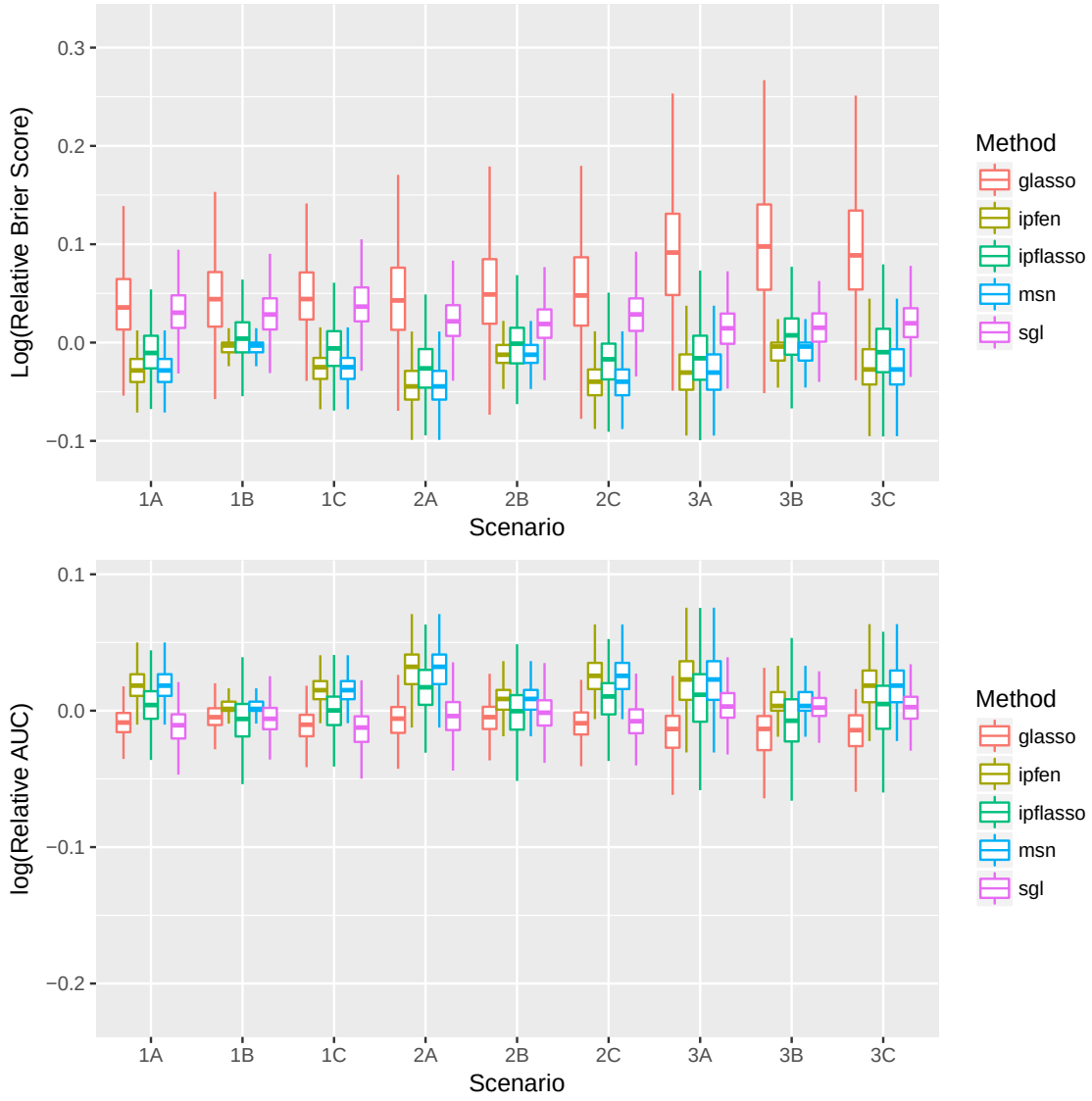

FIGURE 2 *AUC and Brier score of IPF-Lasso, IPF-EN, MSN, SGL, and GLASSO, log-scaled relative to the elastic net, n = 1000*

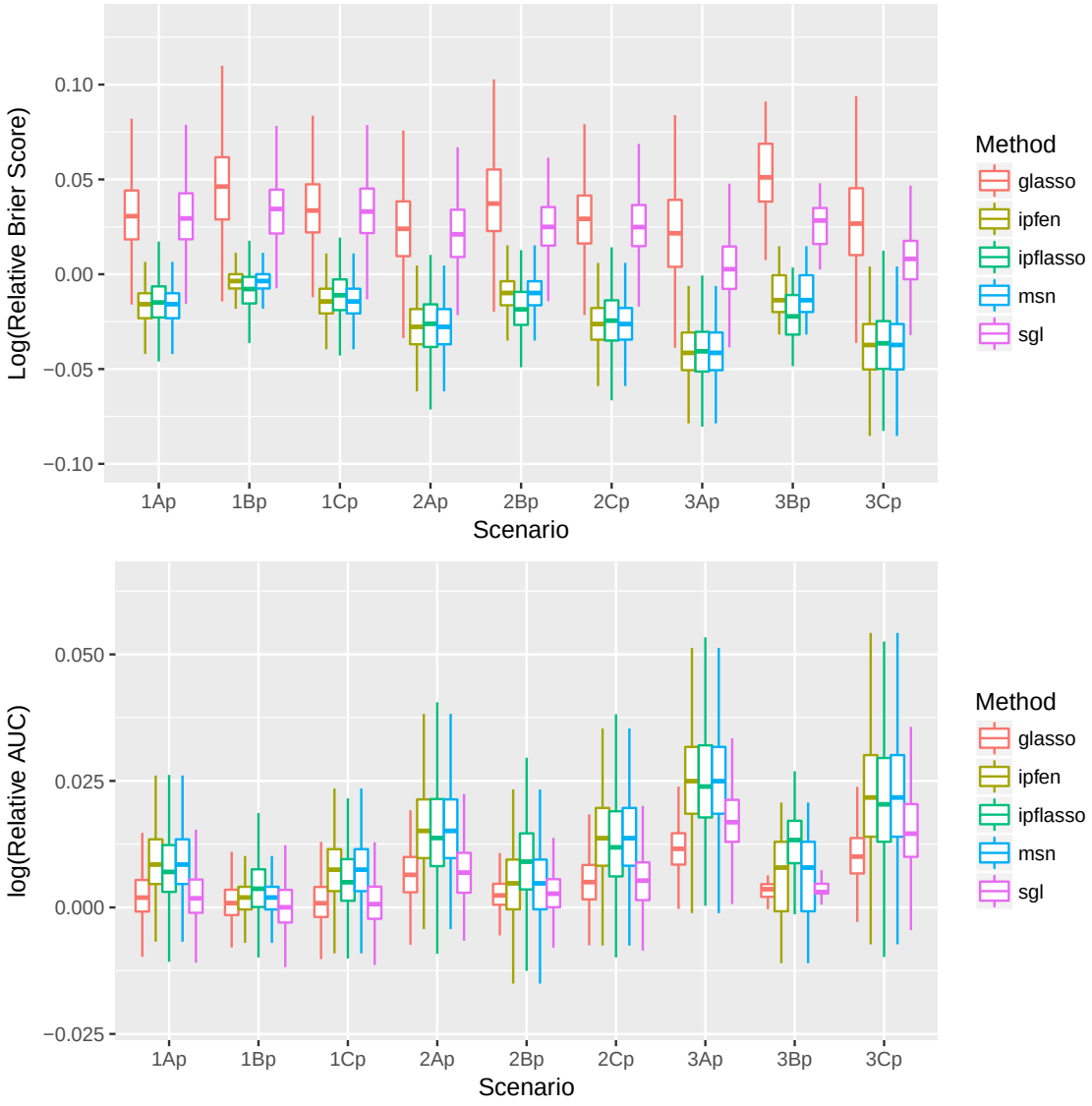

FIGURE 3 *rMSE of IPF-Lasso, IPF-EN, MSN, SGL, and GLASSO, log-scaled relative to the elastic net, n = 200*

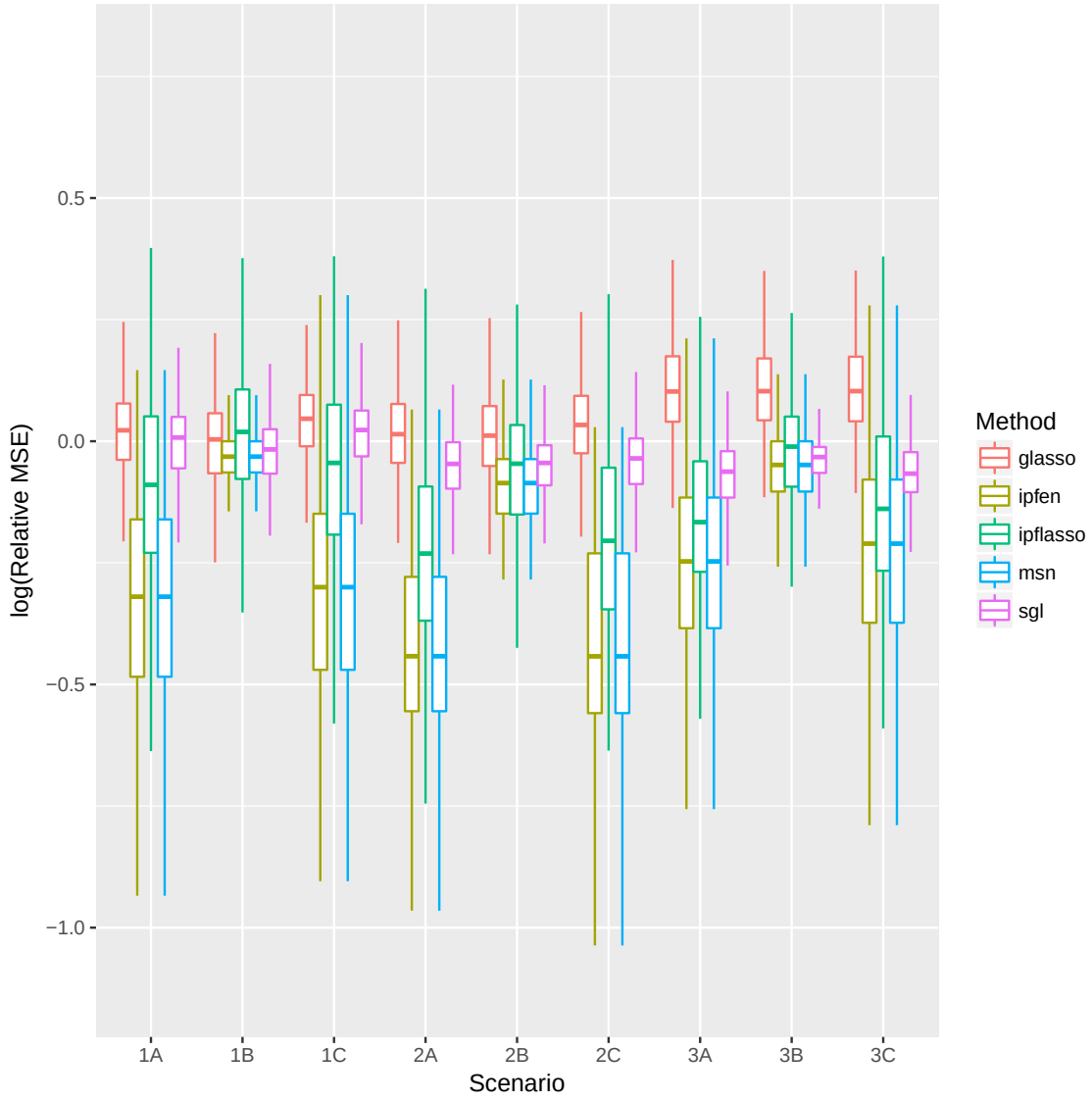

FIGURE 4 *rMSE of IPF-Lasso, IPF-EN, MSN, SGL, and GLASSO, log-scaled relative to the elastic net, n = 1000*