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**

Multivariable models for prediction or estimating associations with an outcome are rarely built in isolation. Instead, they are based upon a mixture of covariates that have been evaluated in earlier studies (e.g. age, sex, or common biomarkers) and covariates that were collected specifically for the current study (e.g. a panel of novel biomarkers or other hypothesized risk factors). For that context, 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 vs. unestablished predictors. 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 Integrative LASSO with Penalty Factors, the sparse group lasso, and the group lasso, and we investigate the importance of not penalizing the established predictors at all. We demonstrate the MSN to update a prediction model for pediatric ECMO patient mortality.



**KEYWORDS:**

penalized regression, nested models, grouped data, lasso, grouped lasso

## 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 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 prediction and estimation performance of the final model.

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.

We suspect that this method, or something like it, may already be used in practice; e.g. placing no penalty on the established predictors and full penalization on the unestablished predictors. We have used an ad hoc approach like this in our own research, as have several of our colleagues (22, 23). A key motivation of this project was to learn more about the empirical properties of this and related approaches and to make recommendations accordingly. Therefore, the aims of this paper are twofold: to propose our heuristic for modifying of the elastic net, which we call the multi-step elastic net (MSN), and also to explore the performance of the MSN and other methods for incorporating prior research in penalized regression.

Some formal 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 Integrative Lasso with Penalty Factors (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 five 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, instead of five categories, 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 may 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 MSN in Section 2. In Section 3, we will compare the empirical properties of our proposed method to the elastic net, the IPF-Lasso, the IPF-LASSO extended to the elastic net setting, the group lasso, the sparse group lasso, and an elastic net with zero penalization on the established predictors, using a simulation study.

We will then demonstrate the utility of the MSN while building a mortality prediction model for pediatric ECMO patients in Section 4. Section 5 concludes with a discussion of our findings in contrast with existing approaches as well as some limitations of the MSN.



## 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 $\lambda$ is a tuning parameter that controls the overall degree of penalization and $\alpha$ is a tuning parameter to control the mixture between ridge ($L_2$) penalization and lasso ($L_1$) penalization. Both $\alpha$ and $\lambda$ 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, 32).

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:

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

where $\phi$ is a tuning parameter to control the amount of penalization on the established predictors relative to the unestablished predictors. In general, we would expect $\phi$ to be less than or equal to 1, as the established predictors should be penalized less than the unestablished predictors. To select $\phi$, for a fixed grid of $\alpha$ and $\lambda$, 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. (See Remark 2 for an explanation of this choice.)
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, $\phi$, in addition to $\lambda$ and $\alpha$, it can still be straightforwardly implemented in the glmnet function, as demonstrated in the provided code (https://github.com/psboonstra/MSEN). 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 $\lambda$ values, at a fixed value of $\alpha$. In FFCV, the data are partitioned into five 'folds,' and the model at each value of $\lambda$ 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 $\lambda$ 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 $\alpha$, one then profiles this process across a grid of $\alpha$s to be tested, using an identical set of partitions. For MSN, we further profiled over the set of four $\phi$ values. For example, with a grid of three values of $\alpha$ 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:

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

Note that the IPF-Lasso only uses $L_1$ penalization, and the method requires that $\phi_1, \phi_2$ are both strictly greater than 0. Decision-making about the best combination of $\phi_1, \phi_2$ 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 $\phi_2$ at 1 and used cross-validation to select the best of options 2-4 listed above for $\phi_1$.

**Remark 2** The choice of $\phi = \frac{1}{16}$ as a penalty may seem to be minimally different from $\phi = 0$. However, from equation 2, it can be seen that, for a fixed value of $\alpha$, the penalty on $\beta_1$ is proportional to $\lambda\phi$, whereas the penalty on $\beta_2$ is proportional to $\lambda$. When $\phi = 0$, $\lambda$ can be made arbitrarily large without affecting the shrinkage of $\beta_1$, while with $\phi = \frac{1}{16}$, increasing $\lambda$ will always shrink both $\beta_1$ and $\beta_2$. To quantify the importance of this distinction, we included both a small but strictly positive $\phi$ and $\phi = 0$, and we extended the IPF-Lasso, which assumes strictly positive values of $\phi$, to the elastic net setting (IPF-EN) in order to disentangle the importance of the zero penalty vs. the lasso-only penalty. As executed here, the IPF-EN minimizes the same criterion as in equation 2, but without option 1 (zero penalization on the established predictors); it selects the best of options 2-4. The IPF-EN is disallowed from choosing $\phi = 0$, and thus any substantive differences between the MSN and the IPF-EN will be due to the addition of $\phi = 0$, while any substantive difference between the IPF-EN and the IPF-Lasso will be due to differences in structure of the penalty functions.

In addition, a reviewer suggested that we also evaluate the method that never penalizes the established predictors, i.e. always sets $\phi = 0$, because this may be another common ad hoc approach for this problem. We call this method the Auto-Zero.

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

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

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. (29). The GLASSO minimizes the same criterion as above, but with $\alpha$ 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 two exploratory approaches (the IPF-EN and Auto-Zero) in a binary outcome setting using logistic regression. We constructed twelve scenarios that varied in the number of covariates (10-20 established predictors; 30-480 unestablished predictors), whether or not the established predictors were correctly identified (the true log-odds ratios [ORs] are non-zero), and magnitude of log-ORs. All scenarios were performed at n = 200 and n = 1000. 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 and the intercept was -1.39, which corresponds to a population prevalence of 0.2. We simulated 500 replicates for each scenario. The 12 scenarios, repeated under the two sample size configurations, are described in Table 1 .

**TABLE 1** *Simulation study settings describing the generating logistic regression model. The second and fourth columns give the number of established and unestablished predictors, respectively, and the third and fifth columns give the magnitudes of the log-odds ratios in these groups.*

| Scenario | $p_{established}$ | Magnitude | $p_{unestablished}$ | Magnitude |
|---|---|---|---|---|
| 1A | 10 | all 0.26 | 30 | 0 |
| 1B | 10 | all 0.2 | 30 | one 0.6, rest 0 |
| 1C | 10 | all 0.25 | 30 | five 0.05, rest 0 |
| 2A | 10 | all 0.26 | 90 | 0 |
| 2B | 10 | all 0.2 | 90 | one 0.6, rest 0 |
| 2C | 10 | all 0.25 | 90 | five 0.05, rest 0 |
| 3A | 20 | half 0.26, half 0 | 480 | 0 |
| 3B | 20 | half 0.2, half 0 | 480 | one 0.6, rest 0 |
| 3C | 20 | half 0.25, half 0 | 480 | five 0.05, rest 0 |
| 4A | 20 | all 0.13 | 480 | 0 |
| 4B | 20 | all 0.1 | 480 | one 0.6, rest 0 |
| 4C | 20 | all 0.13 | 480 | five 0.05, rest 0 |

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 follows:

**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). The code used to create the simulation, run all methods, and process results is accessible on GitHub (https://github.com/psboonstra/MSEN).

Results are presented in Figures 1 -2 . 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 fitted model. Therefore, we only present the results for the MSN. Across all three metrics, the MSN and IPF-EN performed well. For the smaller sample size (Figure 1 ), the MSN and IPF-EN did best in all scenarios for all metrics, followed by the IPF-Lasso and the Auto-Zero. As sample size increased to 1000 (Figure 2 ), the performance of the MSN, IPF-EN, IPF-Lasso, and Auto-Zero became more similar, which is to be expected; however, the SGL and GLASSO still lagged behind. For all methods, performance appeared to be worst in the B type of scenarios, in which the unestablished coefficients are all zero, except for one extremely large coefficient. Even when some of the established predictors were actually zero (the 3A, 3B, 3C scenarios), it had little effect on the performance of the MSN/IPF-EN/IPF-Lasso, suggesting that even when prior research on the established covariates is wrong, our method would be a fine choice. In terms of estimation, the MSN, IPF-EN, IPF-Lasso, and Auto-Zero had much more variability in rMSE than the EN, SGL, and GLASSO, and at the smaller sample size, the Auto-Zero performed very poorly. We hypothesize that both of these features may actually be caused by the reduced shrinkage that these four methods place on some of the predictors. The lack of shrinkage increases the variability in the estimates of the established predictors and, especially for the Auto-Zero–which imposes no shrinkage at all–may cause them to be biased upwards. For the MSN, IPF-EN, and IPF-Lasso, the slightly reduced shrinkage on the established predictors generally seems to improve rMSE, albeit with increased variability, but the total absence of shrinkage, as in the Auto-Zero, seems to become a liability at smaller sample sizes. This may also explain why the zero penalization option was never selected for the MSN.

## 4 | DATA EXAMPLE

We applied 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 (33). 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 2019, Barbaro et al. updated Ped-RESCUERS with additional data from eleven such biometric variables, the "unestablished" covariates, collected across a non-overlapping cohort of 178 ECMO patients (34). 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 ($n = 178$), it is crucial to incorporate the knowledge that the eleven established covariates have already been used in an existing model for the same outcome.

There was sporadic missingness in the predictors (about 4% of the 178*22 datapoints were missing). To account for this, we used multiple imputation with chained equations to impute 25 datasets (35), and then fit an elastic net, IPF-EN, Auto-Zero, MSN, IPF-Lasso, SGL, and GLASSO 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 coefficients are presented in Tables 2 and 3 . Predictive performance of each method is presented in Table 4 .

For this example, the MSN and IPF-EN again performed identically, suggesting that zero penalization was still not an attractive option (we omitted the IPF-EN's results because of this). Another immediate observation is that the Auto-Zero's estimates of the established predictors were often substantially larger than any of the other methods under consideration, while its estimates of the unestablished predictors were often smaller. That difference aside, 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

**TABLE 2** *Standardized estimates of the log-odds ratios for mortality for the **established** covariates in ECMO example*

| Variable | EN | Auto-Zero | MSN | IPF-Lasso | SGL | GLASSO | Bayes |
|---|---|---|---|---|---|---|---|
| Admitted hours pre-ECMO (log) | 0.024 | 0.071 | 0.025 | 0.000 | 0.000 | 0.000 | 0.020 |
| Intubated hours pre-ECMO (log) | 0.326 | 0.651 | 0.333 | 0.344 | 0.290 | 0.422 | 0.811 |
| pH | -0.028 | -0.325 | -0.035 | -0.016 | -0.004 | -0.018 | -0.151 |
| $PaCO_2$ | 0.049 | -0.011 | 0.049 | 0.007 | 0.005 | 0.012 | 0.020 |
| MAP (CMV), cm $H_2O$ | 0.006 | 0.162 | 0.006 | 0.000 | 0.000 | 0.000 | 0.122 |
| MAP (HFOV), cm $H_2O$ | 0.030 | 0.195 | 0.034 | 0.007 | 0.001 | 0.003 | 0.140 |
| Malignancy | 0.007 | 0.006 | 0.008 | 0.000 | 0.000 | 0.000 | 0.030 |
| Asthma diagnosis | -0.114 | -2.017 | -0.123 | -0.027 | -0.002 | -0.015 | -0.030 |
| Bronchiolitis diagnosis | -0.073 | -0.168 | -0.070 | -0.028 | -0.009 | -0.053 | -0.562 |
| Pertussis diagnosis | 0.220 | 0.392 | 0.225 | 0.217 | 0.188 | 0.252 | 0.399 |
| Milrinone | 0.004 | 0.007 | 0.004 | 0.000 | 0.000 | 0.000 | -0.010 |

**TABLE 3** *Standardized estimates of the log-odds ratios for mortality for the **unestablished** covariates in ECMO example*

| Variable | EN | Auto-Zero | MSN | IPF-Lasso | SGL | GLASSO | Bayes |
|---|---|---|---|---|---|---|---|
| Bilirubin, mg/dL (log) | 0.204 | 0.192 | 0.203 | 0.137 | 0.155 | 0.179 | 0.113 |
| ALT, U/L (log) | 0.515 | 0.452 | 0.508 | 0.640 | 0.625 | 0.702 | 1.53 |
| Leukocytosis (log) | 0.086 | 0.023 | 0.083 | 0.040 | 0.052 | 0.077 | 0.049 |
| Leukopenia (log) | -0.070 | -0.025 | -0.066 | -0.011 | -0.019 | -0.054 | -0.041 |
| Thrombocytopenia (log) | 0.024 | 0.017 | 0.024 | 0.000 | 0.000 | 0.000 | 0.010 |
| INR | 0.057 | 0.099 | 0.058 | 0.007 | 0.010 | 0.024 | 0.049 |
| VIS (log) | 0.019 | 0.009 | 0.018 | 0.000 | 0.000 | 0.000 | 0.020 |
| Lactate, mMol/L (log) | 0.308 | 0.308 | 0.305 | 0.313 | 0.299 | 0.403 | 0.678 |
| PF-ratio (log) | -0.135 | -0.173 | -0.135 | -0.044 | -0.050 | -0.127 | -0.128 |
| Abnormal pupillary response | -0.011 | -0.007 | -0.009 | 0.000 | 0.000 | 0.000 | -0.010 |
| pre-ECMO kidney injury | 0.014 | 0.008 | 0.014 | 0.000 | 0.000 | 0.000 | 0.000 |

**TABLE 4** *Model Performance on ECMO Dataset*

| Method | AUC | Brier |
|---|---|---|
| EN | 0.828 | 0.143 |
| Auto-Zero | 0.837 | 0.141 |
| MSN | 0.828 | 0.143 |
| IPF-Lasso | 0.815 | 0.147 |
| SGL | 0.812 | 0.148 |
| GLASSO | 0.822 | 0.144 |

that bilirubin, ALT, lactate, and PF-ratio were important predictors of mortality. However, the MSN, IPF-EN, EN, and Auto-Zero 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 (36); the estimated coefficients from that analysis (using the sensible adaptive Bayes with optimistic prior approach) are presented in Tables 2 and 3 in the "Bayes" column for comparison.

## 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. We found that the sensible ad hoc approach of placing no penalty at all on the established predictors (Auto-Zero) does not perform well in this setting, and may result in inflated estimates of the established predictors and underestimates of the unestablished predictors. We believe that this is caused by the total lack of shrinkage on the established predictors–even when they truly are nonzero, the large number of covariates relative to sample size makes some kind of shrinkage desirable. When considering more formal approaches, we found that 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. In the same way that Zou and Hastie (2005) and others have found the standard elastic net preferable to the standard lasso, the smoother shrinkage and potential for selecting groups of correlated predictors may correspondingly make the MSN preferable to the IPF-Lasso in this context. In initial exploration of the MSN, we used the full [0, 1] $\alpha$ sequence in 0.1 increments. However, after our preliminary numerical studies found that it rarely selected $\alpha > 0.2$, we restricted 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 option of zero penalization on the established predictors did not yield substantial differences. As discussed, it was never selected in our simulations or in our application, and the Auto-Zero's forced zero penalization on the established predictors often resulted in poor estimates. After further investigation, we found that the established covariates' association with the outcome had to be very large to make zero penalization a viable choice for the cross-validation procedure, and in the scenarios considered in our simulation study, the established covariates were not large enough relative to the unestablished covariates to make zero penalization attractive. Crucially, however, performance of the MSN did not seem to suffer due to this additional consideration. We are less attached to the use of only two groupings. Limiting the number of groups to two is the simplest option, and having a more clear-cut decision may make our method more

appealing to researchers. In addition, using only two groupings means that our method can rely on just one tuning parameter, $\phi$, 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 $\lambda$ 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 equivalent.) With more than two groups, though, multiple $\phi$ 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.

A fully Bayesian approach that incorporates historical information directly via prior distributions on the established predictors' coefficients would need to account for potential model misspecification across nested models and differences in coefficient values, as described by Robinson and Jewell (37). Boonstra and Barbaro consider one such Bayesian approach that approximates and accounts for this model misspecification (36). In contrast, the MSN method proposed in this paper would be suitable when prior research suggests that a subset of predictors are likely to be associated with the outcome, but there is considerable uncertainty about the true magnitude of these associations. Such a scenario may occur, for example, when the previous models were fit to a different target population. The MSN provides a less assumptive way to automatically take this prior knowledge into account. In this sense, it represents a middle ground - and a third option - between completely ignoring any prior research versus formally incorporating the previous estimates of association and their uncertainty.

One limitation of our approach, though, is that our method may be ill-equipped for the scenario in which the established covariates are mediators for unestablished covariates. We saw evidence of this in the ECMO application: a key difference between our findings and those of Boonstra and Barbaro, who previously reported on these data, was that we found $PaCO_2$ (an established predictor) to be positively associated with mortality, while Boonstra and Barbaro found this association to be nearly zero (36). $PaCO_2$ is correlated with lactate (an unestablished predictor), both measuring the degree of acidosis, and likely to be a mediator for the association between lactate and mortality. If so, the estimated association between $PaCO_2$ and mortality is probably zero. In general, when the marginal associations between the established covariates and the outcome are much different, or even in the opposite direction, from the same associations after conditioning on the unestablished covariates, the MSN will not be expected to perform as well.

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 and estimation accuracy.

**Acknowledgments** This work was supported by the National Institutes of Health (P30 CA046592; R01 CA129102; T32 CA083654) and the National Science Foundation (DGE-1256260).

**Data Availability** The code used to create the simulation, run all methods, and process results is accessible on GitHub (https://github.com/psboonstra/MSEN).

## 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;37:3468-3497.

[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. *Proceedings of the 26th International Conference on Machine Learning.* Montreal;2009:433-440.

[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;41:1111-1141.

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

[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? https://www.cdc.gov/cancer/lung/basic_info/risk_factors.htm, visited on April 20, 2018; 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;98:270-276.

[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;5:834-846.

[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;51:1422-1436.

[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;37:1515-1530.

[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;9:1488-1496.

[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. *Statistics in Medicine.* 2014;33:5310-5329.

[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;Article ID 7691937.

[28] ACS . Bladder Cancer Risk Factors. https://www.cancer.org/cancer/bladder-cancer/causes-risks-prevention/risk-factors.html, visited on April 20, 2018; 2017.

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

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

[31] Team R Core. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria; 2018.

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

[33] 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.

[34] Barbaro R, Boonstra P, Kuo K, et al. Evaluating mortality risk adjustment among children receiving extracorporeal support for respiratory failure. *ASAIO Journal.* 2019;65:277-284.

[35] Buuren S, Groothuis-Oudsboorn K. MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software.* 2011;45:1-67.

[36] Boonstra PS, Barbaro RP. Incorporating historical models with adaptive Bayesian updates. *Biostatistics.* 2018;kxy053.

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

**FIGURE 1** *AUC, Brier score,and rMSE of Auto-Zero, GLASSO, IPF-Lasso, MSN, and SGL, log-scaled relative to the elastic net, n = 200. For AUC, larger is better; for Brier score and rMSE, smaller is better.*

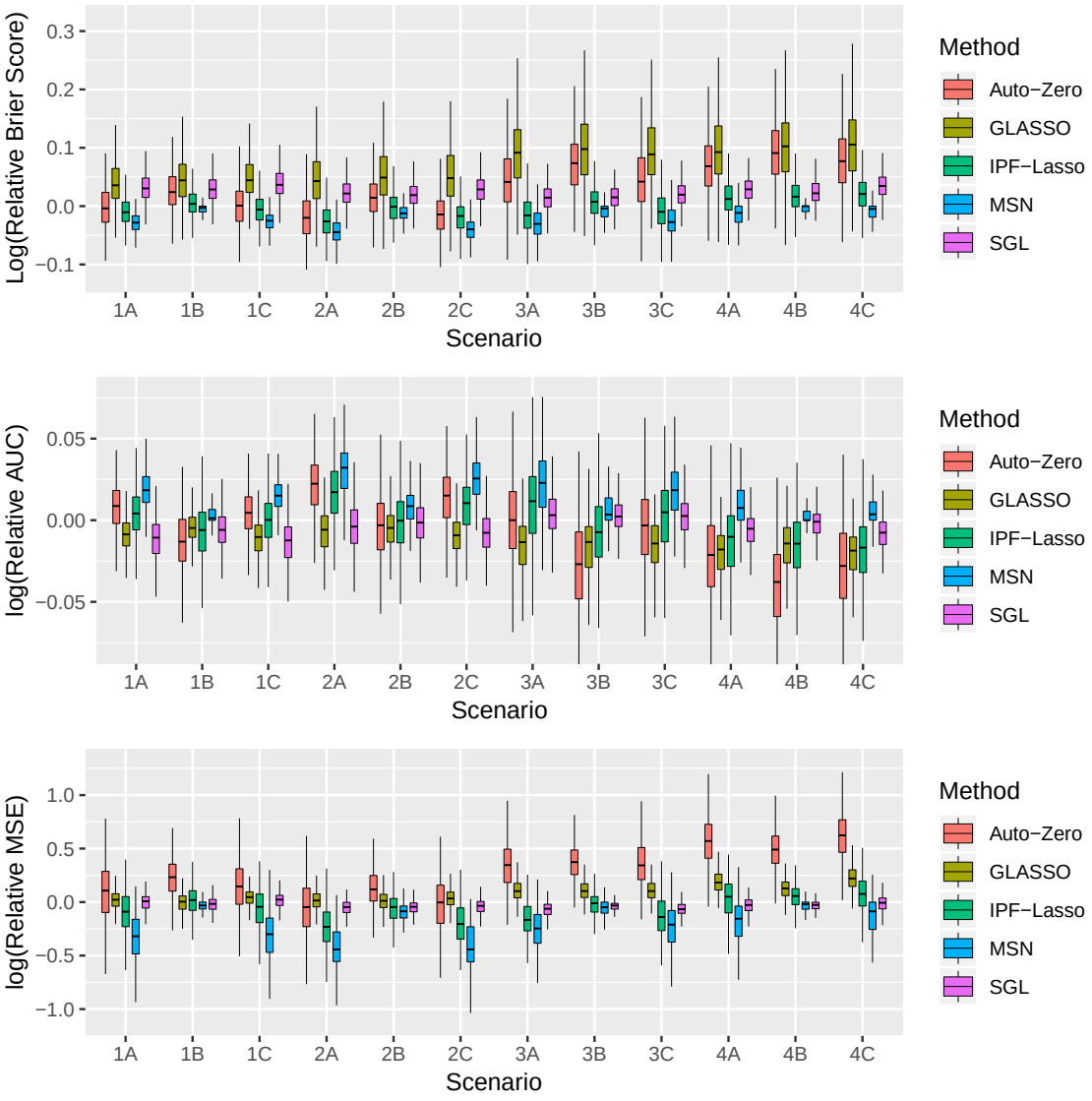

**FIGURE 2** *AUC, Brier score, and rMSE of Auto-Zero, GLASSO, IPF-Lasso, MSN, and SGL, log-scaled relative to the elastic net, n = 1000. For AUC, larger is better; for Brier score and rMSE, smaller is better.*