

Supplementary Materials for “Penalized regression with negative-unlabeled data: An approach to defining long COVID”

Harrison T. Reeder^{1,2}, Tanayott Thaweethai^{1,2} and Andrea S. Foulkes^{1,2,3}

¹Biostatistics, Massachusetts General Hospital, Boston, MA USA

²Department of Medicine, Harvard Medical School, Boston, MA USA

³Department of Biostatistics, Harvard TH Chan School of Public Health,
Boston, MA USA

December 2, 2025

Abstract

In these supplementary materials we present additional details and results beyond what could be presented in the main manuscript. To distinguish the two documents, alpha-numeric labels are used in this document while numeric labels are used in the main paper. Appendix A provides additional simulation results. Appendix B provides additional results from the data application.

A Additional Simulation Results

To more systematically evaluate the impact of: potential balancing weighting on performance in the presence of confounding, simulation studies are conducted to evaluate discrimination with respect to Y of the LCRI given by Equation (3).

A.1 Balancing weights to address potential confounding

A.1.1 Data Generating Mechanism

For clarity, our simulation studies examine a single binary covariate Z , with some prevalence $\Pr(Z = 1)$. We define the proportion infected within each covariate stratum, i.e., $\alpha_z = \Pr(A = 1 \mid Z = z)$. These probabilities can equivalently be specified by a marginal infected proportion of $\alpha = \alpha_0 \Pr(Z = 0) + \alpha_1 \Pr(Z = 1)$, and a risk ratio corresponding to the association between Z and A is given by $\beta_{Z/A} = \alpha_1/\alpha_0$.

Similarly, for each feature k we still consider a risk ratio β_{1k} corresponding to Long COVID status, along with another corresponding to its association with covariate Z , yielding the multiplicative specification

$$p_{k,y,z} = \Pr(X_k = 1 \mid Y = y, Z = z) = \beta_{0k} (\beta_{1k})^y (\beta_{Z,k})^z,$$

Together, the two risk ratios $\beta_{Z/A}$ and $\beta_{Z,k}$ quantify the strength of Z 's associations with infection and with each symptom, and thus jointly are expected to determine the level of bias in an analysis of the association between A and X_k that does not account for Z .

A.1.2 Simulation Settings

Data were generated with the same parameters as the above simulations with $\alpha = 0.7$ corresponding to a population with 70% having had SARS-CoV-2 infection to allow for differences in infection prevalence by confounder level. Marginally $\Pr(Z = 1) = 0.55$, and the marginal risk ratio for A by Z was allowed to be either 1.7, or 0.65, to capture negative or positive Z/A associations. Twelve symptoms had a non-zero association with Z with marginal risk ratios varying between 2 and 2.6, and the set of symptoms associated with Z was set to either the 12 symptoms with a true Long COVID association ('Z/X Overlapping'), or 12 other symptoms without true Long COVID association ('Z/X Non-Overlapping'). Finally, we considered a setting in which no symptoms were associated with Z marginally. Finally, the associations between Long COVID and symptoms were fixed as in the 'Medium Signal' setting above, yielding a total of 6 simulation settings. The metrics and comparators used were the same as previously.

A.1.3 Simulation Results

Table A.1 reports the selection performance of lasso-penalized logistic regression models with and without balancing weights, across simulation settings. In settings with no Z/X association, there is no mechanism for confounding, and the results are comparable, with the unadjusted approach actually performing slightly better due to increased efficiency. However, in the presence

of both Z/X and Z/A associations, the model fit with balancing weights substantially outperforms the unadjusted model, due to the impact of confounding. When Z was associated with non-Long COVID symptoms, the unadjusted model included many such symptoms, leading to larger model size and lower true negative rate. When Z was associated with the same Long COVID-associated symptoms, but the associations were in opposing directions, the unadjusted model tended to miss truly non-zero symptoms due to confounding, leading to decreased true positive rate and poor rank-correlation between estimated and true coefficients.

Figure A.1 compares the discrimination statistics between the models in these settings, revealing similar patterns. In the absence of a Z/A association, and therefore no impact of confounding, performance was comparable, with slight differences due to decreased efficiency of the weighted approach. When confounding affected a non-overlapping set of symptoms, the unadjusted model included more such symptoms associated with infection due to confounding but not truly associated with Long COVID, leading to degraded performance. The balancing weighted approach better recovered the true Long COVID-associated symptoms, and therefore the Long COVID score achieved better discrimination. Finally, when the true Long COVID-associated symptoms were also subject to confounding, the impact of weighting depended on the direction of confounding bias. In the artificial setting of ‘Z/X Overlapping, Positive Z/A Association’ in which all Long COVID-associated symptoms also had confounding bias in the same direction, the result was that the unadjusted approach performed better, because the true signal was amplified by confounding. More realistically, when confounding bias counteracted the direction of Long COVID association as in ‘Z/X Overlapping, Negative Z/A Association’, then the unadjusted approach was unable to select relevant symptoms, leading to degraded performance.

Table A.1: Simulation results for Lasso penalized feature selection performance under potential confounding

	Selected		TPR		TNR		Est. vs. True	
	Median (IQR)		Mean (SD)		Mean (SD)		Coef.	Rank-Correlation*
	Bal. Weights	Unadj.	Bal. Weights	Unadj.	Bal. Weights	Unadj.	Bal. Weights	Unadj.
No Z/X Assoc.								
Neg. Z/A Assoc.	8 (4-13)	8 (5-12)	0.541 (0.25)	0.598 (0.244)	0.894 (0.156)	0.925 (0.135)	0.588 (0.123)	0.663 (0.132)
Pos. Z/A Assoc.	8 (3.75-13)	8 (5-12)	0.526 (0.259)	0.598 (0.244)	0.89 (0.17)	0.925 (0.135)	0.576 (0.125)	0.663 (0.132)
Pos. Z/X Assoc. Non-Overlapping								
Neg. Z/A Assoc.	8 (4-12)	19 (16-22)	0.532 (0.25)	0.637 (0.197)	0.902 (0.153)	0.597 (0.101)	0.589 (0.126)	0.614 (0.066)
Pos. Z/A Assoc.	7 (3-12)	19 (16-22)	0.51 (0.255)	0.613 (0.195)	0.903 (0.148)	0.585 (0.083)	0.571 (0.128)	0.126 (0.1)
Pos. Z/X Assoc. Overlapping								
Neg. Z/A Assoc.	6 (3-10)	5 (3-8)	0.469 (0.242)	0.382 (0.194)	0.938 (0.13)	0.946 (0.109)	0.581 (0.143)	-0.235 (0.113)
Pos. Z/A Assoc.	6 (3-11)	11 (11-12)	0.486 (0.261)	0.929 (0.077)	0.925 (0.151)	0.995 (0.028)	0.579 (0.15)	0.838 (0.045)

Abbreviations: TNR, True Negative Rate; TPR, True Positive Rate.

*Kendall’s τ between estimated regression coefficients and true Symptom-Long COVID risk ratios.

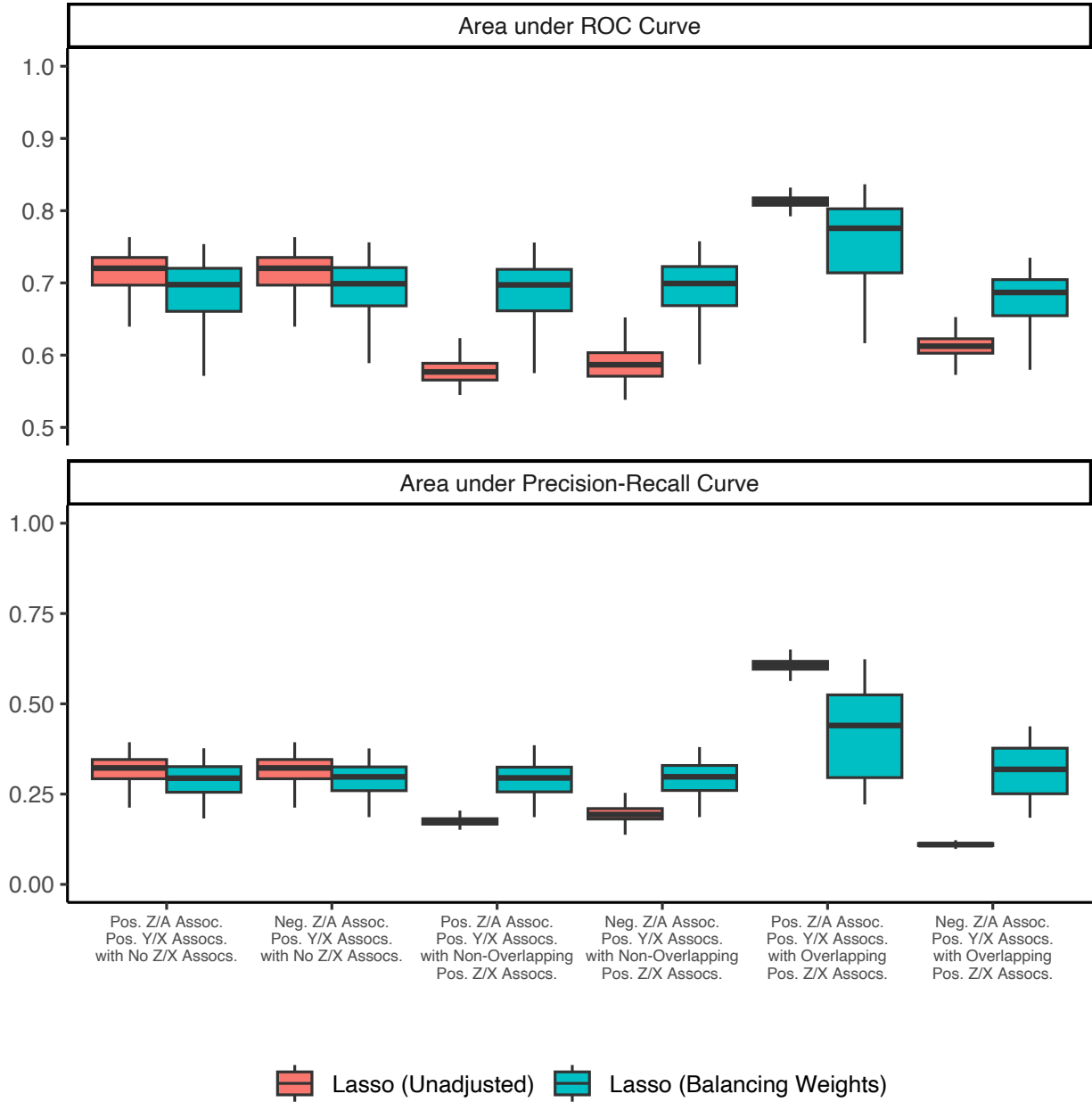


Figure A.1: Discriminative performance of estimated Long COVID indices in distinguishing true Long COVID status under potential confounding. Boxplot whiskers extend to the furthest value within $1.5 \times$ the interquartile range from the upper and lower quartiles. The null value of the area under the ROC curve is 0.5, and for the area under the precision-recall curve is the proportion of individuals truly with LC, which varies slightly across datasets but is expected as $\pi = 0.2$.

A.2 Choice of Penalty

In this section we evaluate the comparative performance of the proposed Lasso penalized logistic regression modeling approach to alternative specification via elastic net penalization (Friedman et al., 2010). The elastic net penalty generalizes ℓ_1 Lasso penalization to incorporate an additional ℓ_2 (‘ridge’) penalty, with a second regularization parameter $0 \leq \alpha \leq 1$ corresponding to the relative weight of each penalty, as follows:

$$p(\gamma) = \lambda \left[\alpha \sum_{j=1}^p |\gamma_j| + (1 - \alpha) \sum_{j=1}^p \gamma_j^2 / 2 \right]$$

The λ regularization parameter can again be selected via cross-validated misclassification error using the ‘one standard error’ rule described in Friedman et al. (2010) and referenced in the main text, while the value of α can either be pre-specified or additionally selected via cross validation. We considered two forms of penalty, one with fixed $\alpha = 0.5$ and one with α selected as the value that minimizes the cross-validated misclassification under the ‘one standard error’ rule from amongst $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$. Note that the latter includes the lasso penalty $\alpha = 1$ and therefore will always have cross-validated misclassification error at least as good as the Lasso specification.

Simulations were performed under identical settings described in Section 2.4 of the main text, evaluating feature selection performance and discriminative performance of the research indices as presented in Figure A.2 and Table A.3. Selection performance results indicated that across the simulation settings, cross-validation of α in the elastic net tended to select larger models, reflecting the ability to choose values that put relatively less weight on the sparsity-inducing ℓ_1 penalty and more weight on the non-sparse ℓ_2 penalty. However, this translated to relatively small increases in discriminative performance in the LCRI, where performance was highly similar across the comparator penalties. This finding may reflect the large sample size of the simulations (as in the motivating RECOVER example), where there is sufficient data richness that the role of the shrinkage provided by the elastic net’s ℓ_2 penalty does not strongly alter the resulting estimates selected by the Lasso penalization. In settings with smaller n or larger p there may be a greater distinction between the Lasso and elastic net penalties, reflecting the known properties of the elastic net in the presence of correlated sets of covariates to select entire groups, rather than a single variable from each group (Zou and Hastie, 2005).

Table A.2: Simulation results for Elastic Net feature selection performance

Setting	# Features Selected	TPR	TNR	Est. vs. True Coef.
	Median (IQR)	Mean (SD)	Mean (SD)	Rank-Correlation*
Elastic Net ($\alpha = 0.5$)				
Low Signal				
Uncorrelated	7 (3-14)	0.475 (0.257)	0.865 (0.183)	0.498 (0.121)
Non-Group Sparse	13 (6-20)	0.451 (0.233)	0.701 (0.228)	0.139 (0.123)
Group Sparse	5 (3-9)	0.341 (0.203)	0.885 (0.187)	0.386 (0.125)
Medium Signal				
Uncorrelated	8 (3-13)	0.529 (0.262)	0.884 (0.179)	0.569 (0.12)
Non-Group Sparse	14 (7-21)	0.484 (0.224)	0.688 (0.23)	0.142 (0.124)
Group Sparse	4 (2-8)	0.349 (0.200)	0.920 (0.155)	0.434 (0.119)
High Signal				
Uncorrelated	8 (4-12)	0.565 (0.251)	0.92 (0.146)	0.631 (0.124)
Non-Group Sparse	14 (8-21)	0.517 (0.211)	0.688 (0.223)	0.143 (0.127)
Group Sparse	4 (3-7)	0.356 (0.195)	0.938 (0.142)	0.465 (0.114)
Elastic Net (α Cross-Validated)				
Low Signal				
Uncorrelated	8 (3-17)	0.524 (0.294)	0.774 (0.327)	0.498 (0.115)
Non-Group Sparse	16 (7-28)	0.553 (0.304)	0.567 (0.346)	0.131 (0.119)
Group Sparse	6 (3-20)	0.46 (0.32)	0.722 (0.389)	0.396 (0.124)
Medium Signal				
Uncorrelated	9 (4-16)	0.582 (0.285)	0.792 (0.321)	0.57 (0.113)
Non-Group Sparse	18 (10-31)	0.606 (0.284)	0.528 (0.343)	0.125 (0.117)
Group Sparse	5 (3-15)	0.453 (0.308)	0.766 (0.373)	0.437 (0.118)
High Signal				
Uncorrelated	9 (5-14)	0.621 (0.269)	0.818 (0.319)	0.633 (0.118)
Non-Group Sparse	18 (10-29.25)	0.622 (0.262)	0.539 (0.333)	0.126 (0.117)
Group Sparse	5 (3-16)	0.465 (0.307)	0.766 (0.384)	0.462 (0.113)

Abbreviations: TPR, True Positive Rate; TNR, True Negative Rate.

*Kendall's τ between estimated regression coefficients and true risk ratios of Long COVID vs each symptom.

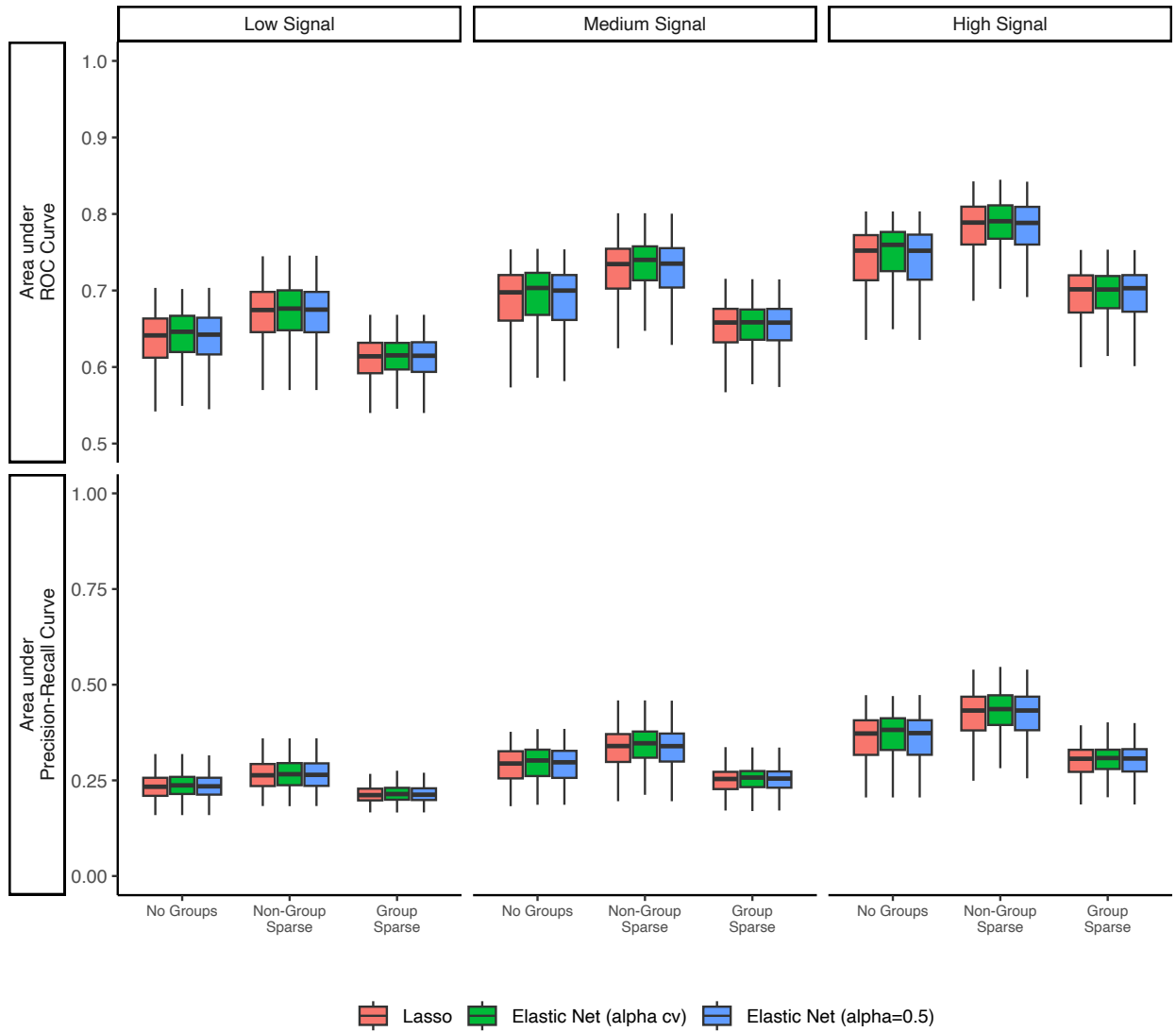


Figure A.2: Discriminative performance of estimated Long COVID indices under three choices of penalty: Lasso penalization, elastic net penalty with $\alpha = 0.5$, and elastic net penalty with α selected via cross-validated misclassification error from amongst $\{0, 0.25, 0.5, 0.75, 1\}$. Boxplot whiskers extend to the furthest value within $1.5 \times$ the interquartile range from the upper and lower quartiles. The null value of the area under the ROC curve is 0.5, and for the area under the precision-recall curve is the proportion of individuals truly with LC, which varies slightly across datasets but is expected as $\pi = 0.2$.

A.3 Misclassification of Infection Status

In this section we evaluate the performance of the proposed approach in the presence of misclassification of infection status. Misclassification may occur either if someone believed to be uninfected actually has a history of infection that was not identified (e.g., if the prior SARS-CoV-2 infection was asymptomatic), or if someone believed to have a history of infection actually is uninfected (e.g., if a supposed prior SARS-CoV-2 infection was actually another viral infection).

Simulations were performed under identical settings described in Section 2.4 of the main text, restricting the true associations between Long COVID and symptoms to the ‘Medium Signal’ setting. However, after simulation of the true infection status, Long COVID status, and symptoms, an additional ‘observed’ infection status was simulated using specified non-differential misclassification rates of for both uninfected-to-infected misclassification and infected-to-uninfected misclassification. Model fits and symptom counts were then constructed using this ‘observed’ infection status subject to misclassification.

The results indicate that both the feature selection performance and discrimination of the research index mildly degrade in the presence of misclassified infection status, which is to be expected in that infection serves as a pseudo-label for the true Long COVID status. However, in these simulation settings the loss of performance is relatively minimal, which may be due to the large sample size, or the size of the true associations. Performance may be more sensitive to misclassification in smaller samples or with weaker signal, or as the level of misclassification continues to increase. In these results, we do observe that under the most extreme setting of high misclassification and no symptom correlation, LCRI discriminative performance approached that of the symptom count. Nevertheless, these results broadly indicate a level of robustness to some misclassification of infection status across a range of settings.

Table A.3: Simulation results for Lasso penalized feature selection performance under misclassification of infection status

	# Features Selected	TPR	TNR	Est. vs. True Coef. Rank-Correlation*
Setting	Median (IQR)	Mean (SD)	Mean (SD)	Mean (SD)
No Misclassification				
Uncorrelated	8 (3-13)	0.523 (0.264)	0.887 (0.179)	0.568 (0.122)
Non-Group Sparse	13 (7-20)	0.482 (0.222)	0.698 (0.226)	0.142 (0.124)
Group Sparse	4 (2-8)	0.339 (0.202)	0.922 (0.154)	0.426 (0.12)
5% Misclassification				
Uncorrelated	7 (3-14)	0.489 (0.266)	0.862 (0.191)	0.512 (0.122)
Non-Group Sparse	11 (4-20)	0.44 (0.238)	0.724 (0.231)	0.142 (0.132)
Group Sparse	4 (2-8)	0.319 (0.192)	0.915 (0.158)	0.396 (0.107)
10% Misclassification				
Uncorrelated	7 (3-14)	0.436 (0.256)	0.861 (0.187)	0.464 (0.123)
Non-Group Sparse	11 (4-19)	0.416 (0.244)	0.729 (0.232)	0.141 (0.134)
Group Sparse	4 (2-8)	0.306 (0.197)	0.896 (0.17)	0.361 (0.122)

Abbreviations: TPR, True Positive Rate; TNR, True Negative Rate.

*Kendall’s τ between estimated regression coefficients and true risk ratios of Long COVID vs each symptom.

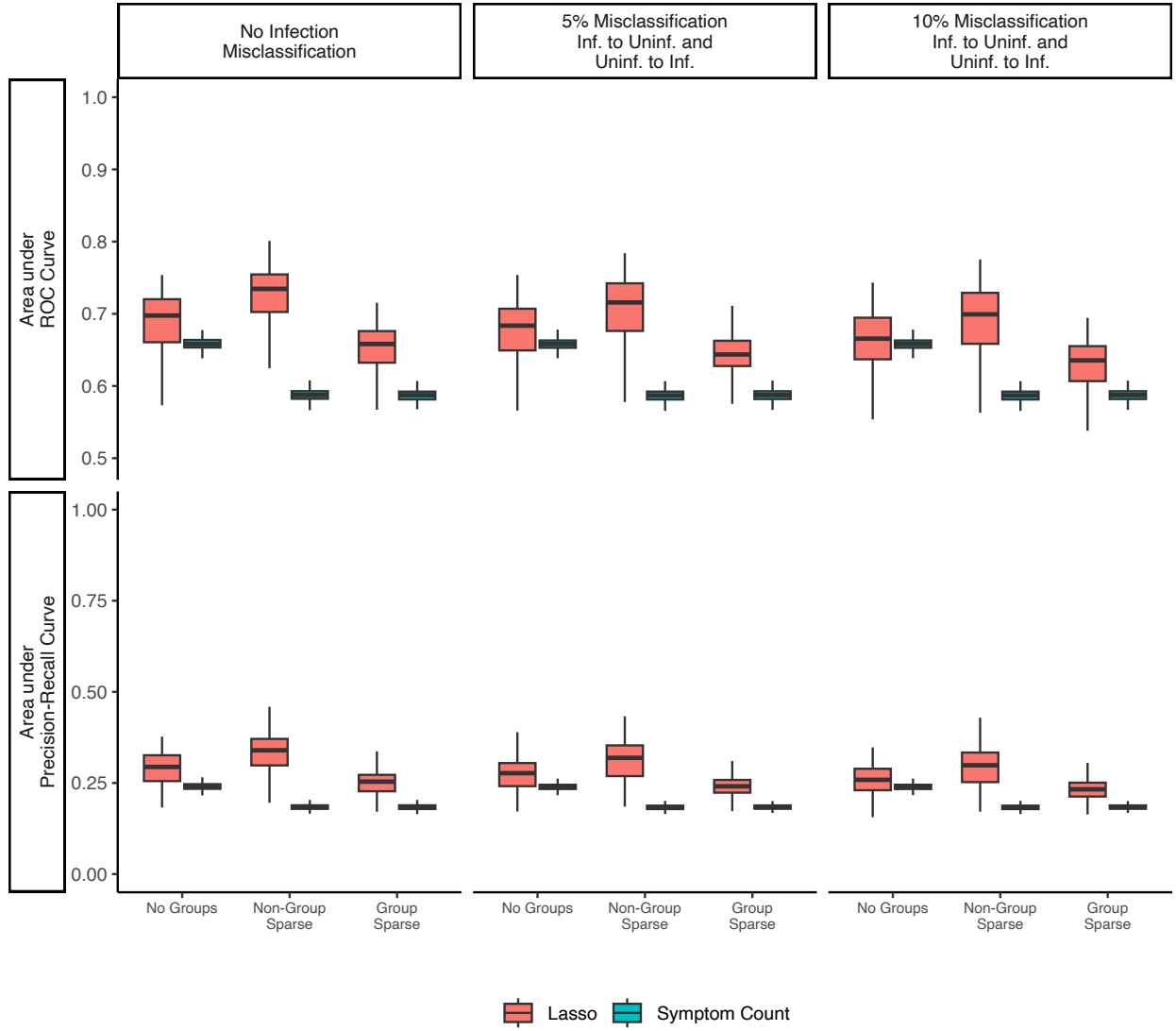


Figure A.3: Discriminative performance of estimated Long COVID indices under three levels of infection misclassification. Boxplot whiskers extend to the furthest value within $1.5 \times$ the interquartile range from the upper and lower quartiles. The null value of the area under the ROC curve is 0.5, and for the area under the precision-recall curve is the proportion of individuals truly with LC, which varies slightly across datasets but is expected as $\pi = 0.2$.

B Additional Data Application Results

Table B.1: Analysis cohort demographic characteristics by infection status.

Category	Overall ($n = 9702$)	Infected ($n = 8593$)	Uninfected ($n = 1109$)
<i>Sex assigned at birth</i>			
Male	2773 (28.6%)	2374 (27.6%)	399 (36%)
Female	6929 (71.4%)	6219 (72.4%)	710 (64%)
<i>Age category</i>			
18-45	4745 (48.9%)	4369 (50.8%)	376 (33.9%)
46-65	3656 (37.7%)	3158 (36.8%)	498 (44.9%)
>65	1301 (13.4%)	1066 (12.4%)	235 (21.2%)
<i>Race/Ethnicity</i>			
Non-Hispanic White	5705 (58.8%)	5021 (58.4%)	684 (61.7%)
Non-Hispanic Black	1416 (14.6%)	1219 (14.2%)	197 (17.8%)
Non-Hispanic Asian	501 (5.2%)	428 (5%)	73 (6.6%)
Hispanic	1591 (16.4%)	1472 (17.1%)	119 (10.7%)
Mixed race/Other/Missing	489 (5%)	453 (5.3%)	36 (3.2%)

Table B.2: Symptom relative frequencies by infection status.

Group	Symptom	Infected, %	Uninfected, %	Uninfected (Balancing Weights), %
General	Fatigue	38.4	16.7	17.9
	Fever/sweats/chills	12	3.7	3.9
	P-E malaise	27.7	6.9	7.2
	Swelling of legs	11	5.9	5.1
Cardiac	Chest pain	7.8	1.1	1
	Palpitations	21	6.7	7.1
Dermatologic	Hair loss	18.9	9.9	10
	Skin color changes	7	2.4	2.9
	Skin pain	2.8	1	1.1
	Skin rash	8.4	4.8	5
Eye	Vision	10	3.2	3
Ear	Hearing	21.3	13.7	11.1
Gastrointestinal	Abdominal pain	5.2	2	2.8
	Dry mouth	14.5	6	5.1
	GI	25.3	10	11.1
	Teeth	12.1	6.9	5.8
Metabolic	Thirst	13.9	4	4
Musculoskeletal	Back pain	15.2	9.2	8.9
	Foot pain	7.6	4.4	4.3
	Joint pain	16.6	9.6	8.5
	Muscle pain	14	5.8	5.9
	Weakness	13.4	5	4.5
Neurologic	Abnormal movements	4.4	0.7	0.6
	Brain fog	20.4	4.4	5.4
	Dizziness	22.9	7.1	8.1
	Headache	13.1	3.4	3.7
	Smell/taste	12.6	1.9	2
	Tremor	6.7	3	2.6
Psychiatric	Anxiety	11.5	6	6.9
	Depression	11.3	5.1	4.9
	Sexual desire/capacity	18.8	8.1	8.1
	Sleep disturbance	11.8	3.8	3.7
Respiratory	Chronic cough	11.7	3.2	3
	Shortness of breath	11.2	2.3	2.5
	Sleep apnea	17.5	10	9.5
	Throat pain	3	0.5	0.6
Urinary	Bladder	13	7.3	7.1

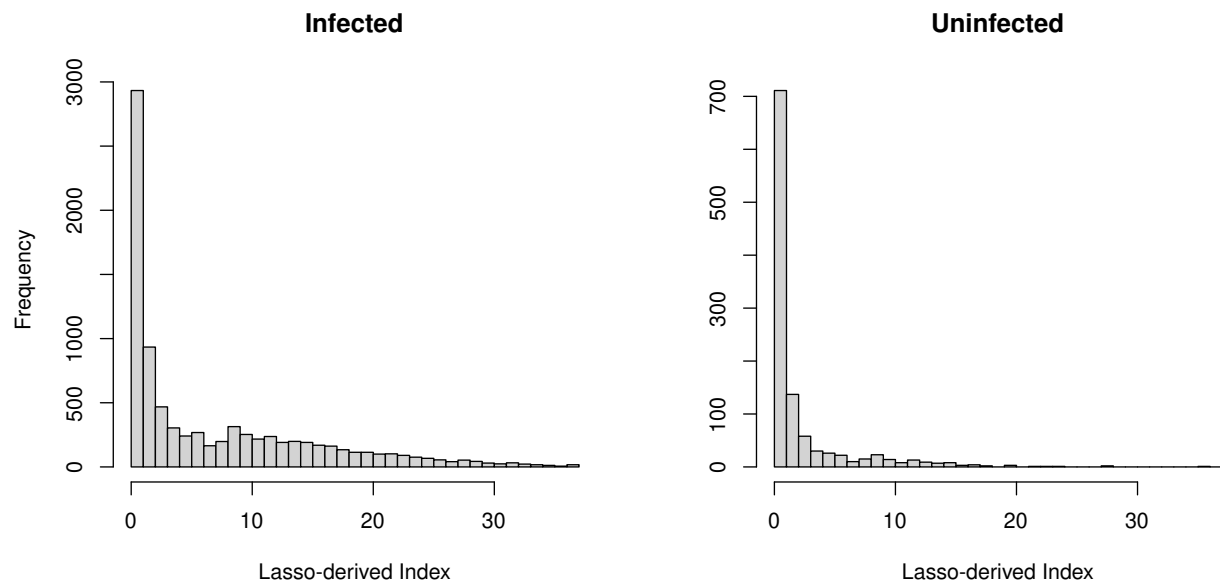


Figure B.1: Histogram of LCRI derived from lasso-penalized logistic regression with balancing weights, stratified by infection status.

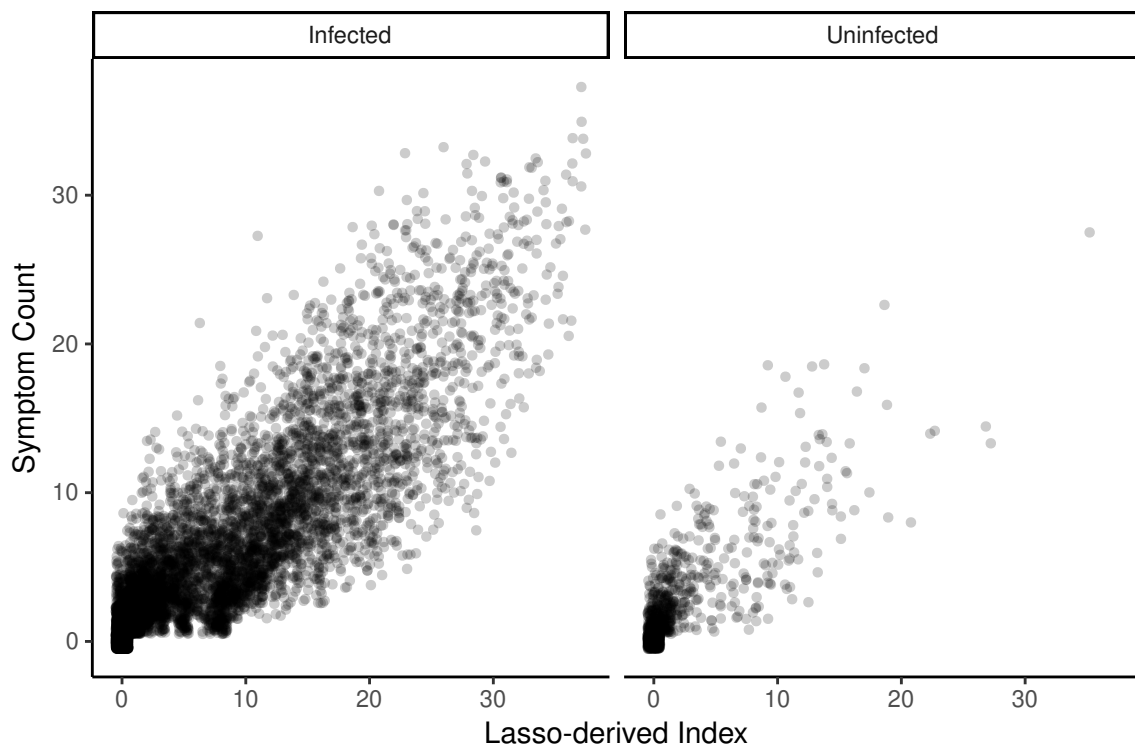


Figure B.2: Jittered scatterplot of LCRI derived from lasso-penalized logistic regression with balancing weights versus symptom count, stratified by infection status.

References

- Friedman, J. H., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320.