Risk Prediction for Heterogeneous Populations with Application to Hospital Admission Prediction

Jared Huling

Joint work with
Menggang Yu, Muxuan Liang, and Maureen Smith

Department of Statistics

University of Wisconsin–Madison
www.stat.wisc.edu/~huling

The Big Picture

Challenges in health system risk modeling

- Heterogeneity of the study population
  - Chronic conditions define heterogeneity

    $\Longrightarrow$

    underlying hierarchical structure

- 1000s of variables
  - Makes estimation a challenge when combined with heterogeneity

The Big Picture

Our solution

- Flexible model for heterogeneity
- Borrow strength across subpopulations using hierarchy constraints on variable importance
- Robustness to our key assumptions
- Code available publicly:

`github.com/jaredhuling/vennLasso`

Huling, J.D., Yu, M., Liang, M., Smith M. (2018), "Risk prediction for heterogeneous populations with application to hospital admission prediction," to appear in *Biometrics*.

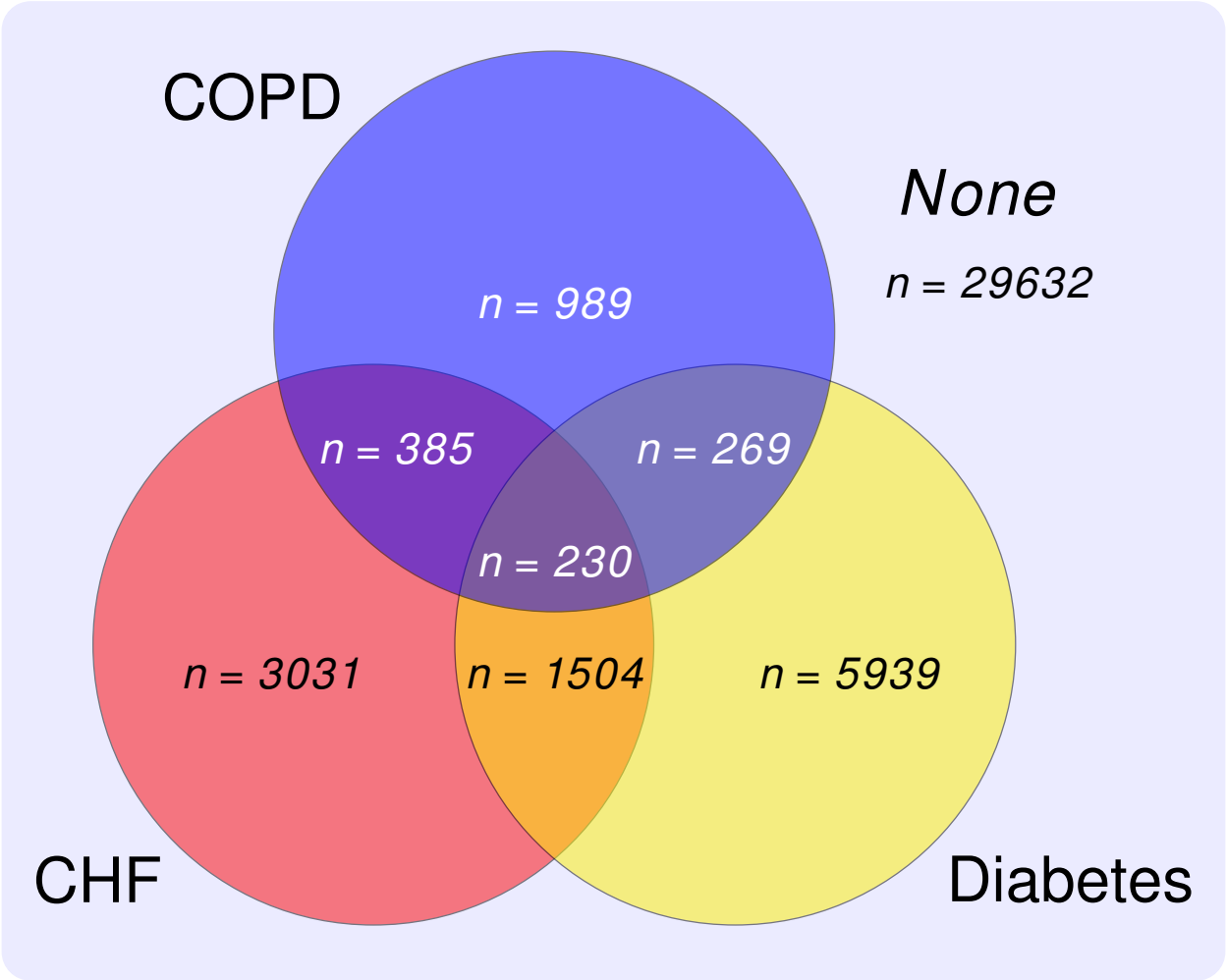
Risk Modeling for Health Systems

- Inpatient services constitute 29% of total health care spending in the US in 2009 (Pfuntner et al., 2013)
- Annual hospital cost of patients with **any** readmission is twice as high (Friedman et al., 2008)
- Many hospitalizations and readmissions are preventable (Minott, 2008)
- Focused care can improve outcomes/readmission and hospitalization rates

UWHealth Hospital Admissions Data

- Covariates from Medicare claims and EHR including
  - Health care payment information
  - Clinic and hospital visit
  - Pharmacy
  - Lab values such as A1c level
  - Demographic
- 12 months of baseline data collected and outcome of interest is hospitalization within 90 days

Profile of a Health System Population

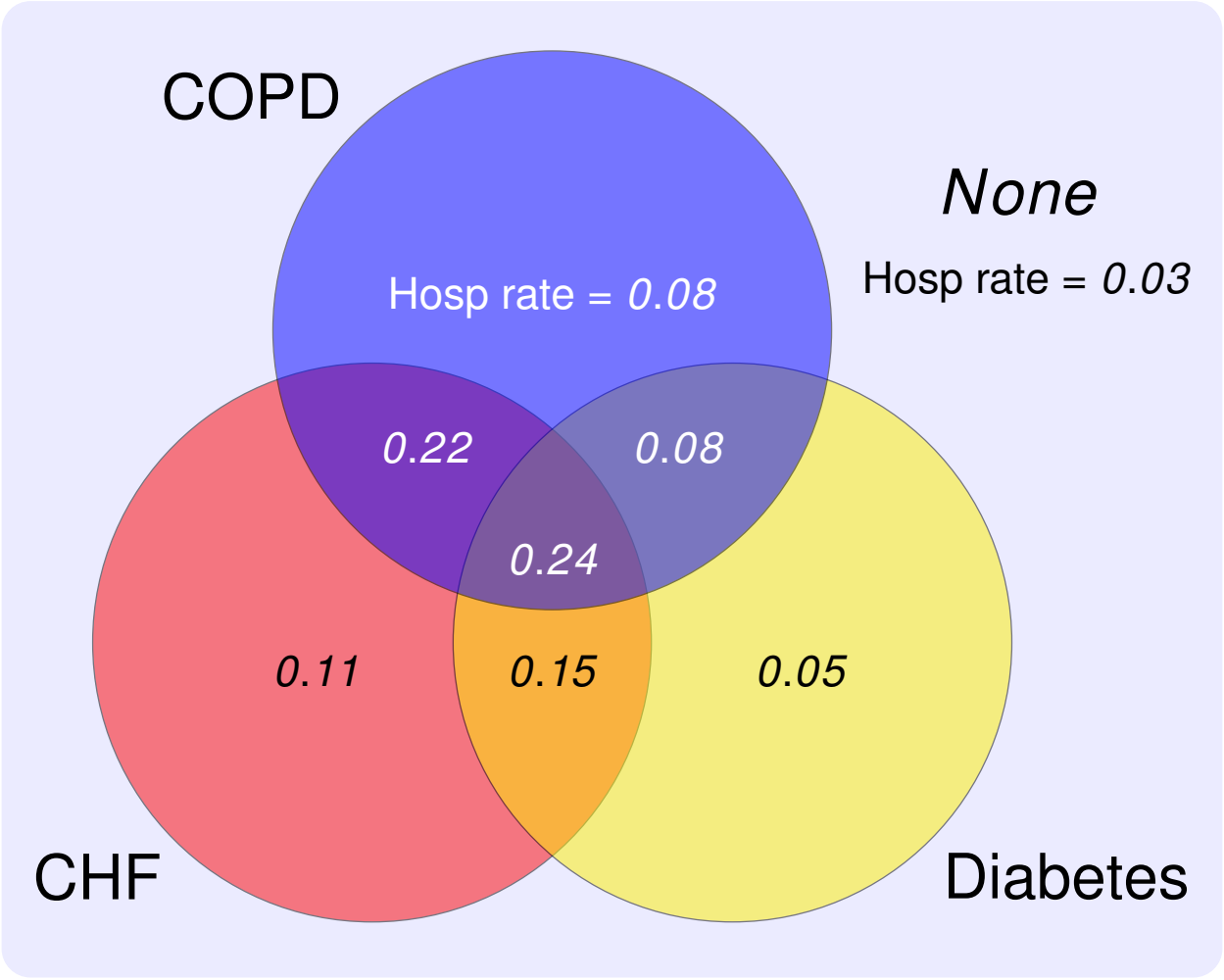


Sample sizes for each subpopulation in the UW Health cohort

CHF = congestive heart failure

COPD = chronic obstructive pulmonary disorder

Chronic Conditions and Risk of Hospitalization



90-day hospitalization rates for various subpopulations

Biological Plausibility of Heterogeneity

Among patients with diabetes:

Usage of medications for gastrointestinal issues may reflect severity of disease

$$\Longrightarrow \uparrow \text{risk of hospitalization}$$

Among patients without diabetes:

Such usage may reflect that a patient actively seeks to resolve health issues

$$\Longrightarrow \downarrow \text{risk of hospitalization}$$

Our Modeling Framework

Model: $\text{logit}(E[Y_{ik}|\mathbf{X}_{ik}]) = \mathbf{X}_{ik}\boldsymbol{\beta}_{k,\bullet}, i = 1, \ldots, n_k,$

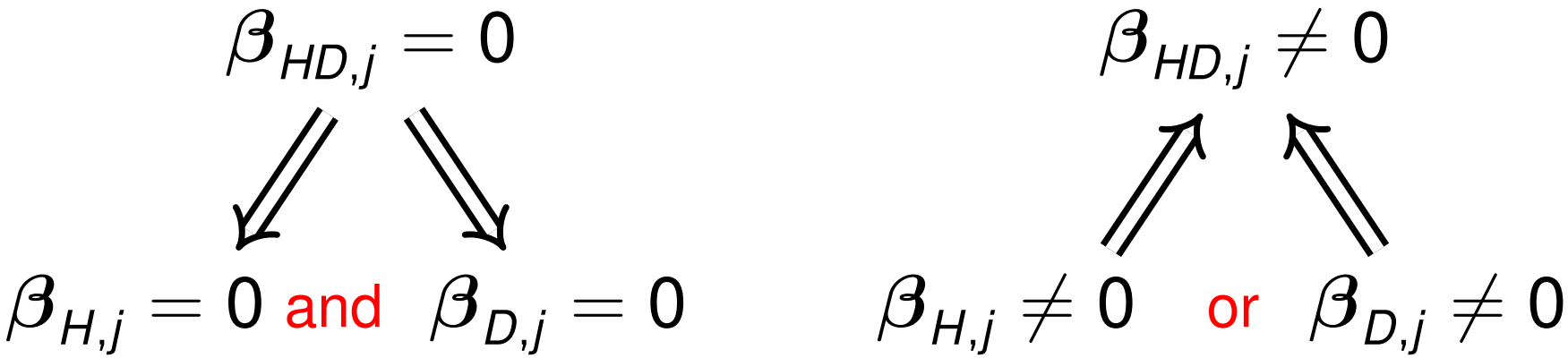
- $k \in \{H, P, D, HP, HD, PD, HPD, none\}$

$$
\begin{aligned}
H &= \text{Congestive } \textbf{Heart} \text{ Failure} \\
P &= \text{Chronic Obstructive } \textbf{Pulmonary} \text{ Disease} \\
D &= \textbf{Diabetes} \\
HP &= \text{C}\textbf{H}\text{F} + \text{CO}\textbf{P}\text{D} \\
&\ldots \\
none &= \text{None of } H, P, \text{ or } D
\end{aligned}
$$

- $\mathbf{X}_k$ is of dimension $n_k \times p$
- $\boldsymbol{\beta}_{k,\bullet} = (\beta_{k,1}, \ldots, \beta_{k,p})$

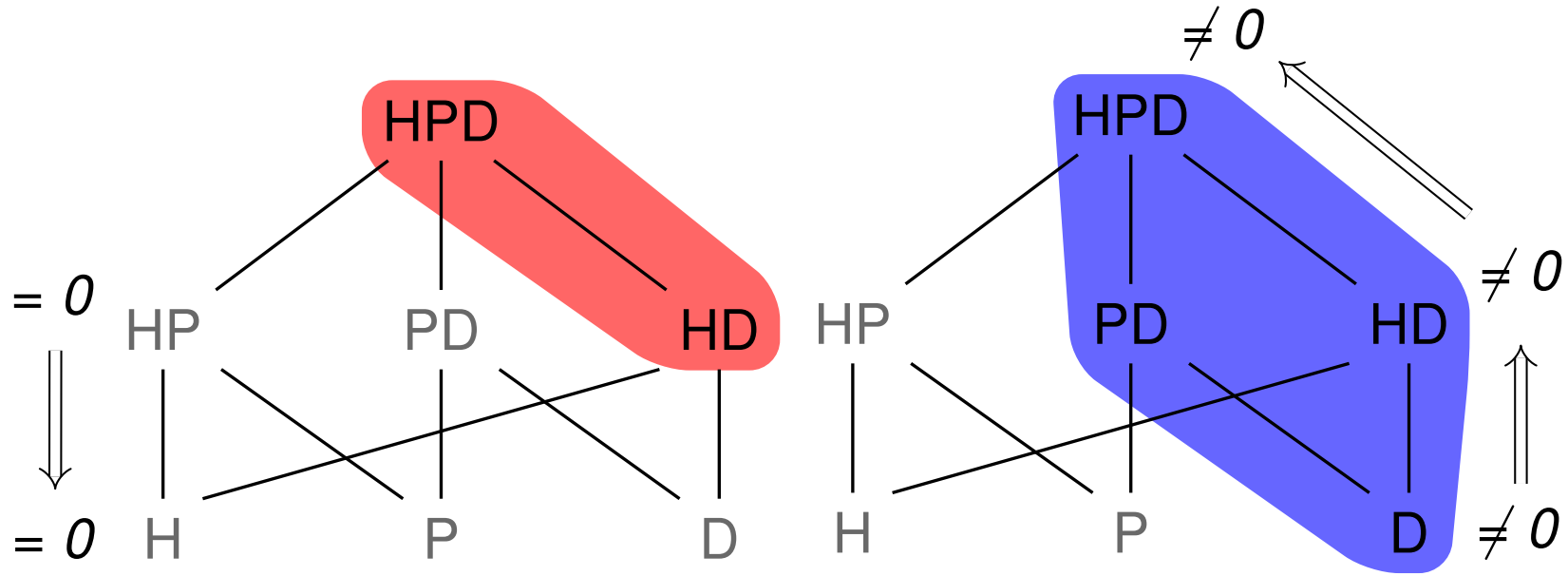
Borrowing Strength via Hierarchical Importance

For the j^{th} variable



Example: Pioglitazone and other similar diabetes medications may cause or worsen CHF (Tannen et al., 2013). Hence this medication information may only be predictive for diabetic patients *and* CHF

Hierarchical Selection Patterns



The two highlighted groups represent hierarchical selection patterns.

Proposed Method for Hierarchical Selection

We maximize the following penalized likelihood:

$$f(\boldsymbol{\beta}) = \sum_{k=1}^{K} \ell_k(\boldsymbol{\beta}_{k,\bullet}) - \lambda P(\boldsymbol{\beta})$$

where ℓ_k are log-likelihood functions (or negative loss) and P is an overlapping group lasso penalty with special structure to induce hierarchical selection patterns.

$$\boldsymbol{\beta}_{k,\bullet} = (\beta_{k,1}, \ldots, \beta_{k,p})$$
$$\boldsymbol{\beta} = (\boldsymbol{\beta}_{H,\bullet}, \boldsymbol{\beta}_{P,\bullet}, \ldots, \boldsymbol{\beta}_{HPD,\bullet}, \boldsymbol{\beta}_{none,\bullet})$$

Hierarchy via Overlapping Group Lasso

Specifically,

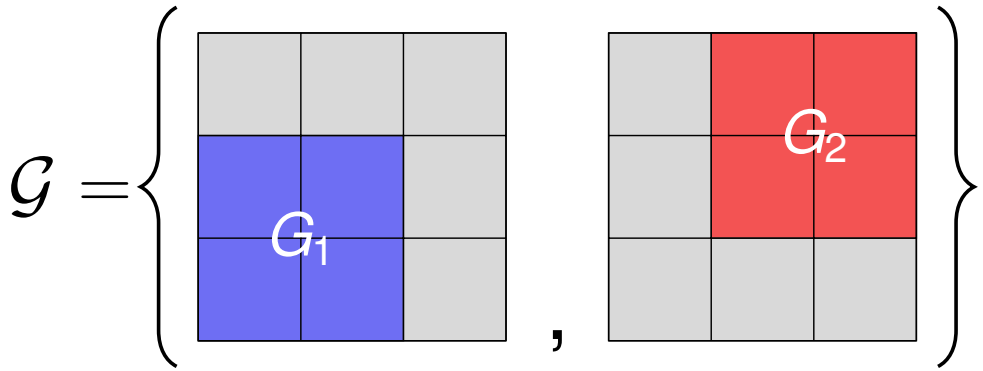
$$P(\beta) = \sum_{j=1}^{p} \sum_{G \in \mathcal{G}} \lambda_{G,j} ||\beta_{G,j}||_2,$$

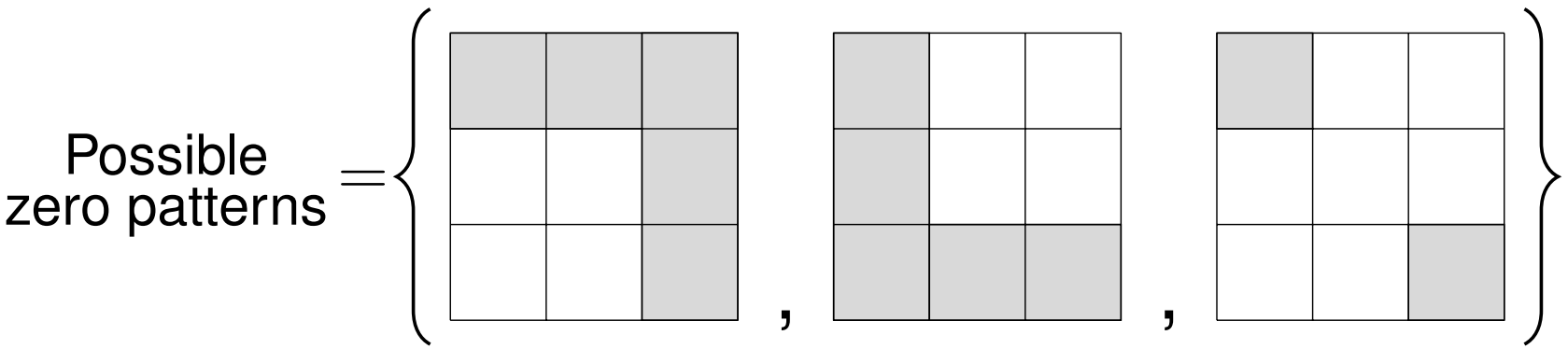
where $\beta_{G,j} \equiv \{\beta_{k,j}, k \in G\}$

- The structure of the groups in $\mathcal{G}$ determines patterns of selection
- Our main contribution is in constructing a $\mathcal{G}$ that borrows strength across subpopulations

Group Structure and Zero Patterns

Possible zero patterns are unions of groups (Jenatton et al., 2011)





Each square represents a coefficient

White squares $=$ 0 coefficients

Blue squares $=$ coefficients in group 1

Red squares $=$ coefficients in group 2

Group Structure and Zero Patterns

The possible zero patterns $\mathcal{Z}$ are all unions of groups

- $$\mathcal{Z} = \left\{ \bigcup_{G \in \mathcal{G}'} G; \mathcal{G}' \subseteq \mathcal{G} \right\}.$$
- The non-zero patterns are $\mathcal{Z}^c \equiv \{Z^c : Z \in \mathcal{Z}\}$
- In our setting,

$$\mathcal{G} = \{\overline{HPD}, \overline{HP}, \overline{HD}, \overline{PD}, \overline{H}, \overline{P}, \overline{D}, none\}$$

  - $\overline{HPD} = \{HPD, HP, HD, PD, H, P, D\}$
  - $\overline{HP} = \{HP, H, P\}$
  - $\ldots$
  - $\overline{P} = \{P\}$

Misspecified Nonzero Pattern and Recovery

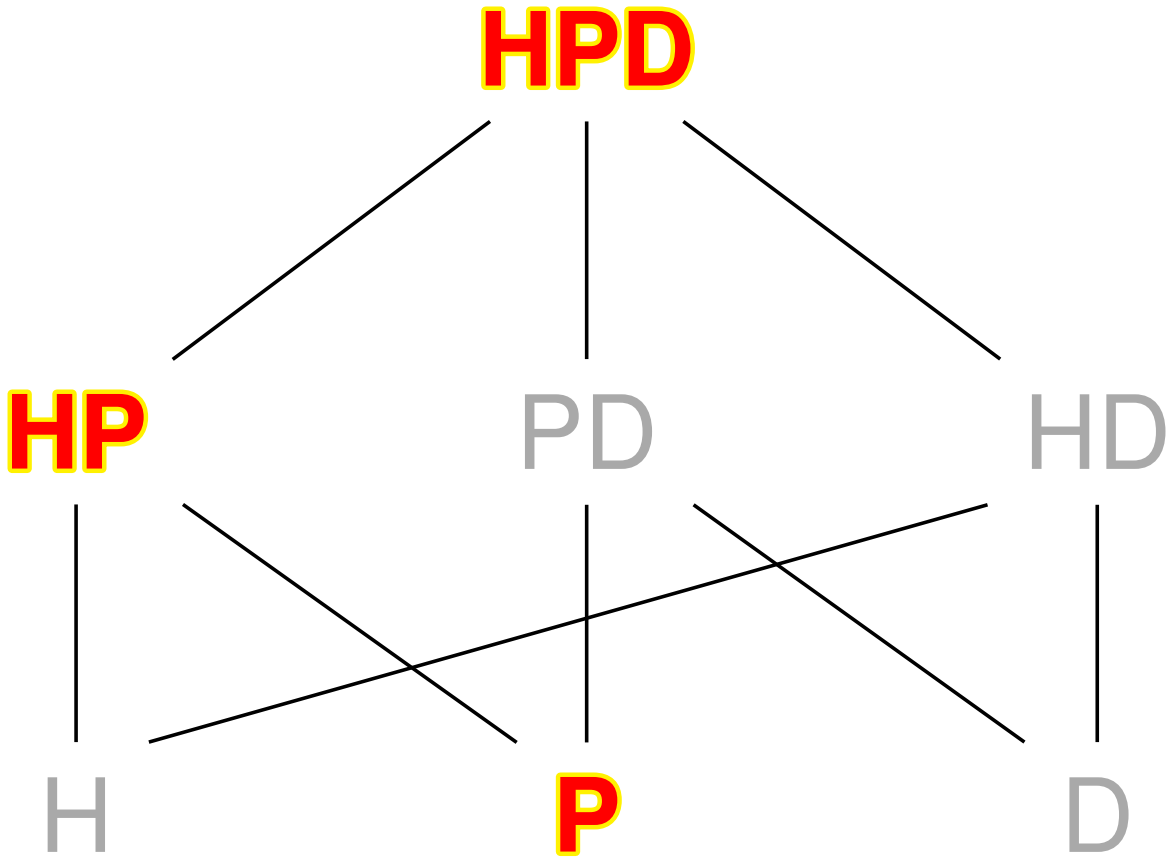
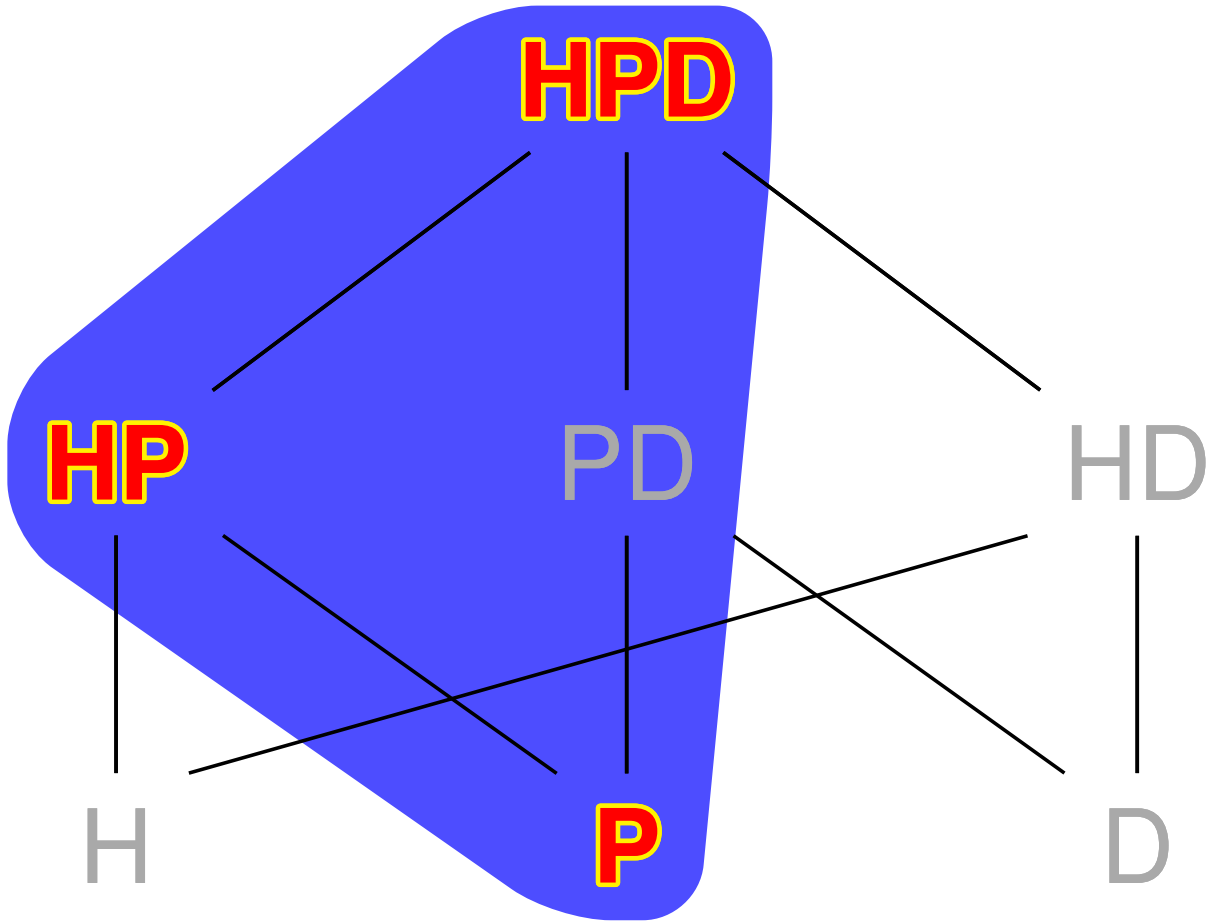


Illustration of a non-zero pattern which violates our hierarchy assumption

Misspecified Nonzero Pattern and Recovery



Our penalty will select the smallest nonzero pattern induced by $\mathcal{G}$ which covers the true nonzero pattern.

Investigating Small Sample Performance

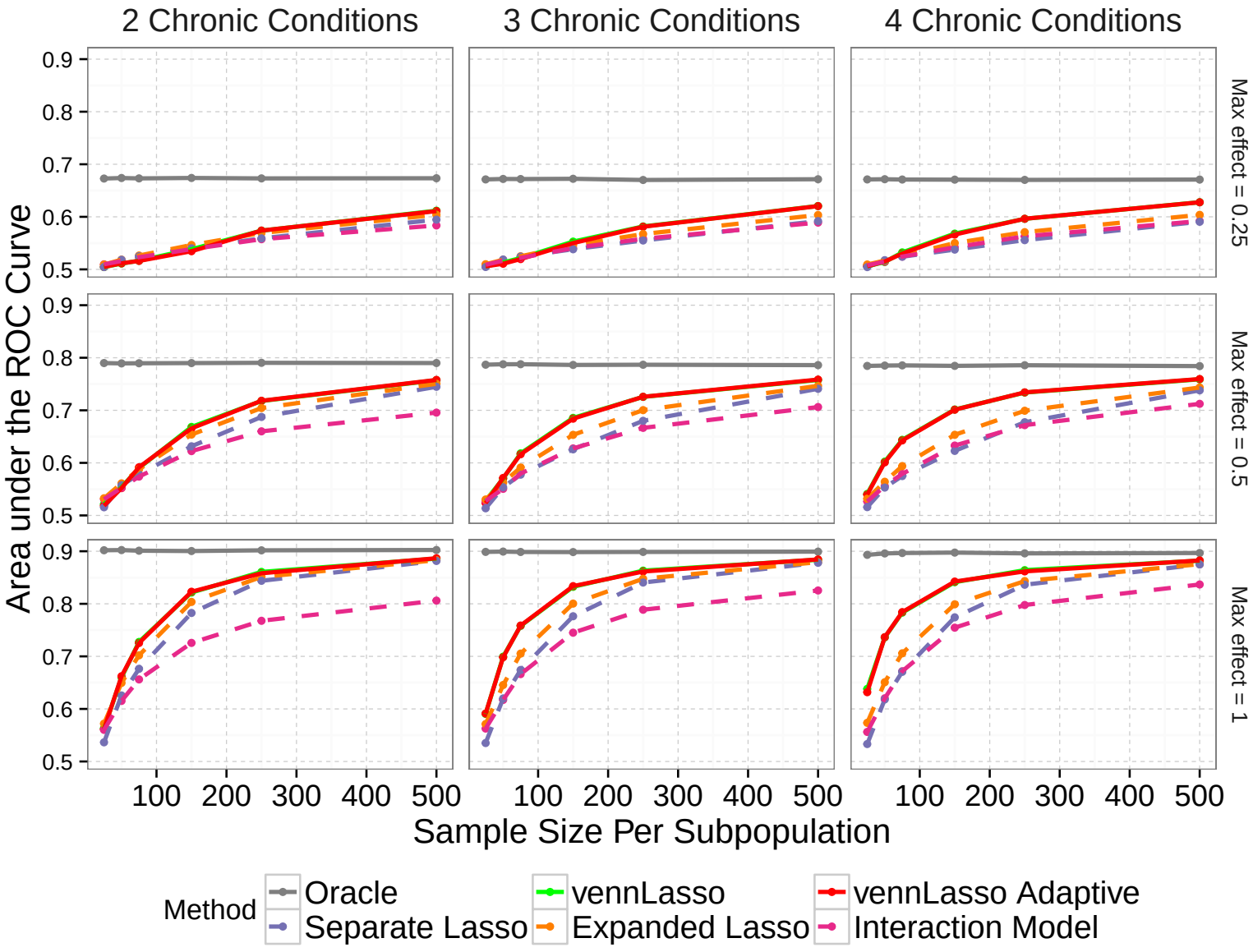
Does leveraging hierarchy help prediction?

- Outcomes simulated from heterogeneous logistic model
- Covariates meet our hierarchical assumptions
- We vary:
  - Number of conditions
  - Sample size
  - Strength of signal
- Evaluate predictive performance on test data

Methods to Compare

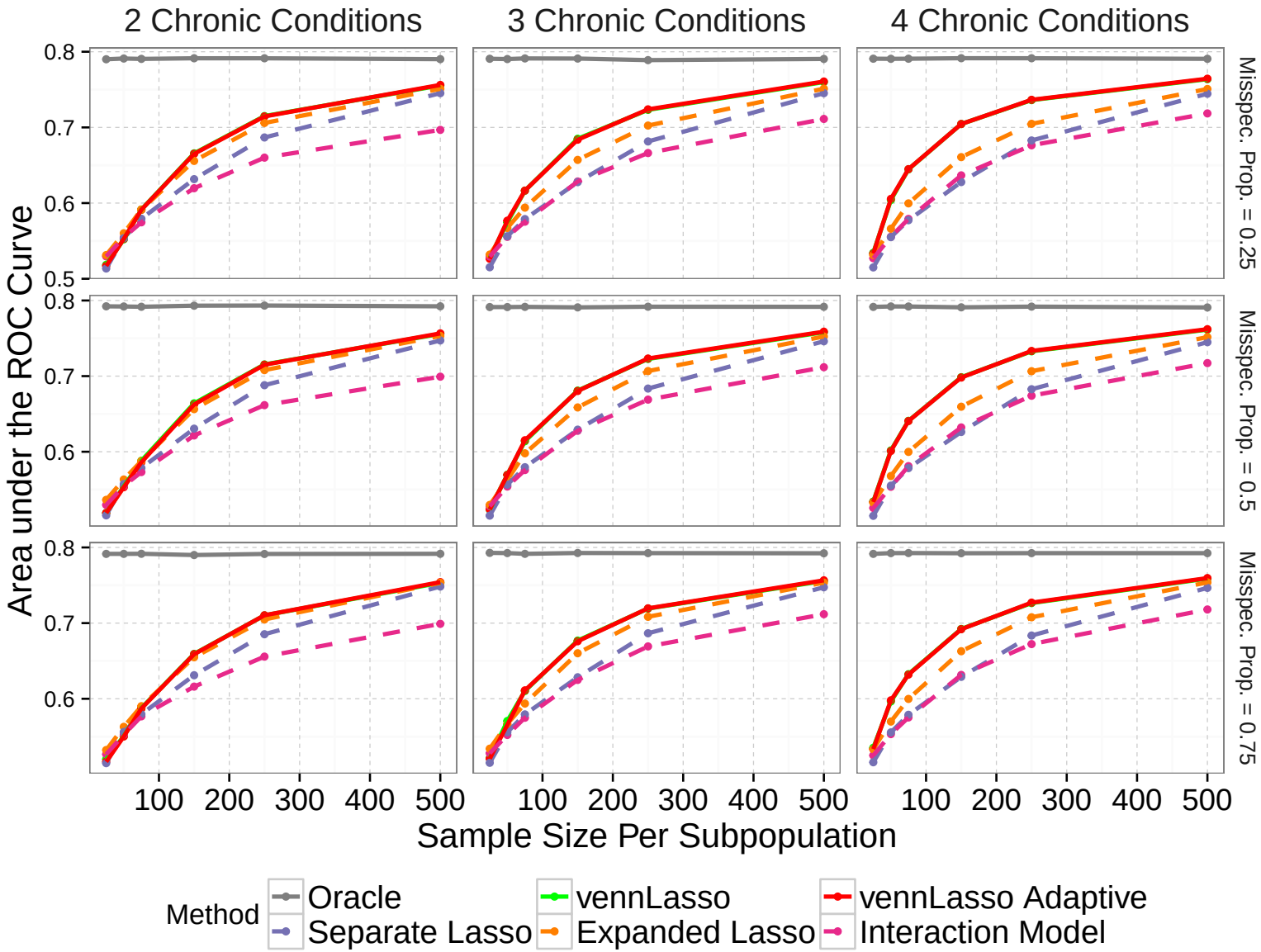
Method	Description
Oracle	The true coefficients
vennLasso	Our method, $\lambda_{G,j} = \|G\|^{1/2}$
vennLasso adaptive	Our method, $\lambda_{G,j} = \|\|\hat{\beta}^{MLE}_{G,j}\|\|_2^{-\gamma}$
Separate Lasso	$\sum_{k=1}^{K} \left(\ell_k(\beta_{k,\bullet}) - \lambda_k \|\|\beta_{k,\bullet}\|\|_1\right)$
Expanded Lasso	$\sum_{k=1}^{K} \ell_k(\beta_{k,\bullet}) - \lambda \|\|\beta\|\|_1$
Interaction Model	$\ell(\beta) - \lambda \|\|\beta\|\|_1$

Simulation Results



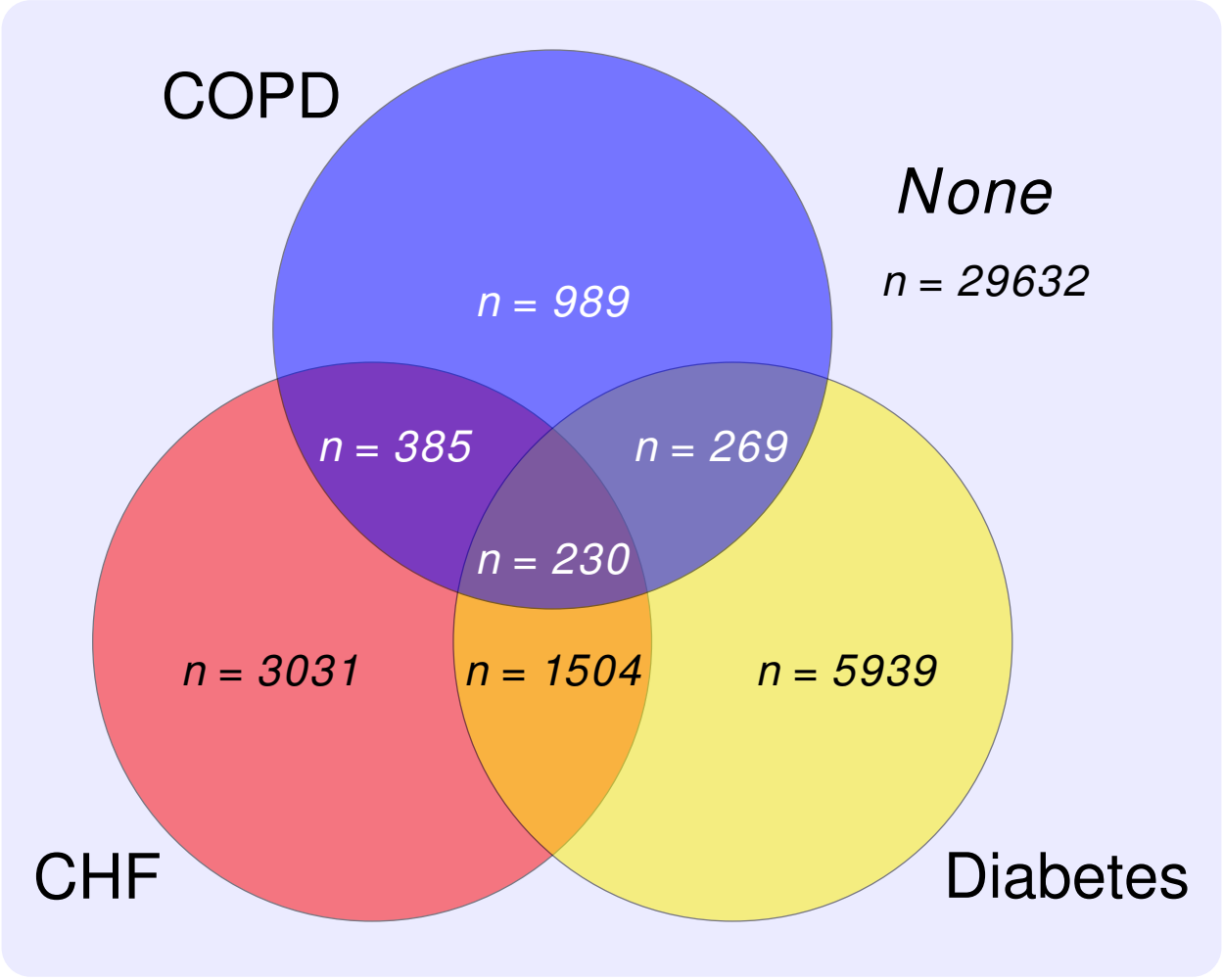
The average sparsity of the coefficients is 0.875 for this simulation.

Simulation - Hierarchy Misspecification



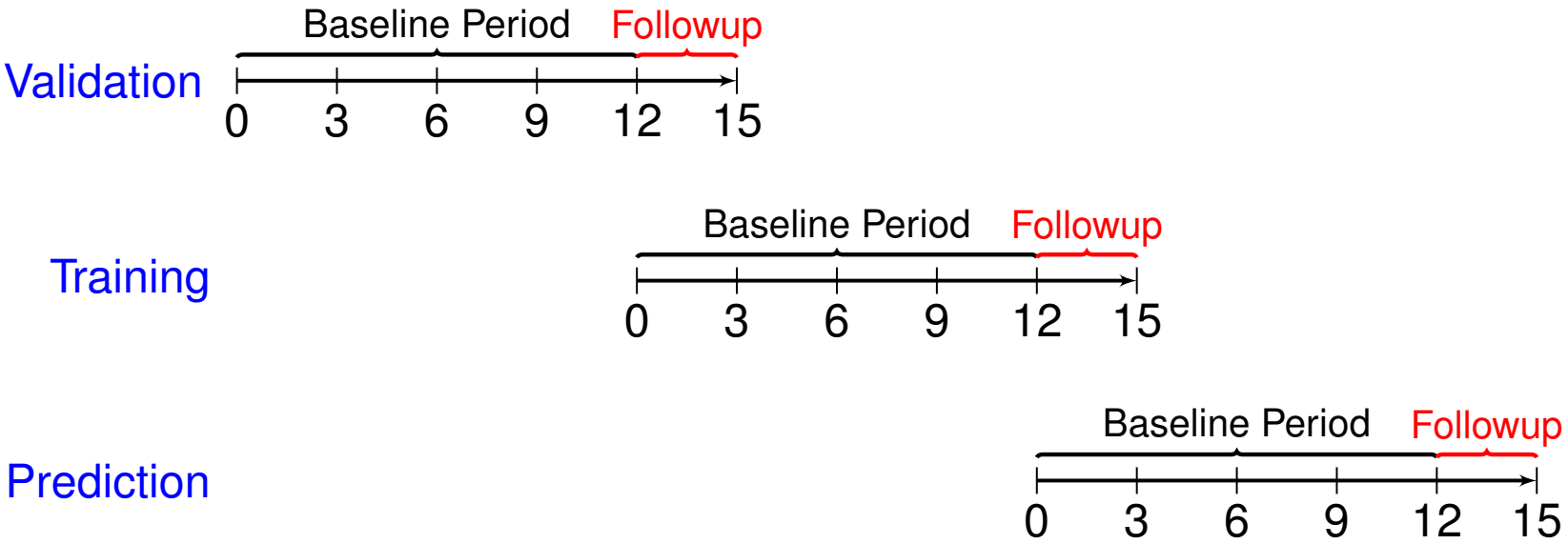
The max effect size is 0.5 for this simulation.

UWHealth Hospital Admissions Data



Sample sizes for each of the subpopulations in the UW Health admissions modeling cohort

UWHealth Hospital Admissions Data Timeline

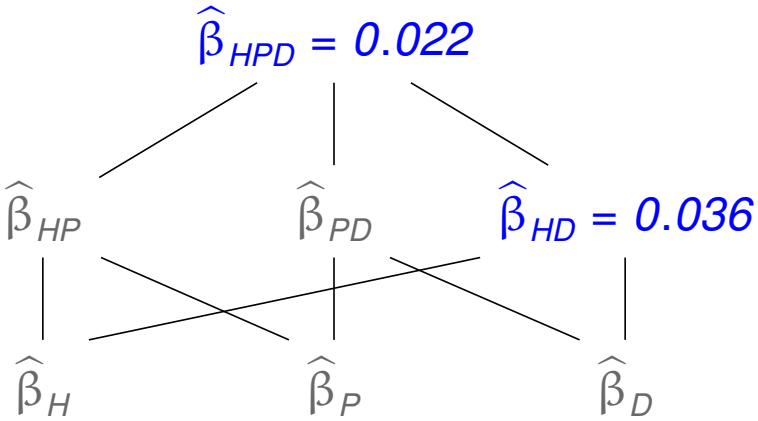


Timelines of the validation, training, and prediction datasets

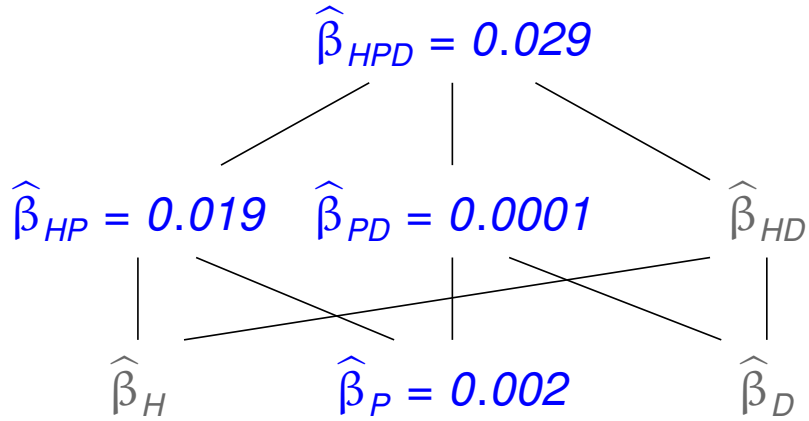
Results by Subpopulation

Subpopulation			Validation AUC			
(CHF, COPD, Diabetes)	Sample Size Train	Validation	vennLasso	Interaction Model	Separate Lasso	Expanded Lasso
(N, N, N)	29,632	28,940	0.756	0.744	**0.758**	0.676
(Y, N, N)	3,031	2,435	**0.688**	0.685	**0.688**	0.662
(N, Y, N)	989	1,047	**0.738**	0.721	0.690	0.639
(N, N, Y)	5,939	5,568	**0.726**	0.711	0.720	0.709
(Y, Y, N)	385	356	0.702	**0.717**	0.563	0.638
(Y, N, Y)	1,504	1,190	**0.705**	0.681	0.701	0.676
(N, Y, Y)	269	286	**0.779**	0.763	0.746	0.635
(Y, Y, Y)	230	204	0.599	0.601	0.619	**0.622**

Some Selected Variables



(a) Indicator of Cardiomyopathy



(b) Number of times brain natriuretic peptide (BNP) was measured during baseline

(a) Diabetic cardiomyopathy may only be relevant or predictive for patients with both diabetes and CHF

(b) BNP is often used to diagnose heart failure. Also predictive of exacerbation of stable COPD (Inoue et al., 2009)

Discussion

- Heterogeneity is common in health system risk modeling
- Introduced a hierarchical penalty to borrow strength across subpopulations with common underlying structure
- Helps substantially in modeling small subpopulations with many conditions (often these are of great interest)
- Models with interpretation specific to subpopulations

References I

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122.

Friedman, B., Jiang, H. J., and Elixhauser, A. (2008). Costly hospital readmissions and complex chronic illness. *Inquiry*, 45(4):408–421.

Gabay, D. and Mercier, B. (1976). A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2(1):17–40.

References II

Glowinski, R. and Marroco, A. (1975). Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires. *Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique*, 9(2):41–76.

Inoue, Y., Kawayama, T., Iwanaga, T., and Aizawa, H. (2009). High plasma brain natriuretic peptide levels in stable copd without pulmonary hypertension or cor pulmonale. *Internal Medicine*, 48(7):503–512.

Jenatton, R., Audibert, J.-Y., and Bach, F. (2011). Structured variable selection with sparsity-inducing norms. *The Journal of Machine Learning Research*, 12:2777–2824.

Minott, J. (2008). Reducing hospital readmissions. *Academy Health*, 23(2):1–10.

References III

Pfuntner, A., Wier, L., and Steiner, C. (2013). Costs for hospital stays in the united states, 2010: Statistical brief #146. *Healthcare Cost and Utilization Project (HCUP) Statistical Briefs*, pages 1–11.

Tannen, R., Xie, D., Wang, X., Yu, M., and Weiner, M. G. (2013). A new comparative effectiveness assessment strategy using the thin database: comparison of the cardiac complications of pioglitazone and rosiglitazone. *Pharmacoepidemiology and Drug Safety*, 22(1):86–97.

Thanks!

Code:

`github.com/jaredhuling/vennLasso`

Computation

Computation for the group lasso with overlapping groups is non-trivial.

- We utilize an alternating direction method of multipliers (ADMM) (Glowinski and Marroco, 1975; Gabay and Mercier, 1976; Boyd et al., 2011) algorithm.
- The ADMM algorithm works by decomposing an objective function and solving the decomposed subproblems iteratively.

  Example: minimize $\frac{1}{2}||\boldsymbol{Y} - \boldsymbol{X}\beta||^2 + \lambda P(\beta)$

  ADMM: minimize $\frac{1}{2}||\boldsymbol{Y} - \boldsymbol{X}\beta||^2 + \lambda P(\gamma)$ s.t. $A\beta = \gamma$

ADMM Algorithm

ADMM solves problems of the form

$$\begin{aligned} &\text{minimize } f(\boldsymbol{\beta}) + P(\boldsymbol{\gamma}) \\ &\text{subject to } A\boldsymbol{\beta} + B\boldsymbol{\gamma} = c \end{aligned}$$

where $\boldsymbol{\beta} \in \mathbb{R}^{Kp}$, $\boldsymbol{\gamma} \in \mathbb{R}^{m}$, $A \in \mathbb{R}^{r \times Kp}$, $B \in \mathbb{R}^{r \times m}$, and $c \in \mathbb{R}^{r}$

To solve the above problem, the augmented Lagrangian is formed as:

$$\begin{aligned} L_\rho(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\nu}) = f(\boldsymbol{\beta}) + P(\boldsymbol{\gamma}) + \boldsymbol{\nu}^\top (A\boldsymbol{\beta} + B\boldsymbol{\gamma} - c) \\ + (\rho/2)||A\boldsymbol{\beta} + B\boldsymbol{\gamma} - c||_2^2 \end{aligned}$$

Example: minimize $\frac{1}{2}||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \lambda||\boldsymbol{\gamma}||_1$ s.t. $\boldsymbol{\beta} = \boldsymbol{\gamma}$

ADMM Algorithm

Alternatingly minimize L_ρ with respect to β and γ and update the Lagrangian parameter ν

$$\beta^{(t+1)} = \underset{\beta}{\operatorname{argmin}}\, L_\rho(\beta, \gamma^{(t)}, \nu^{(t)}) \tag{1}$$

$$\gamma^{(t+1)} = \underset{\gamma}{\operatorname{argmin}}\, L_\rho(\beta^{(t+1)}, \gamma, \nu^{(t)}) \tag{2}$$

$$\nu^{(t+1)} = \nu^{(t)} + \rho(A\beta^{(t+1)} + B\gamma^{(t+1)} - c)$$

where t indexes the iteration number. ADMM has been shown to converge for any $\rho > 0$.

$$L_\rho(\beta, \gamma, \nu) = f(\beta) + P(\gamma) + \nu^\top(A\beta + B\gamma - c) + (\rho/2)||A\beta + B\gamma - c||_2^2$$

ADMM Algorithm for Overlapping Group Lasso

Let $A = (A_1, \ldots, A_g)$ be an $m \times Kp$ matrix where A_l is a $|G_l| \times Kp$ matrix with the (i, j)th entry equal to 1 if j is the i^{th} element of group G_l, and 0 otherwise.

For example, if p=1, K=3, $\beta = (\beta_1, \beta_2, \beta_3)^\top$ and $\mathcal{G} = \{\{1, 2\}, \{2, 3\}\}$,
then

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \text{ and } A\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_2 \\ \beta_3 \end{pmatrix}.$$

In this example, the penalty $P(\gamma)$ is

$$P(\gamma) = \lambda(||(\gamma_1, \gamma_2)||_2 + ||(\gamma_3, \gamma_4)||_2).$$

Asymptotics for Generalized Linear Models

The density of a generalized linear model with canonical link given a single observation $(y_k, \mathbf{x}_k)$ for subpopulation k can be written as:

$$f_k(y_k|\mathbf{x}_k, \theta_k) = h(y_k)\exp(y_k\theta_k - \phi(\theta_k)), \tag{3}$$

where $\theta_k = \mathbf{x}_k\beta^0_{k,\cdot}$, $\mathbf{x}_k = (x_{k,1}, \ldots, x_{k,p})$, and $\beta^0_{k,\cdot}$ are the true coefficients.

$$\underset{\beta}{\operatorname{argmin}}\Big[\sum_{k=1}^{K}\frac{1}{N}\big\{-\mathbf{Y}_k^\top(\mathbf{X}_k\beta_{k,\cdot}) + \boldsymbol{e}_k^\top\phi(\mathbf{X}_k\beta_{k,\cdot})\big\}\Big] + \lambda P(\beta), \tag{4}$$

Asymptotics for Generalized Linear Models

(C.1) $\mathbf{I}^k = \mathrm{E}_k[\phi''(\mathbf{x}_k\beta^0_{k,\cdot})\mathbf{x}_k\mathbf{x}_k^\top]$ is finite and postive definite, where $\mathrm{E}_k[\cdot]$ is the expectation w.r.t $\mathbf{x}_k$ under the measure of subpopulation k.

(C.2) For subpopulation k, there is a sufficiently large enough open set $\mathcal{O}_k$ that contains $\beta^0_{k,\cdot}$ such that $\forall\beta_{k,\cdot} \in \mathcal{O}_k$,

$$|\phi'''(\mathbf{x}_k\beta_{k,\cdot})| \leq M_k(\mathbf{x}_k) < \infty,$$

and

$$\mathrm{E}_k[M_k(\mathbf{x}_k)|x_{k,j}x_{k,l}x_{k,m}|] < \infty,$$

for all $1 \leq j, l, m \leq p$.

(C.3) $0 < \inf_{k=1,\ldots,K} \liminf_{N\to+\infty} \frac{n_k}{N} \leq \sup_{k=1,\ldots,K} \limsup_{N\to+\infty} \frac{n_k}{N} < 1$.

Asymptotics for Generalized Linear Models

Group structure is correct

Theorem 1

Assume the data are generated under the model represented by (3) and that our estimator is given by (4). Furthermore, assume that the non-zero patterns $\mathcal{Z}$ induced by the specified group structure $\mathcal{G}$ contain the true zero pattern. Let $\lambda_{G,j} = ||\hat{\beta}^{MLE}_{G,j}||_2^{-\gamma}$ for some $\gamma > 0$ such that $N^{(\gamma+1)/2}\lambda \to \infty$. If $\sqrt{N}\lambda \to 0$ and our regularity conditions hold, then we have the following:

$$P(\hat{J}_{\cdot,j} = J_{\cdot,j}) \to 1 \text{ as } N \to \infty, \tag{5}$$

and

$$\sqrt{n_k}(\hat{\beta}_{k,\cdot} - \beta^0_{k,\cdot}) \xrightarrow{d} \mathbf{Z}_k, \tag{6}$$

where $\mathbf{Z}_{k,J_{k,\cdot}} \sim N_{|J_{k,\cdot}|}(0, (\mathbf{I}^k_{J_{k,\cdot}J_{k,\cdot}})^{-1})$ *and* $\mathbf{Z}_{k,J^c_{k,\cdot}} = \mathbf{0}$.

Asymptotics for Generalized Linear Models

Group structure is misspecified

Theorem 2

Assume the data are generated under the model represented by (3) and that our estimator is given by (4). Here we do not necessarily assume that the group structure is correctly specifiied. Let $\lambda_{G,j} = ||\hat{\beta}^{MLE}_{G,j}||_2^{-\gamma}$ *for some* $\gamma > 0$ *such that* $N^{(\gamma+1)/2}\lambda \to \infty$. *If* $\sqrt{N}\lambda \to 0$ *and our regularity conditions hold, then we have the following:*

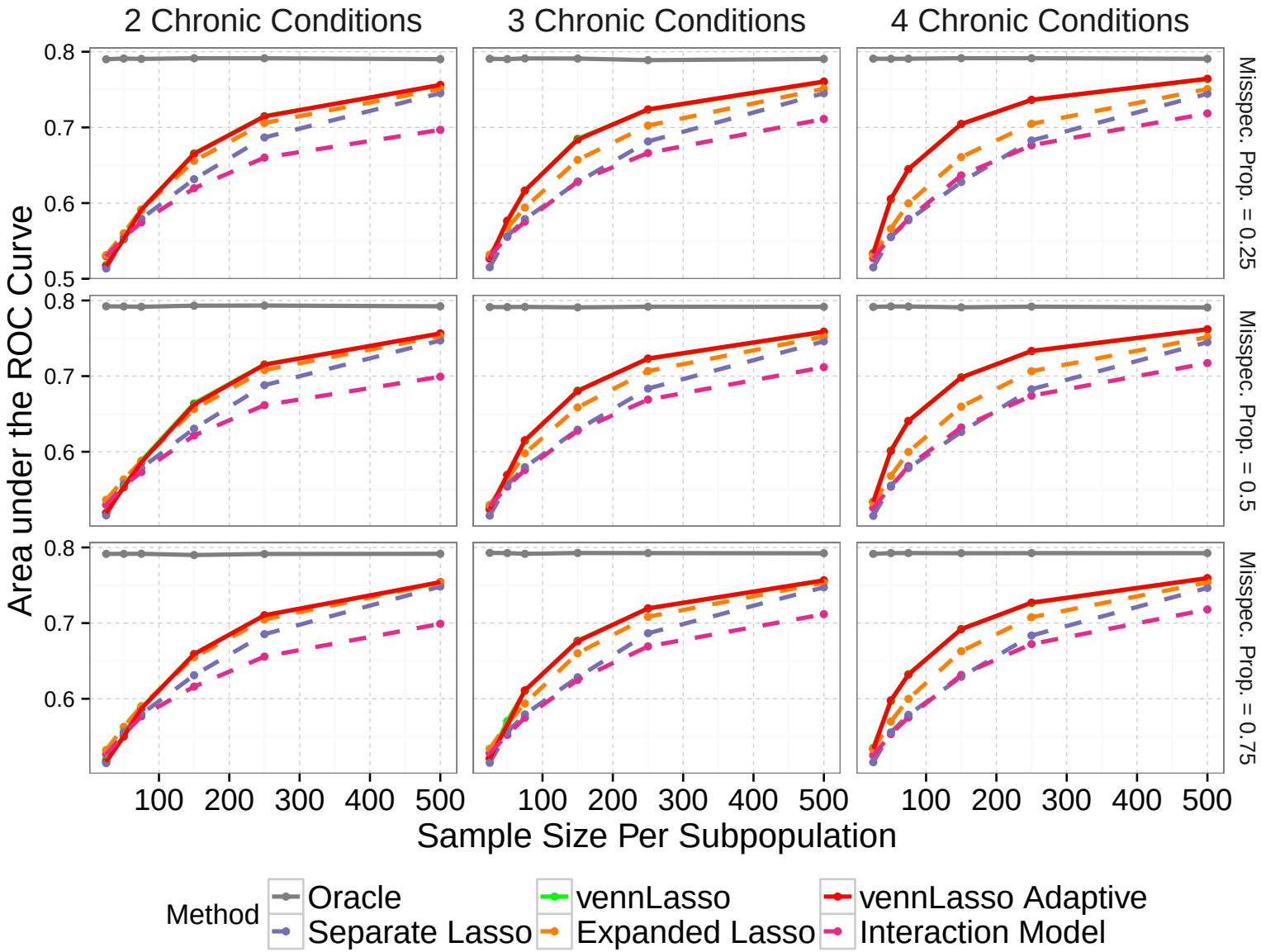
$$P(\hat{J}_{\cdot,j} = Hull(J_{\cdot,j})) \to 1 \text{ as } N \to \infty, \tag{7}$$

and

$$\sqrt{n_k}(\hat{\beta}_{k,\cdot} - \beta^0_{k,\cdot}) \xrightarrow{d} \mathbf{Z}_k, \tag{8}$$

where $\mathbf{Z}_{k,H_{k,\cdot}} \sim N_{|H_{k,\cdot}|}(0, (\mathbf{I}^k_{H_{k,\cdot}H_{k,\cdot}})^{-1})$ *and* $\mathbf{Z}_{k,H^c_{k,\cdot}} = \mathbf{0}$.

Simulation - Hierarchy Misspecification



The max effect size is 0.5 for this simulation.

Empirical coverage for all nonzero coefficients

N	Conditions	Signal-to-Noise Ratio 0.5	1	2
150	2	0.910	0.982	0.946
	3	0.994	0.982	0.957
	4	0.998	0.988	0.971
250	2	0.963	0.930	0.925
	3	0.968	0.946	0.944
	4	0.976	0.960	0.956
500	2	0.926	0.924	0.930
	3	0.940	0.939	0.942
	4	0.954	0.949	0.950

Empirical coverage results for 95% confidence intervals

Results by Subpopulation - Random Split

Subpopulation	Sample Size		Validation AUC			
(CHF, COPD, Diabetes)	Train	Validation	vennLasso	Interaction Model	Separate Lasso	Expanded Lasso
(N, N, N)	14,939	14,693	0.760	0.769	**0.770**	0.701
(Y, N, N)	1,488	1,543	**0.692**	0.687	0.683	0.665
(N, Y, N)	471	518	**0.727**	0.667	0.604	0.687
(N, N, Y)	2,917	3,022	**0.699**	0.690	0.679	0.649
(Y, Y, N)	196	189	0.587	**0.609**	0.583	0.512
(Y, N, Y)	720	784	0.752	**0.760**	0.706	0.722
(N, Y, Y)	138	131	**0.727**	0.688	0.569	0.510
(Y, Y, Y)	120	110	**0.619**	0.567	0.501	0.533