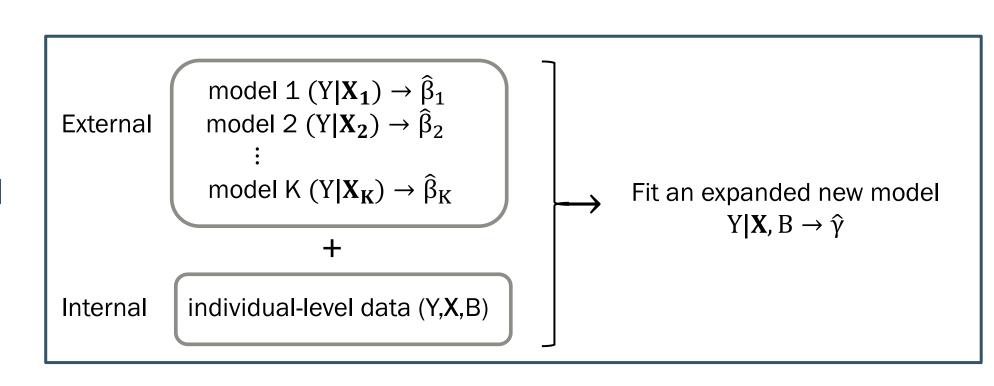
# An Ensemble meta-prediction framework to integrate multiple regression models into a current study





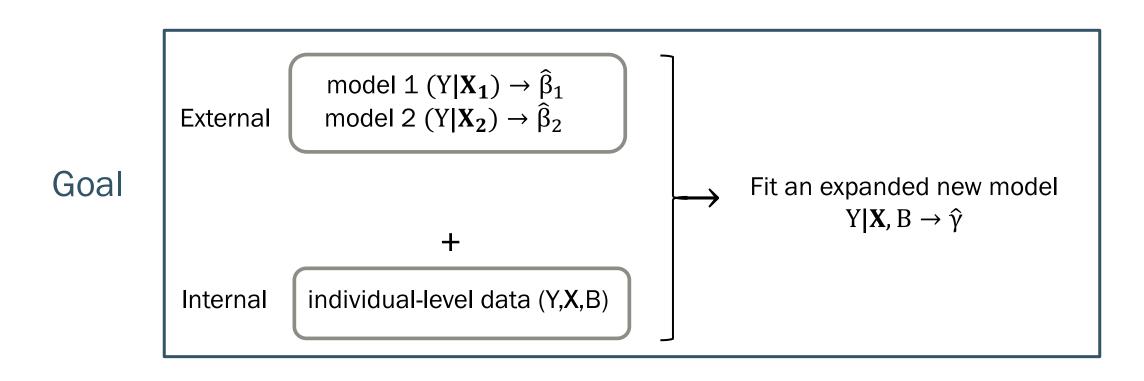
# Background & Settings

### Background in statistical language



Goal

# For simplicity, all the following illustration will be using the case of 2 external models



# Proposed Framework & Estimates

### **Key notations**

 $\hat{\gamma}_I$  maximum likelihood estimate (MLE) using the complete internal data

 $\hat{\gamma}_{CML}$  the constrained ML (CML) estimate

 $\hat{\gamma}_{EB}$  the empirical Bayes (EB) estimate

### 2-Step Framework

### Proposed Framework

#### Step 1:

External 
$$\hat{\boldsymbol{\beta}}_1 \xrightarrow{\text{Chatterjee's}} \hat{\boldsymbol{\gamma}}_{CML^1} \xrightarrow{\text{Estes'}} \hat{\boldsymbol{\gamma}}_{EB^1} = \hat{W}_1 \hat{\boldsymbol{\gamma}}_I + (I - \hat{W}_1) \hat{\boldsymbol{\gamma}}_{CML^1}$$
External  $\hat{\boldsymbol{\beta}}_2 \xrightarrow{\text{Chatterjee's}} \hat{\boldsymbol{\gamma}}_{CML^2} \xrightarrow{\text{Estes'}} \hat{\boldsymbol{\gamma}}_{EB^2} = \hat{W}_2 \hat{\boldsymbol{\gamma}}_I + (I - \hat{W}_2) \hat{\boldsymbol{\gamma}}_{CML^2}$ 

Step 2:

$$egin{bmatrix} \hat{m{\gamma}}_{EB^1} \ \hat{m{\gamma}}_{EB^2} \end{bmatrix} \sim m{N}(m{\gamma}_0, m{V}),$$

# Constrained Maximum Likelihood Estimation for Model Calibration Using Summary-Level Information From External Big Data Sources

Nilanjan Chatterjee, Yi-Hau Chen, Paige Maas, and Raymond J. Carroll

Proposed Framework

External 
$$\hat{\boldsymbol{\beta}}_1$$
 Chatterjee's  $\hat{\boldsymbol{\gamma}}_{CML^1} \xrightarrow{\text{Estes'}} \hat{\boldsymbol{\gamma}}_{EB^1} = \hat{W}_1 \hat{\boldsymbol{\gamma}}_I + (I - \hat{W}_1) \hat{\boldsymbol{\gamma}}_{CML^1}$  External  $\hat{\boldsymbol{\beta}}_2 \xrightarrow{\text{Chatterjee's}} \hat{\boldsymbol{\gamma}}_{CML^2} \xrightarrow{\text{Estes'}} \hat{\boldsymbol{\gamma}}_{EB^2} = \hat{W}_2 \hat{\boldsymbol{\gamma}}_I + (I - \hat{W}_2) \hat{\boldsymbol{\gamma}}_{CML^2}$ 

Step 2:

$$egin{bmatrix} \hat{\gamma}_{EB^1} \ \hat{\gamma}_{EB^2} \end{bmatrix} \sim \textit{N}(\gamma_0, \textit{V}),$$

#### Assumptions

1 Known form of  $g_{\beta}(Y|X)$ 

 $f_{\gamma}(Y|X,B)$  is correctly-specified

F(X,B) is the same in internal and external studies

Untestable
Transportability
assumptions may
lead to bias

Construct constraints

1  $p_i \equiv Pr(X = X_i, B = B_i), \sum_{i=1}^{n} p_i = 1$ 2 Connect external score function to internal model

 $0 = E_{Y,X,B}(\frac{\partial I_{\beta}(Y|X)}{\partial \beta}) = E_{Y,X,B}(U_{\beta}(Y|X))$   $= E_{X,B}\{E_{Y|X,B}(U_{\beta}(Y|X)|X,B)\}$   $= \int_{X,B} \int_{Y|X,B} U_{\beta}(Y|X)f_{\gamma}(Y|X,B)dYdF(X,B)$   $= \sum_{X,B} \int_{Y|X,B} U_{\beta}(Y|X)f_{\gamma}(Y|X,B)dYp_{i}$ 

Maximize the likelihood  $\prod_{i=1}^n f_{\gamma}(Y_i|X_i,B_i)p_i$  w.r.t  $\gamma$  and  $p_i$  subject to two constraints  $\sum_{i=1}^n p_i = 1 \text{ and } \sum_{i=1}^n \int U_{\beta}(Y|X)f_{\gamma}(Y|X,B)dYp_i = 0$ 

$$egin{aligned} \hat{\gamma}_{\mathit{CML}} &= rg \max_{\gamma, 
ho_i} \{\prod_{i=1}^n f_\gamma(Y_i|X_i, B_i) 
ho_i \ &+ \lambda_1 (\sum_{i=1}^n 
ho_i - 1) \ &+ \lambda_2 \sum_{i=1}^n \int U_eta(Y|X) f_\gamma(Y|X, B) dY 
ho_i \} \end{aligned}$$

# Proposed Framework

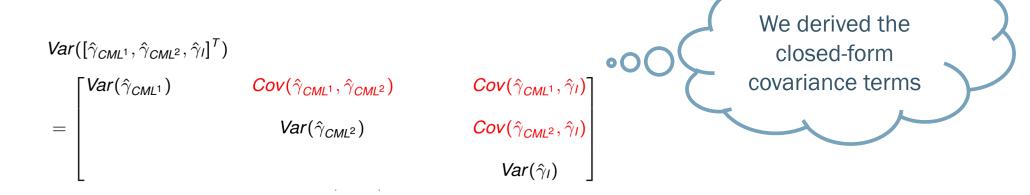
#### Step 1:

External 
$$\hat{\boldsymbol{\beta}}_1$$
 Chatterjee's  $\hat{\boldsymbol{\gamma}}_{CML^1} \xrightarrow{\text{Estes'}} \hat{\boldsymbol{\gamma}}_{EB^1} = \hat{\boldsymbol{W}}_1 \hat{\boldsymbol{\gamma}}_I + (I - \hat{\boldsymbol{W}}_1) \hat{\boldsymbol{\gamma}}_{CML^1}$  External  $\hat{\boldsymbol{\beta}}_2 \xrightarrow{\text{Chatterjee's}} \hat{\boldsymbol{\gamma}}_{CML^2} \xrightarrow{\text{Estes'}} \hat{\boldsymbol{\gamma}}_{EB^2} = \hat{\boldsymbol{W}}_2 \hat{\boldsymbol{\gamma}}_I + (I - \hat{\boldsymbol{W}}_2) \hat{\boldsymbol{\gamma}}_{CML^2}$ 

Step 2:

$$egin{bmatrix} \hat{m{\gamma}}_{EB^1} \ \hat{m{\gamma}}_{EB^2} \end{bmatrix} \sim m{N}(m{\gamma}_0, m{V}),$$

$$\hat{m{\gamma}}_0 = \mathsf{A}$$
 weighted average of  $\hat{m{\gamma}}_{\mathit{EB}^1}$  and  $\hat{m{\gamma}}_{\mathit{EB}^2}$ 



# Proposed Framework

External 
$$\hat{\boldsymbol{\beta}}_1 \xrightarrow{\text{Chatterjee's}} \hat{\boldsymbol{\gamma}}_{CML^1} \xrightarrow{\text{Estes'}} \hat{\boldsymbol{\gamma}}_{EB^1} = \hat{W}_1 \hat{\boldsymbol{\gamma}}_I + (I - \hat{W}_1) \hat{\boldsymbol{\gamma}}_{CML^1}$$
External  $\hat{\boldsymbol{\beta}}_2 \xrightarrow{\text{Chatterjee's}} \hat{\boldsymbol{\gamma}}_{CML^2} \xrightarrow{\text{Estes'}} \hat{\boldsymbol{\gamma}}_{EB^2} = \hat{W}_2 \hat{\boldsymbol{\gamma}}_I + (I - \hat{W}_2) \hat{\boldsymbol{\gamma}}_{CML^2}$ 

Step 2:

$$egin{bmatrix} \hat{m{\gamma}}_{EB^1} \ \hat{m{\gamma}}_{EB^2} \end{bmatrix} \sim N(m{\gamma}_0, V),$$

#### **Empirical Bayes Estimation and Prediction Using** Summary-Level Information From External Big Data Sources Adjusting for Violations of Transportability

Jason P. Estes<sup>1</sup> · Bhramar Mukherjee<sup>1</sup> · Jeremy M. G. Taylor<sup>1</sup>

## Proposed Framework |

#### Step 1:

External 
$$\hat{m{\beta}}_1 \xrightarrow{\text{Chatterjee's}} \hat{m{\gamma}}_{\textit{CML}^1} \xrightarrow{\text{Estes'}} \hat{m{\gamma}}_{\textit{EB}^1} = \hat{\pmb{W}}_1 \hat{m{\gamma}}_I + (I - \hat{\pmb{W}}_1) \hat{m{\gamma}}_{\textit{CML}^1}$$
External  $\hat{m{\beta}}_2 \xrightarrow{\text{Chatterjee's}} \hat{m{\gamma}}_{\textit{CML}^2} \xrightarrow{\text{Estes'}} \hat{m{\gamma}}_{\textit{EB}^2} = \hat{\pmb{W}}_2 \hat{m{\gamma}}_I + (I - \hat{\pmb{W}}_2) \hat{m{\gamma}}_{\textit{CML}^2}$ 

External 
$$\hat{m{eta}}_2 \xrightarrow{ ext{Chatterjee's}} \hat{m{\gamma}}_{\textit{CML}^2} \xrightarrow{ ext{Estes'}} \hat{m{\gamma}}_{\textit{EB}^2} = \hat{\pmb{W}}_2 \hat{m{\gamma}}_\emph{I} + (\emph{I} - \hat{\pmb{W}}_2) \hat{m{\gamma}}_\textit{CML}^2$$

Step 2:

$$egin{bmatrix} \hat{m{\gamma}}_{EB^1} \ \hat{m{\gamma}}_{EB^2} \end{bmatrix} \sim m{N}(m{\gamma}_0, m{V}),$$

$$\hat{m{\gamma}}_0 = {\sf A}$$
 weighted average of  $\hat{m{\gamma}}_{\it EB^1}$  and  $\hat{m{\gamma}}_{\it EB^2}$ 

To correct the potential bias due to incomparability: the larger the difference between internal & external estimates, the more  $\hat{\gamma}_{EB}$  shrinks toward  $\hat{\gamma}_{I}$ 

$$\begin{cases} \gamma \sim \textit{N}(\gamma_0, A) \\ \hat{\gamma}_I | \gamma \sim \textit{N}(\gamma, \Sigma) \end{cases} \Rightarrow \hat{\gamma}_{\textit{posterior}} = \textit{A}(\Sigma + A)^{-1} \hat{\gamma}_I + \Sigma(\Sigma + A)^{-1} \gamma_0$$

$$\hat{\gamma}_{\textit{EB}} = \hat{\textit{A}}(\hat{\Sigma} + \hat{\textit{A}})^{-1} \hat{\gamma}_I + \hat{\Sigma}(\hat{\Sigma} + \hat{\textit{A}})^{-1} \hat{\gamma}_{\textit{CML}}$$

$$= \hat{\textit{W}} \hat{\gamma}_I + (\textit{I} - \hat{\textit{W}}) \hat{\gamma}_{\textit{CML}}$$

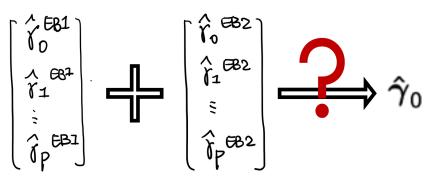
$$\begin{cases} \hat{\textit{A}} = (\hat{\gamma}_I - \hat{\gamma}_{\textit{CML}})(\hat{\gamma}_I - \hat{\gamma}_{\textit{CML}})^T \\ \hat{\Sigma} = \textit{Var}(\hat{\gamma}_I) \end{cases}$$

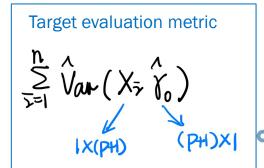
### Proposed Framework

External 
$$\hat{\boldsymbol{\beta}}_1 \xrightarrow{\text{Chatterjee's}} \hat{\boldsymbol{\gamma}}_{CML^1} \xrightarrow{\text{Estes'}} \hat{\boldsymbol{\gamma}}_{EB^1} = \hat{W}_1 \hat{\boldsymbol{\gamma}}_I + (I - \hat{W}_1) \hat{\boldsymbol{\gamma}}_{CML^1}$$
External  $\hat{\boldsymbol{\beta}}_2 \xrightarrow{\text{Chatterjee's}} \hat{\boldsymbol{\gamma}}_{CML^2} \xrightarrow{\text{Estes'}} \hat{\boldsymbol{\gamma}}_{EB^2} = \hat{W}_2 \hat{\boldsymbol{\gamma}}_I + (I - \hat{W}_2) \hat{\boldsymbol{\gamma}}_{CML^2}$ 

Step 2:

$$egin{bmatrix} \hat{\gamma}_{EB^1} \ \hat{\gamma}_{EB^2} \end{bmatrix} \sim \textit{N}(\gamma_0, \textit{V}),$$





Goal: combine 2  $\hat{\gamma}_{EB}$ 's to minimize the expected prediction variance, which reduce the p-dimension optimization to a scalar

# Proposed Framework

External 
$$\hat{\boldsymbol{\beta}}_1 \xrightarrow{\text{Chatterjee's}} \hat{\boldsymbol{\gamma}}_{CML^1} \xrightarrow{\text{Estes'}} \hat{\boldsymbol{\gamma}}_{EB^1} = \hat{W}_1 \hat{\boldsymbol{\gamma}}_I + (I - \hat{W}_1) \hat{\boldsymbol{\gamma}}_{CML^1}$$
External  $\hat{\boldsymbol{\beta}}_2 \xrightarrow{\text{Chatterjee's}} \hat{\boldsymbol{\gamma}}_{CML^2} \xrightarrow{\text{Estes'}} \hat{\boldsymbol{\gamma}}_{EB^2} = \hat{W}_2 \hat{\boldsymbol{\gamma}}_I + (I - \hat{W}_2) \hat{\boldsymbol{\gamma}}_{CML^2}$ 

Step 2:

$$egin{bmatrix} \hat{\hat{\gamma}}_{EB^1} \ \hat{\hat{\gamma}}_{EB^2} \end{bmatrix} \sim \textit{N}(\gamma_0, \textit{V}),$$

$$\begin{bmatrix}
\hat{\gamma}_0 & \oplus 2 \\
\hat{\gamma}_1 & \oplus 2 \\
\hat{\gamma}_1 & \oplus 2
\end{bmatrix}$$

$$\hat{\gamma}_0 & \oplus 2 \\
\hat{\gamma}_1 & \oplus 2
\end{bmatrix}$$

$$\hat{\gamma}_0 & \oplus 2$$

#### Naïve approach:

Inverse variance weighed estimate, ignoring the correlation between  $\widehat{\gamma}_{EB}$  's

$$V = Var \begin{bmatrix} \hat{\gamma}_{EB^1} \\ \hat{\gamma}_{EB^2} \end{bmatrix} = \begin{bmatrix} var(\hat{\gamma}_{EB^1}) & Cov(\hat{\gamma}_{EB^1}, \hat{\gamma}_{EB^2}) \\ Var(\hat{\gamma}_{EB^2}) \end{bmatrix}$$

# Proposed Framework

#### Step 1:

External 
$$\hat{\boldsymbol{\beta}}_1 \xrightarrow{\text{Chatterjee's}} \hat{\boldsymbol{\gamma}}_{CML^1} \xrightarrow{\text{Estes'}} \hat{\boldsymbol{\gamma}}_{EB^1} = \hat{W}_1 \hat{\boldsymbol{\gamma}}_I + (I - \hat{W}_1) \hat{\boldsymbol{\gamma}}_{CML^1}$$
External  $\hat{\boldsymbol{\beta}}_2 \xrightarrow{\text{Chatterjee's}} \hat{\boldsymbol{\gamma}}_{CML^2} \xrightarrow{\text{Estes'}} \hat{\boldsymbol{\gamma}}_{EB^2} = \hat{W}_2 \hat{\boldsymbol{\gamma}}_I + (I - \hat{W}_2) \hat{\boldsymbol{\gamma}}_{CML^2}$ 

Step 2:

$$egin{bmatrix} \hat{\hat{\gamma}}_{EB^1} \ \hat{\hat{\gamma}}_{EB^2} \end{bmatrix} \sim \textit{N}(\gamma_0, \textit{V}),$$

### Proposed **Estimates**

Waight 1 
$$\begin{pmatrix} \hat{\gamma}_0 & \text{EB1} \\ \hat{\gamma}_0 & \text{EB1} \\ \hat{\gamma}_1 & \text{EB1} \end{pmatrix} + \begin{pmatrix} 1 - \text{weight 1} \end{pmatrix} \begin{bmatrix} \hat{\gamma}_0 & \text{EB2} \\ \hat{\gamma}_1 & \text{EB2} \\ \hat{\gamma}_1 & \text{EB1} \end{bmatrix}$$

$$\text{Weight 1} \begin{bmatrix} \hat{\gamma}_0 \text{ EB1} \\ \hat{\gamma}_1 \text{ EB1} \\ \hat{\gamma}_p \text{ EB1} \end{bmatrix} + (1 - \text{Weight 1}) \begin{bmatrix} \hat{\gamma}_0 \text{ EB2} \\ \hat{\gamma}_1 \text{ EB2} \\ \hat{\gamma}_p \text{ EB2} \end{bmatrix}$$

$$\text{Weight 2} \hat{\gamma}_1 \text{ EB1} + (1 - \text{Weight 2}) \hat{\gamma}_1 \text{ EB2}$$

$$\text{Weight PH} \hat{\gamma}_p \text{ EB1} + (1 - \text{Weight PH}) \hat{\gamma}_p \text{ EB2}$$

$$\text{Weight PH} \hat{\gamma}_p \text{ EB1} + (1 - \text{Weight PH}) \hat{\gamma}_p \text{ EB2}$$

### Proposed Framework

#### Step 1:

External 
$$\hat{\boldsymbol{\beta}}_1 \xrightarrow{\text{Chatterjee's}} \hat{\boldsymbol{\gamma}}_{CML^1} \xrightarrow{\text{Estes'}} \hat{\boldsymbol{\gamma}}_{EB^1} = \hat{W}_1 \hat{\boldsymbol{\gamma}}_I + (I - \hat{W}_1) \hat{\boldsymbol{\gamma}}_{CML^1}$$
External  $\hat{\boldsymbol{\beta}}_2 \xrightarrow{\text{Chatterjee's}} \hat{\boldsymbol{\gamma}}_{CML^2} \xrightarrow{\text{Estes'}} \hat{\boldsymbol{\gamma}}_{EB^2} = \hat{W}_2 \hat{\boldsymbol{\gamma}}_I + (I - \hat{W}_2) \hat{\boldsymbol{\gamma}}_{CML^2}$ 

Step 2:

$$egin{bmatrix} \hat{\gamma}_{EB^1} \ \hat{\gamma}_{EB^2} \end{bmatrix} \sim \textit{N}(\gamma_0, \textit{V}),$$

$$\hat{m{\gamma}}_0 = {\sf A}$$
 weighted average of  $\hat{m{\gamma}}_{\it EB^1}$  and  $\hat{m{\gamma}}_{\it EB^2}$ 

#### Optimal Covariance Weighted Estimate

#### Selective Coefficient Learner

# Proposed Estimates

Weighting across different  $\hat{\gamma}_{EB}$ 's: same weight for all coefficient within one  $\hat{\gamma}_{EB}$ 

# Proposed Framework

Waight 1 
$$\begin{pmatrix} \hat{\gamma}_0 \\ \hat{\gamma}_0 \\ \hat{\gamma}_1 \end{pmatrix} + (1 - \text{weight 1}) \begin{pmatrix} \hat{\gamma}_0 \\ \hat{\gamma}_0 \\ \hat{\gamma}_1 \end{pmatrix} + \hat{\gamma}_1 \end{pmatrix}$$

$$\begin{bmatrix} \text{weight 1 } \hat{f}_0^{\text{EB1}} + (1-\text{weight 1}) \hat{f}_0^{\text{EB2}} \\ \text{weight 2 } \hat{f}_1^{\text{EB1}} + (1-\text{weight 2}) \hat{f}_1^{\text{EB2}} \\ \\ \text{weight PH} \hat{f}_p^{\text{EB1}} + (1-\text{weight pH}) \hat{f}_p^{\text{EB2}} \end{bmatrix}$$

Weighting across each coefficient: different

weight for coefficients within one  $\hat{\gamma}_{FR}$ 

#### Step 1:

External 
$$\hat{m{\beta}}_1 \xrightarrow{\text{Chatterjee's}} \hat{m{\gamma}}_{CML^1} \xrightarrow{\text{Estes'}} \hat{m{\gamma}}_{EB^1} = \hat{\pmb{W}}_1 \hat{m{\gamma}}_I + (\hat{m{\gamma}}_I + \hat{m{\gamma}}_I)$$

External  $\hat{m{\beta}}_2 \xrightarrow{\text{Chatterjee's}} \hat{m{\gamma}}_{\textit{CML}^2} \xrightarrow{\text{Estes'}} \hat{m{\gamma}}_{\textit{EB}^2} = \hat{\pmb{W}}_2 \hat{m{\gamma}}_{\textit{I}} + (\textit{I} - \hat{\pmb{W}}_2) \hat{m{\gamma}}_{\textit{CML}^2}$ 

Step 2:

$$egin{bmatrix} \hat{\gamma}_{EB^1} \ \hat{\gamma}_{EB^2} \end{bmatrix} \sim \textit{N}(\gamma_0, \textit{V}),$$

#### **Optimal Covariance** Weighted Estimate

### Proposed **Estimates**

$$\boldsymbol{\hat{\gamma}}_0 = \textit{Weight}_1 \hat{\gamma}_{\textit{EB}^1} + \textit{Weight}_2 \hat{\gamma}_{\textit{EB}^2}$$

$$\textit{Weight} = \arg\min_{\textit{Weight}} \textit{Var}(X^T \boldsymbol{\hat{\gamma}}_0)$$

$$s.t.$$
  $\begin{cases} \textit{Weight1}^{T} = 1 \\ \textit{Weight} \in (0,1) \end{cases}$ 

#### Selective Coefficient Learner

$$\hat{\gamma}_{X_j}^* = \sum_{k \in E_j} \frac{\hat{V}ar(\boldsymbol{\hat{\gamma}}_{X_j^k})}{\sum_{k \in E_j} \hat{V}ar(\boldsymbol{\hat{\gamma}}_{X_j^k})} \boldsymbol{\hat{\gamma}}_{X_j^k}$$

$$\hat{\boldsymbol{\gamma}}_{\mathrm{SC-Learner}} = [\hat{\gamma}_{\mathrm{X}_{0}}^{*}, \hat{\gamma}_{\mathrm{X}_{1}}^{*}, ..., \hat{\gamma}_{\mathrm{X}_{\mathrm{p}}}^{*}, \hat{\gamma}_{\mathrm{B}}]^{\mathrm{T}}$$

## Proposed Framework

#### Step 1:

External 
$$\hat{\boldsymbol{\beta}}_1 \xrightarrow{\text{Chatterjee's}} \hat{\boldsymbol{\gamma}}_{CML^1} \xrightarrow{\text{Estes'}} \hat{\boldsymbol{\gamma}}_{EB^1} = \hat{W}_1 \hat{\boldsymbol{\gamma}}_I + (I - \hat{W}_1) \hat{\boldsymbol{\gamma}}_{CML^1}$$
  
External  $\hat{\boldsymbol{\beta}}_2 \xrightarrow{\text{Chatterjee's}} \hat{\boldsymbol{\gamma}}_{CML^2} \xrightarrow{\text{Estes'}} \hat{\boldsymbol{\gamma}}_{EB^2} = \hat{W}_2 \hat{\boldsymbol{\gamma}}_I + (I - \hat{W}_2) \hat{\boldsymbol{\gamma}}_{CML^2}$ 

External 
$$\hat{m{eta}}_2 \xrightarrow{ ext{Chatterjee's}} \hat{m{\gamma}}_{ ext{CML}^2} \xrightarrow{ ext{Estes'}} \hat{m{\gamma}}_{ ext{EB}^2} = \hat{\pmb{W}}_2 \hat{m{\gamma}}_I + (\emph{I} - \hat{\pmb{W}}_2) \hat{m{\gamma}}_{ ext{CML}^2}$$

Step 2:

$$egin{bmatrix} \hat{m{\gamma}}_{EB^1} \ \hat{m{\gamma}}_{EB^2} \end{bmatrix} \sim N(m{\gamma}_0, \, V),$$

$$\hat{m{\gamma}}_0 = {\sf A}$$
 weighted average of  $\hat{m{\gamma}}_{\it EB^1}$  and  $\hat{m{\gamma}}_{\it EB^2}$ 

# Simulation/Real Data Analysis Results and Conclusion

X represents the predictors that have been used at lease in one external model; B represents the predictors that are only available in the internal study; grey indicates that the external model did not use this predictor. Simulation Setting Snapshot

Overall performance: $\overline{V}(Xi\hat{\gamma})$ , SSE,												
Covariate-wise performance: variance decrease for single coefficient				X2	X3	X4	В					
I: homogeneous populations with correct external models	Internal		X1	AL	AJ	Ач	-	-				
	External 1											
		External 2										
	External 3											
			X1	X2	X3	X4	В					
II: homogeneous populations with uncertainties in external models	Internal							1				
	External 1 (small sample size)											
	External 2											
	Ext	ernal 3										
III: heterogenous X distribution			X1	X2	X3	X4	В					
	1	Internal										
		External 1										
		External 2 (different B X)										
		External 3										
	2		X1	X2	X3	X4	В					
		Internal										
		External 1										
		External 2										
		External 3 (different Y X, B)										
IV: homogeneous populations with correct external models	1		X1	X2	X3	X4	X5	X6	X7	X8	X9	В
		Internal										
		External 1									1	_
		External 2						_				_
		External 3	371	770	772	37.4	D.	D0	D2	D.4	D.f.	
		T-t1	X1	X2	X3	X4	B1	B2	В3	B4	B5	-
	2	Internal										-
		External 1					-					4
		External 2 External 3										-
Abbreviation: $\overline{V}(Xi\hat{v}) = \frac{1}{2} \sum_{i=1}^{N} \widehat{V}(Xi\hat{v})$												

Abbreviation:  $V(Xi\hat{\gamma}) = \frac{1}{N} \sum_{i=1}^{N} V(Xi\hat{\gamma})$ , estimated mean variance of predicted log odds over observations in the validation dataset;

 $\begin{aligned} &\text{SSE} = \sum_{i=1}^{N} (p_{i0} - \hat{p}_i)^2 \text{, sum of squared error over observations in the validation dataset;} \\ &\text{Scaled Brier score} = \sum_{i=1}^{N} (Y_i - \hat{p}_i)^2 / \sum_{i=1}^{N} (Y_i - \overline{Y})^2 \text{, scaled Brier score over observations in the validation dataset} \end{aligned}$ 

#### Summary of simulation & prostate cancer data analysis

- Incorporating 3 correct external model information
  - Overall: ↓ ~25%
  - Covariate-wise: ↓ ~30% each
- Incorporating 2 correct & 1 incorrect external model information
  - Low-weight for the wrong external model information
  - Overall: ↓ 15-25% depending on dimension
  - o Covariate-wise: ↓ 15-30% each
- Regarding both overall and covariate-wise performance, Optimal Covariate Weighted Estimate (OCWE) and
   Selective Covariate Learner (SC-Learner) are better than
  - Direct regression using internal data only
  - EB estimate solely (one-time shrinkage)
  - o Inverse variance weighted estimate
- OCWE vs. SC-Learner
  - SC-Learner has better covariate-wise performance when the dimension difference of external calculators
     was large (i.e. certain external calculators used more X's than other calculators)

e.g.

	X1	X2	X3	X4	X5	X6	X7	X8	X9	В
Internal										
External 1										
External 2										
External 3										

#### In Conclusion

- The proposed framework is flexible and robust
  - It can incorporate external models that use a slightly different set of covariates
  - It can identify the most relevant external information and diminish the influence of information that is less compatible with the internal data
  - It nicely balances the bias-variance tradeoff while preserving the most efficiency gain
- The 2 proposed estimates--Optimal Covariate Weighted Estimate (OCWE) and Selective Covariate Learner (SC-Learner) are efficient
  - More efficient than the naïve analysis of the internal data and other naïve combinations of external estimators
  - More efficient than incorporating single external information

#### Reference

Chatterjee, N., Chen, Y.-H., Maas, P. & Carroll, R. J. (2016). Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association*, 111(513), 107–117.

Estes, J.P. & Mukherjee, B. & Taylor, J.M.G. (2018). Empirical Bayes estimation and prediction using summary-Level information from external big data sources adjusting for violations of transportability. *Statistics in Biosciences*, 10(3), 568–586.

Gu, T. & Taylor, J.M.G. & Cheng, W. & Mukherjee, B. (2019). Synthetic data method to incorporate external information into a current study. *Canadian Journal of Statistics*, 47(4), 580-603.

Another approach to solve the same genre of problem by converting the external information into the synthetic data (no external model coefficient required)

# Thank you!