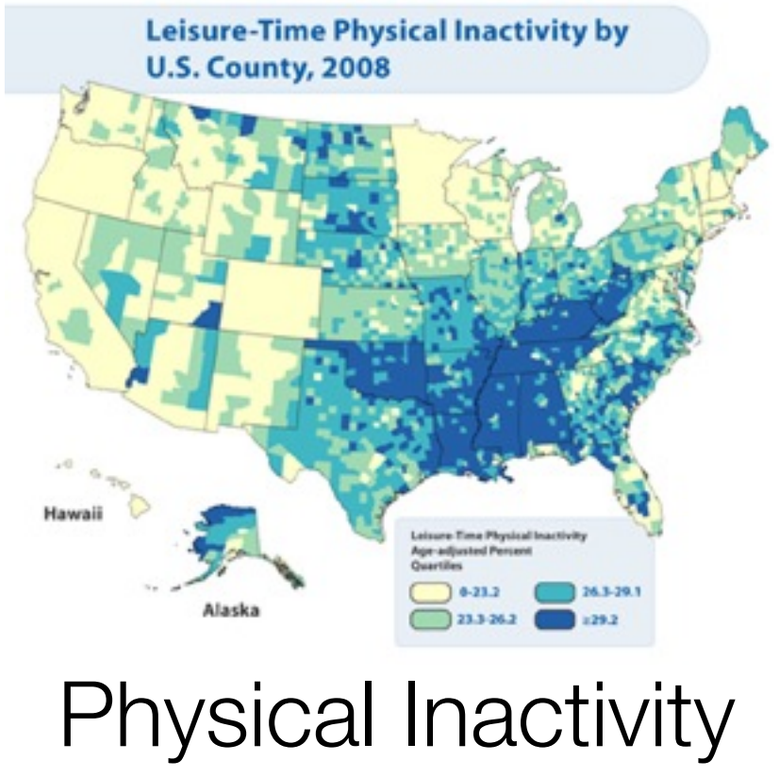
# LUDIA

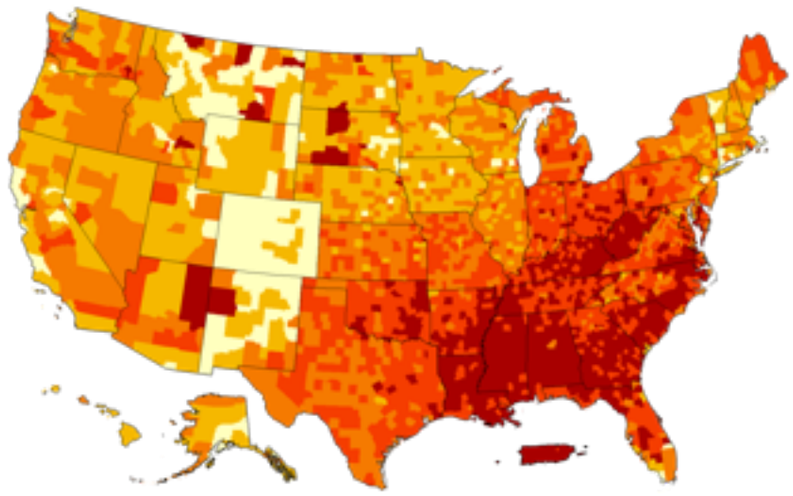
## An Aggregate-Constrained Low-Rank Reconstruction Algorithm to Leverage Publicly Released Health Data

Yubin Park and Joydeep Ghosh

yubin@accordionhealth.com
ghosh@ece.utexas.edu




Physical Inactivity



Diabetes Rate

**What can we say about the relationship between physical inactivity and diabetes at individual-level?**

# Aggregate Data

| State | Avg. Age | Avg. BMI | Diabetes Rate |
|---|---|---|---|
| CA | 54.2 | 125.2 | 0.45 |
| TX | 63.5 | 142.4 | 0.68 |
| FL | 72.7 | 153.2 | 0.72 |
| NY | 56.4 | 134.1 | 0.81 |
| ... | | | |

Can we study the “individual-level” relationship between (Age, BMI) and (Diabetes)?

# Ecological Fallacy

- Results from aggregate data ≠ Results from individual-level data

- Ecological fallacy occurs when aggregate-level statistics are misinterpreted as individual-level inferences

- For example, the high correlation between “per capita consumption of dietary fat” and “breast cancer” in different countries does not imply that dietary fat causes breast cancer (Carroll, 1975)

# Traditional Approaches

- **The neighborhood model**, proposed by Freedman, will imply that diabetes rates are more influenced by geographical attributes rather than the age and BMI variables.

- **Ecological regression**, suggested by Goodman, will assume that the effect strengths of age and BMI are the same across different states, based on the constancy assumption. According to the constancy assumption, geographical partitions are treated as different batches of i.i.d. experiments.

# In fact,

- Both the neighborhood model and ecological regressions are partially true.

- Diabetes rates are influenced by the states “and” the age and BMI variables.

- Thus, a more reasonable (perhaps accurate) model would be:

$$\text{Diabetes} \sim c_{\text{State}} + \beta_{1,\text{State}}\text{Age} + \beta_{2,\text{State}}\text{BMI}$$

# Statistical Underdetermination

- Including the state variable implies that we have more parameters than the number of observations

- For example,
  - 50 observations for 50 states
  - The number of parameters: 150 (intercept, age, and BMI)

# What if…

- If we have an individual-level dataset that maps to the aggregate data, can we leverage multi-level parameter estimation by combining the two datasets?

- Suppose

| State | Avg. Age | Avg. BMI | Diabetes Rate |
|---|---|---|---|
| CA | 54 | 142 | 0.45 |
| | 57 | 102 | |
| | 34 | 98 | |
| | … | | |
| … | | | |

Additional individual-level data

In fact, we can reconstruct the aggregate diabetes rates to individual-level diabetic indicators, which can be used in multi-level modeling.

# LUDIA

- **L**ow-rank factorization **U**sing **DI**fferent levels of **A**ggregation

- Reconstruct the individual-level diabetes variable using:
  - low-rank approximation
  - average statistics e.g. state-level diabetes rate
  - auxiliary individual-level variables e.g. age and BMI

# LUDIA Details

Low-rank approximation

$$\begin{bmatrix}\mathbf{X} & \mathbf{y}\end{bmatrix} = \mathbf{U}\mathbf{V}^\top + \mathbf{E} = \mathbf{U}\begin{bmatrix}\mathbf{V}_x^\top & \mathbf{v}_y^\top\end{bmatrix} + \mathbf{E}$$

Age, BMI  Diabetes

Aggregation Constraint

$$\min_{\mathbf{y},\mathbf{U},\mathbf{V}} \left\| \begin{bmatrix}\mathbf{X} & \mathbf{y}\end{bmatrix} - \mathbf{U}\mathbf{V}^\top \right\|_F^2$$

$$\text{subject to} \quad \mathbf{A}\mathbf{y} = \mathbf{s}$$

We do not observe individual-level "**y**", but only know the aggregate statistics of **y.**

# More details are in Appendix.

# LUDIA Algorithm

Using 1) KKT conditions and 2) relaxation, we obtain:

$$\min_{\mathbf{U},\mathbf{V},\mathbf{\Pi}} \|\mathbf{X} - \mathbf{U}\mathbf{V}_x^\top\|_F^2 + \|\sqrt{\mathbf{W}}(\mathbf{s} - \mathbf{\Pi}\mathbf{v}_y^\top)\|_2^2 + \|\mathbf{A}\mathbf{U} - \mathbf{\Pi}\|_F^2$$

Iterate until convergence:

$$\mathrm{vec}(\mathbf{U}) = (\mathbf{V}_x^\top \mathbf{V}_x \otimes \mathbf{I}_n + \mathbf{I}_r \otimes \mathbf{A}^\top \mathbf{A})^{-1} \mathrm{vec}(\mathbf{X}\mathbf{V}_x + \mathbf{A}^\top \mathbf{\Pi})$$

$$\mathrm{vec}(\mathbf{\Pi}) = (\mathbf{I}_r \otimes \mathbf{I}_p + \mathbf{v}_y^\top \mathbf{v}_y \otimes \mathbf{W})^{-1} \mathrm{vec}(\mathbf{W}\mathbf{s}\mathbf{v}_y + \mathbf{A}\mathbf{U})$$

$$\mathbf{V}_x = (\mathbf{U}^\top \mathbf{U})^{-1} \mathbf{U}^\top \mathbf{X}$$

$$\mathbf{v}_y^\top = (\mathbf{\Pi}^\top \mathbf{W} \mathbf{\Pi})^{-1} \mathbf{\Pi}^\top \mathbf{W}\mathbf{s}$$

# More details are in Appendix.

# Experimental Results

- Dataset: Texas Inpatient Public Use File from Texas Department of State Health Services

- Variables: hospital charges, length of stay, severity of disease, etc.

- We aggregate the “hospital charges” variable at
  - County-, Hospital-, and ZIP-levels.

# Reconstruction Accuracy

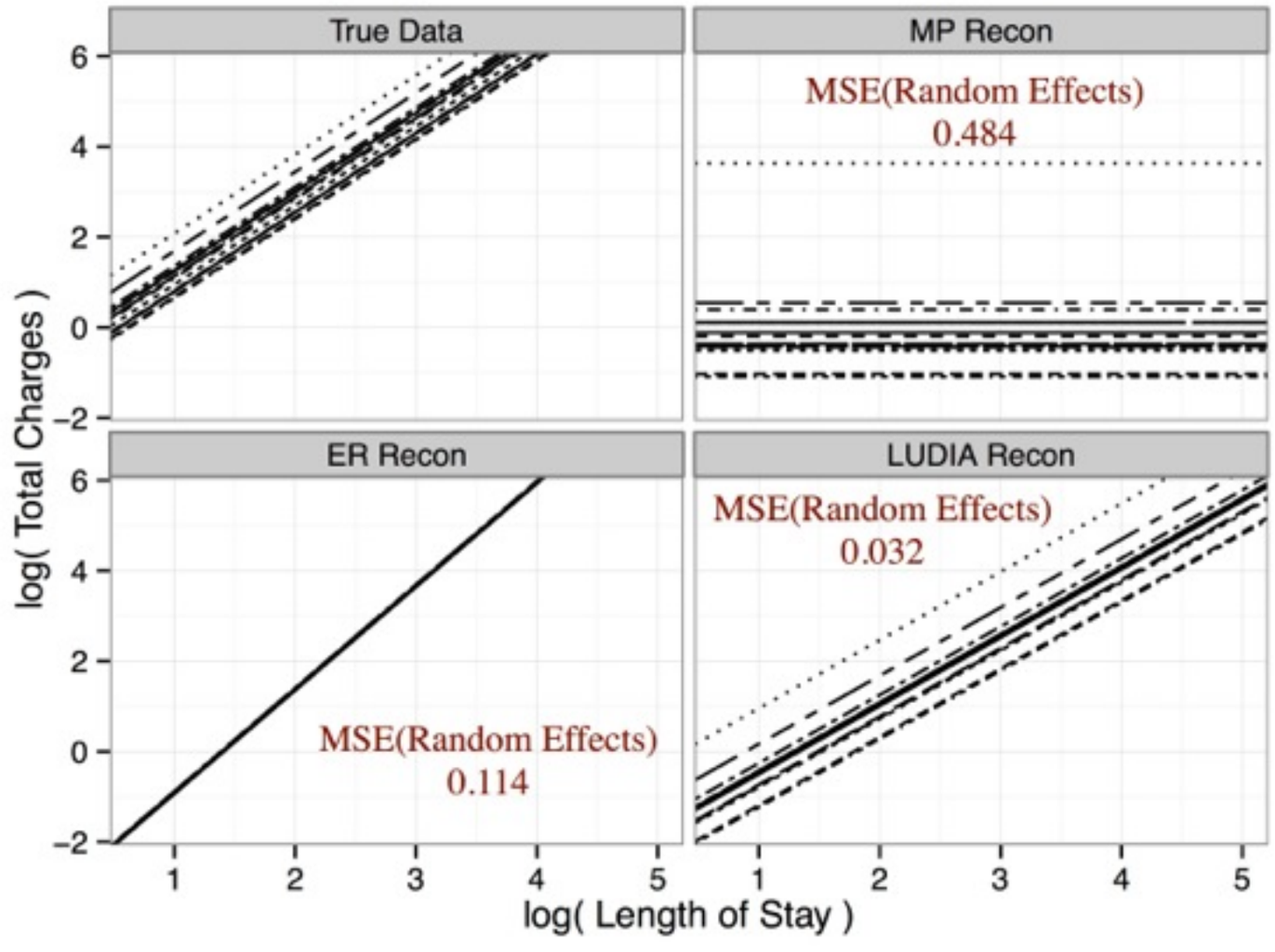
| | Model | MAE | MSE |
|---|---|---|---|
| County-level (not enough to provide accurate reconstruction) | MP | 0.648 (±0.75) | 0.976 (±3.28) |
| | ER | 0.466 (±0.45) | 0.422 (±0.87) |
| | LUDIA | 0.514 (±0.48) | 0.497 (±1.14) |
| Hospital-level | MP | 0.609 (±0.69) | 0.851 (±2.92) |
| | ER | 0.513 (±0.49) | 0.501 (±1.10) |
| | LUDIA | 0.435 (±0.40) | 0.348 (±0.68) |
| ZIP-level | MP | 0.589 (±0.69) | 0.824 (±2.92) |
| | ER | 0.319 (±0.28) | 0.184 (±0.38) |
| | LUDIA | 0.289 (±0.26) | 0.152 (±0.34) |

LUDIA provides the best reconstruction in the hospital- and ZIP-level aggregate settings.

# Multi-level Modeling

$$\log \mathrm{HC} = \log \beta_{\mathrm{hospital}} + \alpha \log \mathrm{LoS} + \ldots + \mathrm{Error}'$$

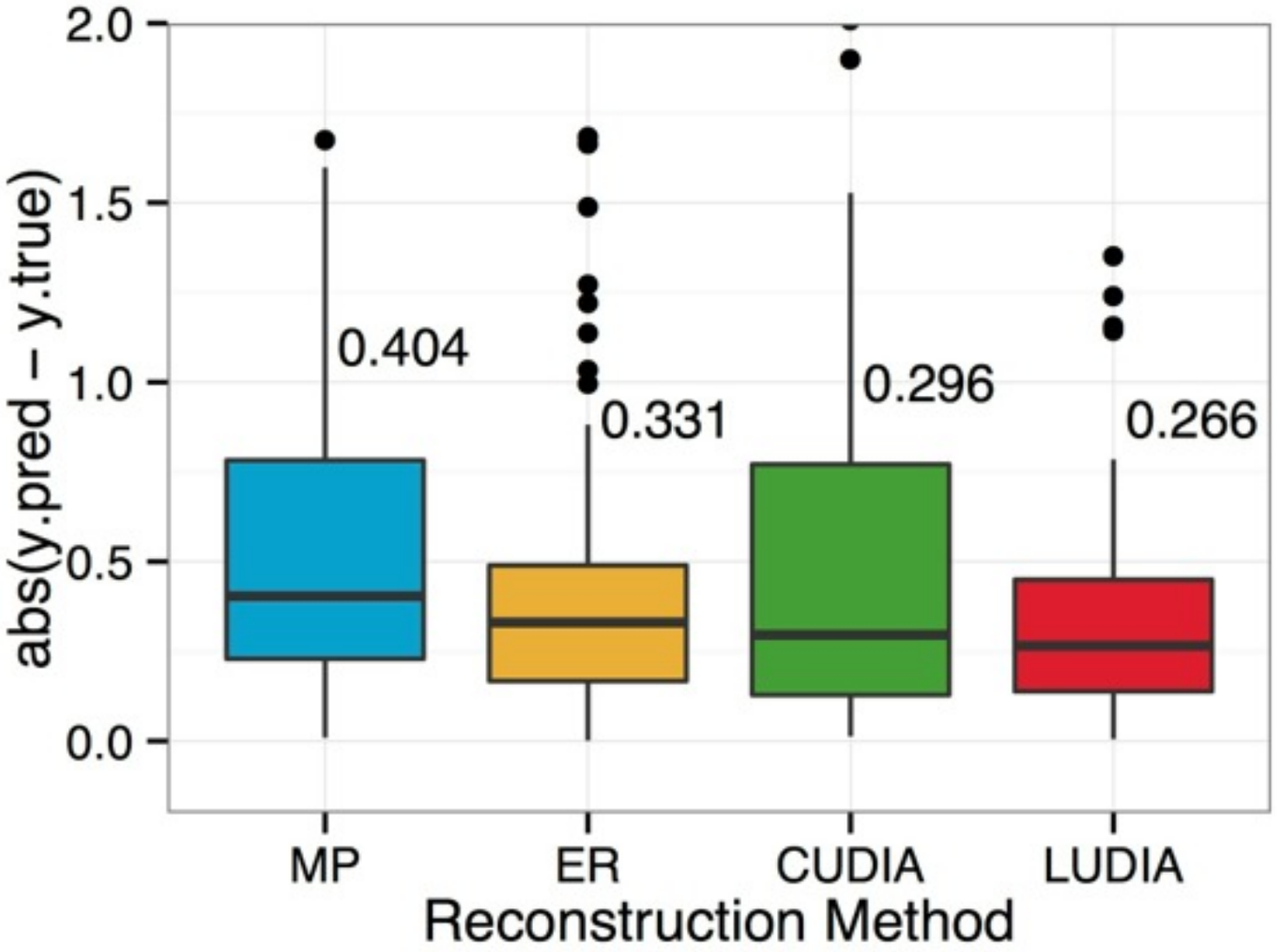
Multi-level coefficients estimated on the original data and the three types of reconstructed data.




# This plot is showing the fitted models, not the data.

LUDIA-reconstructed data provide meaningful multi-level parameters.

# Predictive Performance




Predictive models are trained on the four-types of reconstructed data.

Using LUDIA, you can train a (reasonable) model even if your target variable is aggregated.

# Summary

- When combined with individual-level data, aggregate data can provide informative inferences

- Aggregate data can be reasonably reconstructed when there exists a significant correlation between (already published) individual-level data and aggregate data