AGENCY FOR HEALTHCARE RESEARCH AND QUALITY

# A Machine Learning Approach to SDOH Indices: Optimizing Predictive Power for Health Outcomes


Chandler McClellan, Patricia Keenan, Thomas Selden, Thomas Buchmueller


10/24/2024 – FCSM

# Disclaimer

The views expressed in this presentation are those of the authors, and no official endorsement by the Agency for Healthcare Research and Quality is intended or should be inferred.

# SDOH Indices

- Seek to summarize a given area’s Social Determinants of Health
- Used for policy development, resource allocation, research, community health assessment, risk stratification, etc.
- Examples:
  - ► Area Deprivation Index (ADI) – Mortality w.r.t. Deprivation
  - ► Social Vulnerability Index (SVI) – Disaster response
  - ► Social Deprivation Index (SDI) – Triple Aim (costs, care, health)
  - ► Child Opportunity Index (COI) – Children's development

# Current approaches to indices

- Select Variables
- Reorient/Normalize/Percentiles
- Calculate Index
  - Sum
  - Principal Component Analysis
    - Take first PC as index
      - Linear combination of variables that captures most variation
  - Factor Analysis
    - Linear combination of variables that captures common variation
  - Targeted prediction
    - Predicted value is the index value
- Validate to show index is related to outcomes of interest

# Potential Pitfalls

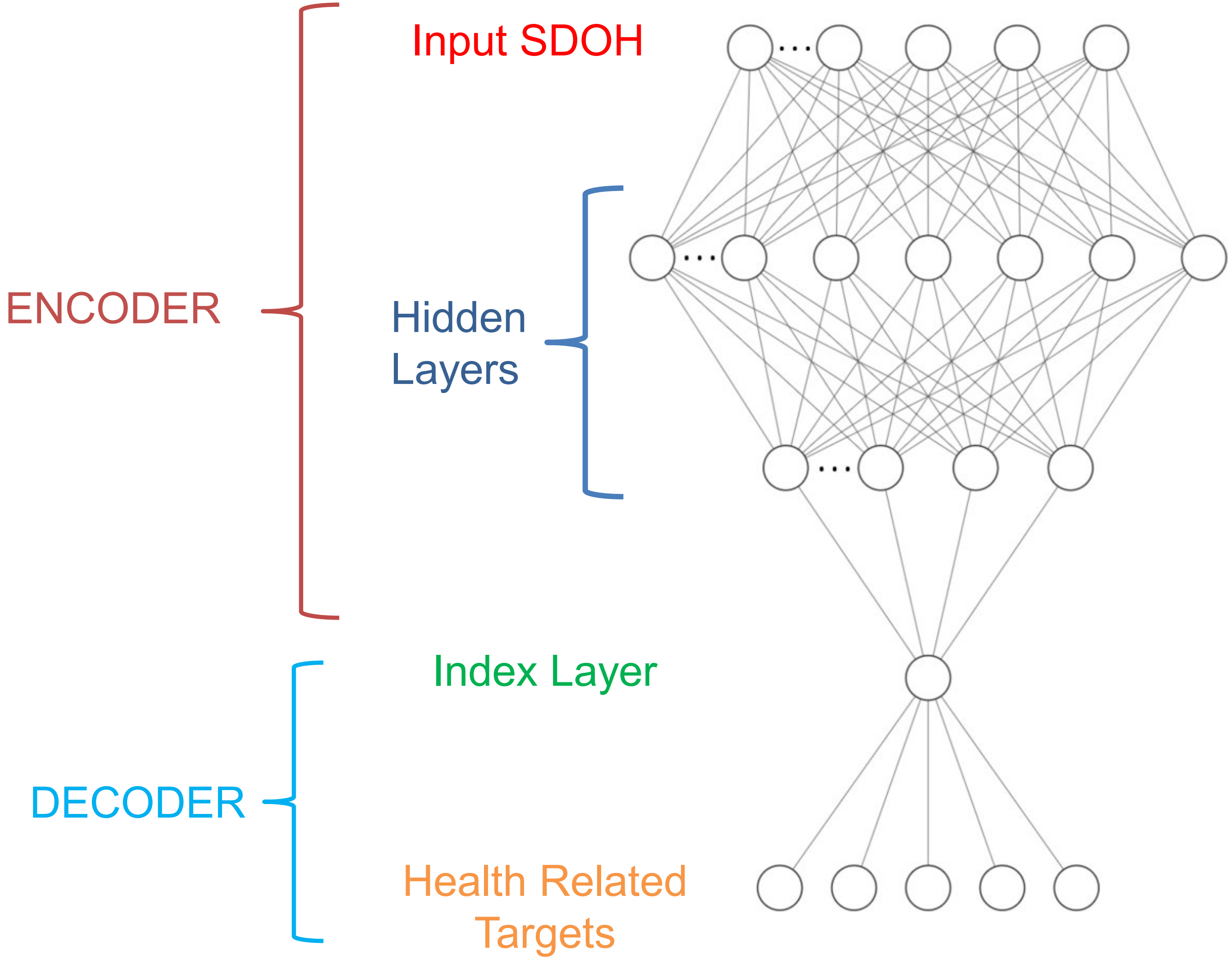
- May leave out important variables (or include unimportant ones)
- Sensitive to data handling choices
- Often linear and may fail to capture important relationships in the data
- May not be related to outcomes of interest
- Often validated on modeled outcomes
  - Area level estimates of outcomes often modeled using SDOH

# Our Approach

- Predict outcomes of interest
  - ► Ensures index is related to outcomes
  - ► Census Tract SDOH used in the ADI (16) – AHRQ SDOH database
  - ► Outcomes: Priority Conditions, Access, Expenditure – MEPS data
- Use a neural network framework to encode SDOH in a single value
  - ► Data driven variable selection
  - ► Captures complex relationships – Universal function approximator
  - ► Multiple outcomes
- Use individual level data to build and validate model

# An Encoding Approach for Indices

- Encoder
  - ▶ Bottlenecks hidden layers to index size.
- Decoder
  - ▶ Uses index to predict outcomes
- Index is constructed vector that explains the most variation in all target outcomes
- Trained as a feed-forward neural network



# SDOH Measures – Census Tract

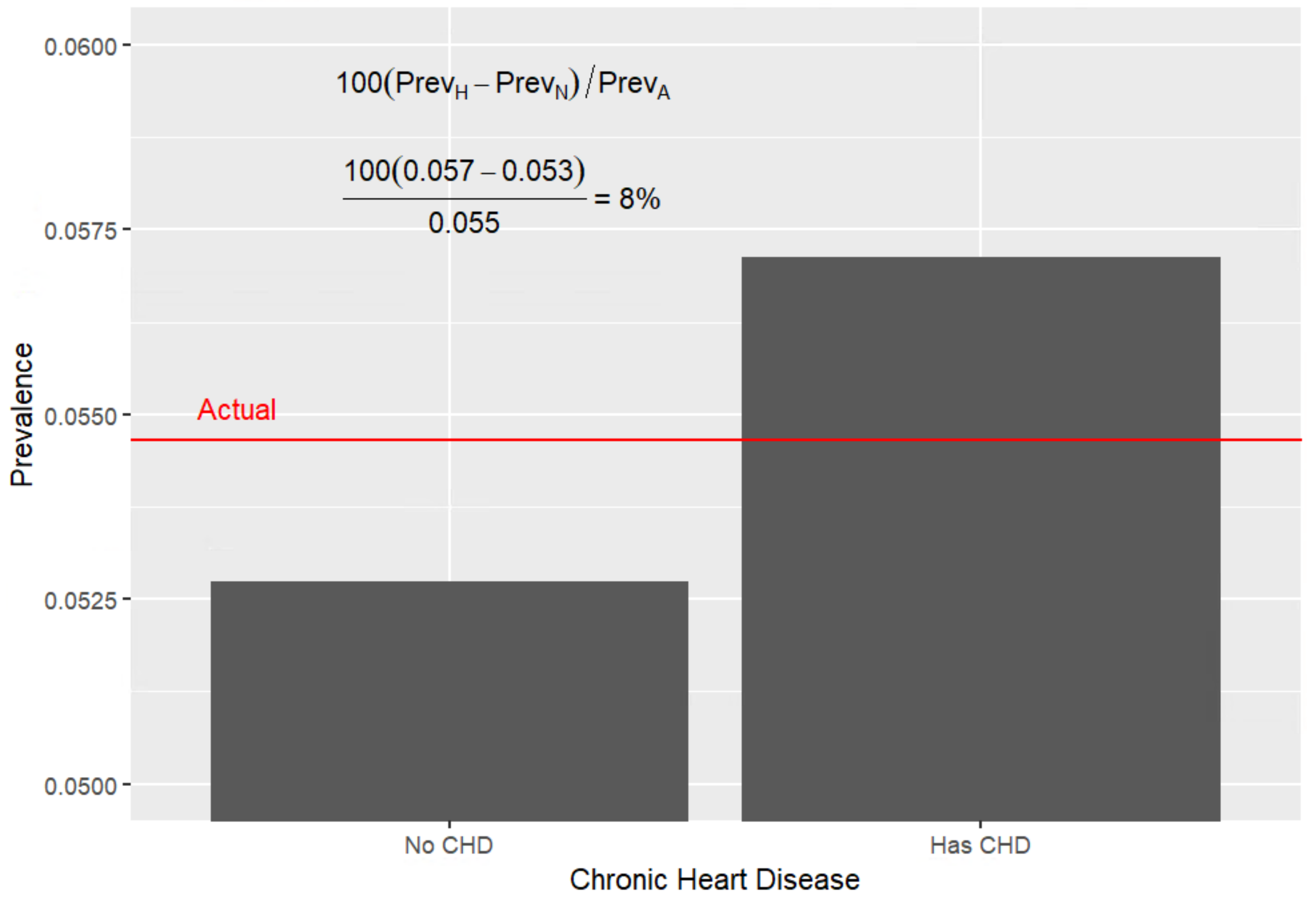
<table>
<tr><th>AHRQ SDOH</th><th>ADI</th></tr>
<tr><td colspan="2">% <100 FPL</td></tr>
<tr><td>% <125 FPL</td><td>% <150% FPL</td></tr>
<tr><td colspan="2">Ln(% <$10k / % >$50k )</td></tr>
<tr><td>% Less than HS</td><td>% less than 9th grade</td></tr>
<tr><td colspan="2">% White Collar</td></tr>
<tr><td colspan="2">% Unemployed</td></tr>
<tr><td colspan="2">% Single Parent</td></tr>
<tr><td colspan="2">% Owner Occupied</td></tr>
<tr><td colspan="2">% Over-crowded</td></tr>
<tr><td colspan="2">% No Vehicle</td></tr>
<tr><td colspan="2">% No Plumbing</td></tr>
<tr><td>%No Computer</td><td>% No Phone</td></tr>
<tr><td colspan="2">Home Value</td></tr>
<tr><td colspan="2">Income</td></tr>
<tr><td colspan="2">Mortgage payment</td></tr>
<tr><td colspan="2">Rent payment</td></tr>
</table>

# MEPS Indices – Individual Level (Census Tracts)

- Targets – Individual Health Related Outcomes
  - Five Priority Conditions: Chronic Heart Disease, High BP, High Cholesterol, Stroke, Diabetes (5)
  - Full set of Priority Conditions (12)
  - Access Measures: No Usual source of care, Cost a barrier to Medical Care/Dental Care/Rx access (4)
  - Total Medical Expenditures (1)
- Architecture
  - "Encoder" – 3 layers: nodes {70,30,1} – with regularization
  - "Decoder" – Linear/logit regression
- Construct Index using 66% of MEPS individuals from 2009-2020
- Hold out 33% for validation and testing
  - No overlapping census tracts
- Compare to ADI
  - Publicly available ADI may not be accurate. Also assess a community deprivation index (CDI)

# Evaluation

- Estimate predictive model in training data
  - ▶ $Y_i = \alpha + \beta * Index_i + \epsilon$
- Predict outcomes in test data
  - ▶ $\widehat{Y_i} = \hat{\alpha} + \hat{\beta} * Index_i$
- Marginalize prediction over actual outcome values.



# Difference in Predictions between those With and Without Condition Relative to Mean – Chronic Condition Targets

| | ADI | CDI | Chronic Condition | Full | Access | Expenditure |
|---|---|---|---|---|---|---|
| Chronic Heart Disease | 3.236 | 0.787 | 8.323 | **8.559** | -0.024 | 2.09 |
| Diabetes | 4.329 | 4.404 | **4.702** | 4.336 | 1.227 | 0.829 |
| High BP | 2.836 | 1.006 | **5.482** | 5.459 | -0.016 | 1.208 |
| High Chol. | 0.077 | 0.108 | **1.916** | 1.737 | 0.373 | 0.642 |
| Stroke | 6.924 | 3.674 | 10.457 | **11.115** | 0.195 | 2.651 |

# Difference in Predictions between those With and Without Condition Relative to Mean – Additional Priority Condition Targets

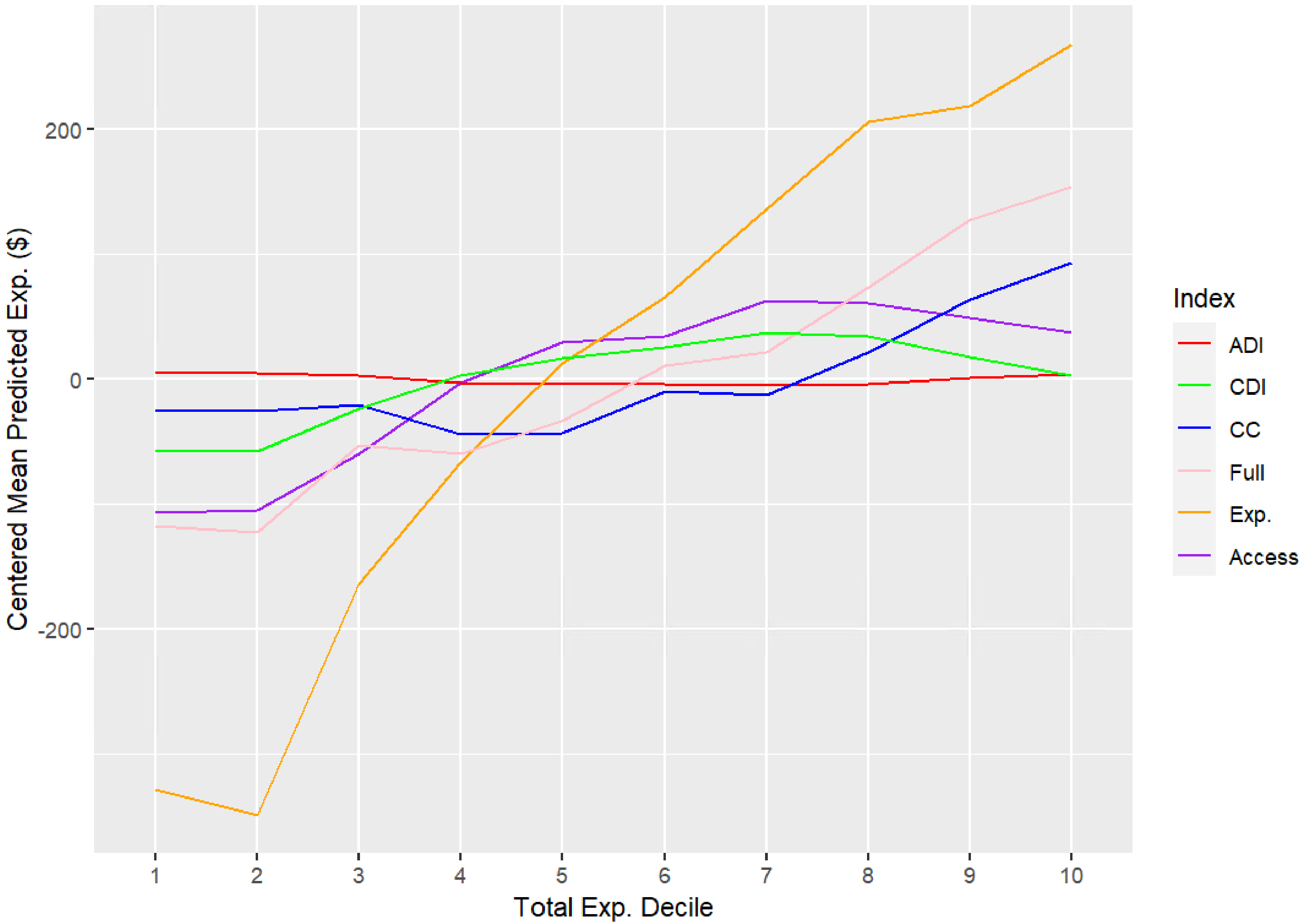
| | ADI | CDI | Chronic Condition | Full | Access | Expenditure |
|---|---|---|---|---|---|---|
| Angina | 5.567 | 1.673 | 11.628 | **15.288** | -0.36 | 2.328 |
| Myocardial Infarction | 8.228 | 2.563 | 12.035 | **14.207** | -0.089 | 3.446 |
| Other Heart | 0.878 | 0.239 | 5.171 | **7.551** | 2.846 | 4.684 |
| Asthma | 0.822 | 0.622 | 1.316 | **2.011** | 0.101 | 4.197 |
| Emphysema | 14.045 | 3.562 | 21.166 | **29.429** | -0.606 | 7.014 |
| Arthritis | 2.943 | 0.209 | 7.520 | **9.278** | 0.803 | 3.892 |
| Cancer | 0.039 | 1.500 | 3.846 | **5.053** | 3.748 | 4.773 |

# Difference in Predictions between those With and Without Relative to Mean – Access Outcomes

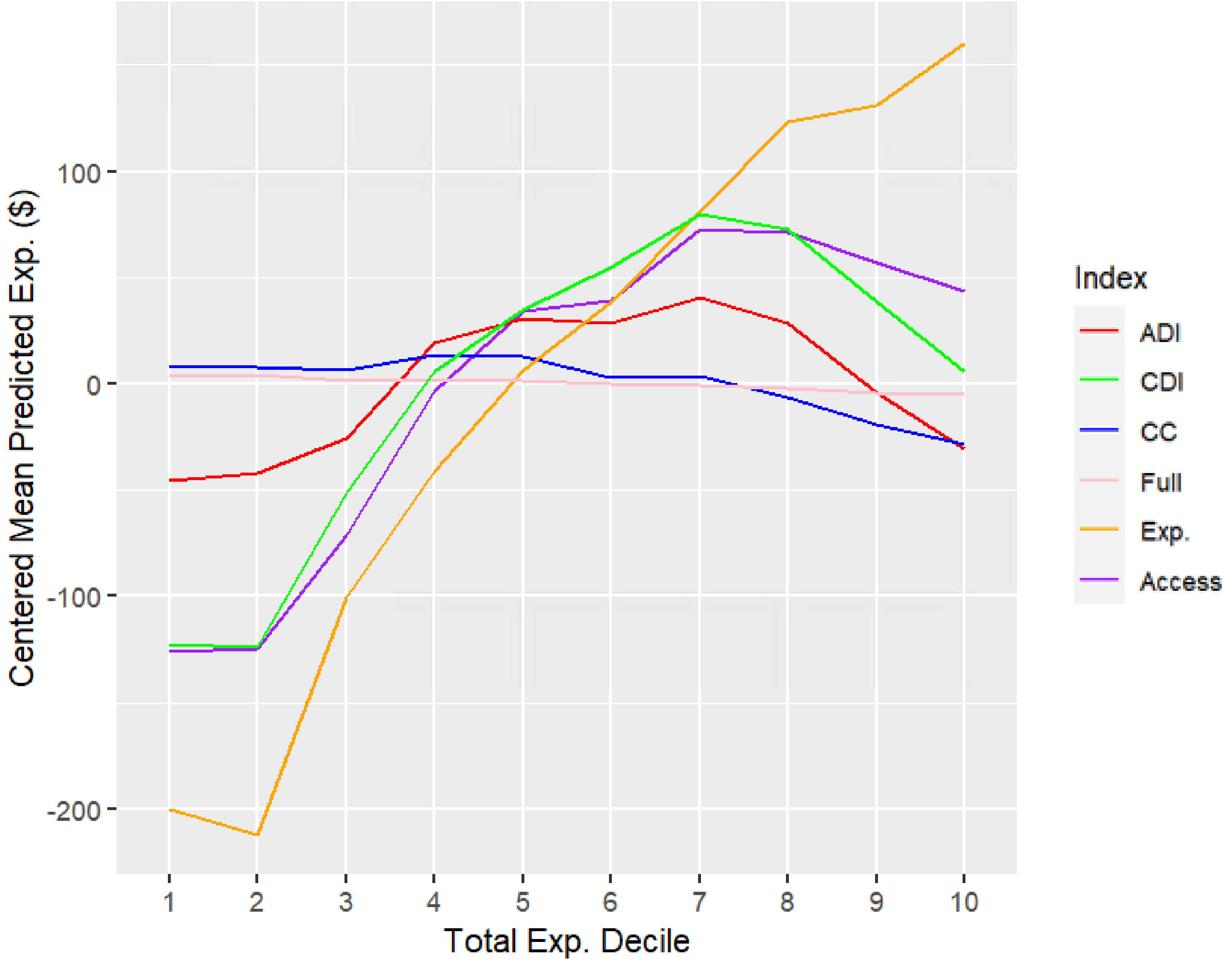
| | ADI | CDI | Chronic Condition | Full | Access | Expenditure |
|---|---|---|---|---|---|---|
| No Usual Source of Care | 0.601 | 2.871 | 0.423 | 1.374 | 2.488 | 2.108 |
| Cost Barrier Medical | 4.577 | 4.026 | 0.329 | 0.108 | 6.121 | 0.751 |
| Cost Barrier Dental | 4.909 | 4.574 | 1.332 | 0.676 | 6.398 | 0.397 |
| Cost Barrier Rx | 10.469 | 10.084 | 4.478 | 2.043 | 8.892 | -0.326 |

*Correlation b/t USC and Medical, Dental, and RX Barriers: 0.05, 0.009, 0.003

# Predicted Mean Expenditure by Actual Expenditure Decile



# Predicted Adjusted Mean Expenditure by Actual Expenditure Decile



# Takeaways

- Neural Network Encoders offer a powerful way to summarize many variables
  - ► Isolates input variation related to outcomes of interest
- Generally better predictions
  - ► Both targeted outcomes and related outcomes
  - ► Prediction of distal outcomes can suffer
  - ► Needs careful consideration about index use
- Tradeoff between prediction power and breadth of targeted outcomes

# Potential Extensions

- Fine-Tuning the encoder
- More encoder input variables
  - Limited for comparison purposes
- More complex “decoder”
  - Non-linearities
  - More index dimensions
- “Decoder” predictors
  - Individual level characteristics

# Thank You

Questions and Comments: chandler.mcclellan@ahrq.hhs.gov