Where are the Health Services Research Opportunities?
A Big Data Analysis with HSRProj

Riyi Qiu\textsuperscript{1}, Yunpeng Zhao\textsuperscript{2}, Yueqi Hu\textsuperscript{3}

\textsuperscript{1}Department of Software and Information Systems, University of North Carolina at Charlotte, Charlotte, NC, USA
\textsuperscript{2}Department of Health Outcomes and Biomedical Informatics, University of Florida, Gainesville, FL, USA
\textsuperscript{3}Department of Computer Science, University of North Carolina at Charlotte, Charlotte, NC, USA

Advisors: Dr. Lixia Yao and Dr. Bian Jiang
Research Question: How to allocate limited medical research resources to address growing patient needs?

The Assumption: Maximal societal benefits can only be achieved when resources are allocated proportional to the disease burden across the full distribution of diseases and conditions

The model: Research Opportunity Index (ROI)
Research Opportunity Index (ROI)

- Measures the degree of (mis)alignment between disease burden and research resources allocation

- Input Variables
  - Disease burden: treatment cost ($b$) from claims data
  - Research focus: number of publication ($p$)
  - Development focus: number of clinical trial ($t$)

- Calculation
  - Normalize every variable: $b' = b / \Sigma b$, where $\Sigma b$ is total treatment cost of all diseases; same calculation for $p'$ & $t'$
  - $ROI = \log_{10}(\frac{b'}{p'} * \frac{b'}{t'})$
Goals

- Examine if disease burden and HSR funding are correlated for each disease
- Calculate the ROI by including HSRProj funding information
- Run topic modeling on HSR project titles and abstracts to understand the topics in those funded projects over time
Figure 1 Flowchart of the analysis on HSRProj data. This figure exhibits the methods, tools, and additional datasets that we used to analyze the HSRProj data.

Abbreviations: CUI – Concept Unique Identifier; PheCode – phenotype code vocabulary; ROI – Research Opportunity Index.
# Correlations between HSRProj Funding and Disease Burden

## Top 5 positively correlated diseases

<table>
<thead>
<tr>
<th>Disease</th>
<th>Pearson Correlation</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acute pancreatitis</td>
<td>0.991</td>
<td>3.954e-10</td>
</tr>
<tr>
<td>Other intestinal obstruction</td>
<td>0.936</td>
<td>7.522e-06</td>
</tr>
<tr>
<td>Encephalitis</td>
<td>0.930</td>
<td>1.187e-05</td>
</tr>
<tr>
<td>Conduct disorders</td>
<td>0.893</td>
<td>9.331e-05</td>
</tr>
<tr>
<td>Other hemoglobinopathies</td>
<td>0.891</td>
<td>9.920e-05</td>
</tr>
</tbody>
</table>

## Top 5 negatively correlated diseases

<table>
<thead>
<tr>
<th>Disease</th>
<th>Pearson Correlation</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blindness and low vision</td>
<td>-0.923</td>
<td>1.814e-05</td>
</tr>
<tr>
<td>Renal failure NOS</td>
<td>-0.910</td>
<td>4.092e-05</td>
</tr>
<tr>
<td>Gonococcal infections</td>
<td>-0.845</td>
<td>5.374e-04</td>
</tr>
<tr>
<td>Chronic hepatitis</td>
<td>-0.837</td>
<td>6.766e-04</td>
</tr>
<tr>
<td>Substance addiction and disorders</td>
<td>-0.803</td>
<td>1.663e-03</td>
</tr>
</tbody>
</table>
Update the ROI

We calculated the ROI by including HSRProj funding \((f)\):

\[
ROI = \log_{10}\left(\frac{b'}{p'} \times \frac{b'}{t'} \times \frac{b'}{f'}\right),
\]

where \(f'\) is the normalized HSRProj funding.
Updated ROI: Top 4 Over-studied Diseases

- Cervical cancer and dysplasia
- Deep vein thrombosis
- Renal failure
- Sepsis and SIRS
Updated ROI: Top 6 Under-studied Diseases

- Calculus of ureter
- Palpitations
- Atherosclerosis of the extremities
- Spondylosis with myelopathy
- Calculus of kidney
- Acne
Topic Modeling

- A statistical model for identifying topical patterns in a large collection of text bodies
- An example

Results

The top 5 topics among the projects related to cervical cancer:

1. Risk factor
2. HPV Infection
3. Cancer control evaluation
4. Mortality
5. Cancer screen
The coverage of the top 5 topics among the projects related to cervical cancer by year
Conclusions

➢ Limitation
  ➢ The assumption may not always be optimal
  ➢ The topic modeling approach cannot identify fine-grained details

➢ Findings
  ➢ We identified the (mis)alignment between disease burden and research resources allocation for 1,337 diseases
  ➢ Under-studied diseases might suggest future research opportunity for the HSR community
Thank you!