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

Riyi Qiu[1], Yunpeng Zhao[2], Yueqi Hu[3]

[1]*Department of Software and Information Systems, University of North Carolina at Charlotte, Charlotte, NC, USA*

[2]*Department of Health Outcomes and Biomedical Informatics, University of Florida, Gainesville, FL, USA*

[3]*Department of Computer Science, University of North Carolina at Charlotte, Charlotte, NC, USA*

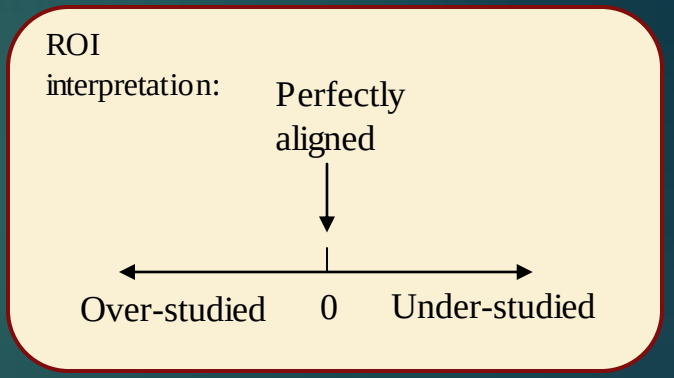
Advisors: Dr. Lixia Yao and Dr. Bian Jiang

Outline

- 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 Σb is total treatment cost of all diseases; same calculation for p' & t'
  - $ROI = \log_{10}(\frac{b'}{p'} * \frac{b'}{t'})$



Yao L, Li Y, Ghosh S, Evans JA, Rzhetsky A. Health ROI as a measure of misalignment of biomedical needs and resources. Nat Biotechnol. 2015 Aug; 33: (8)807-11.

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

Workflow

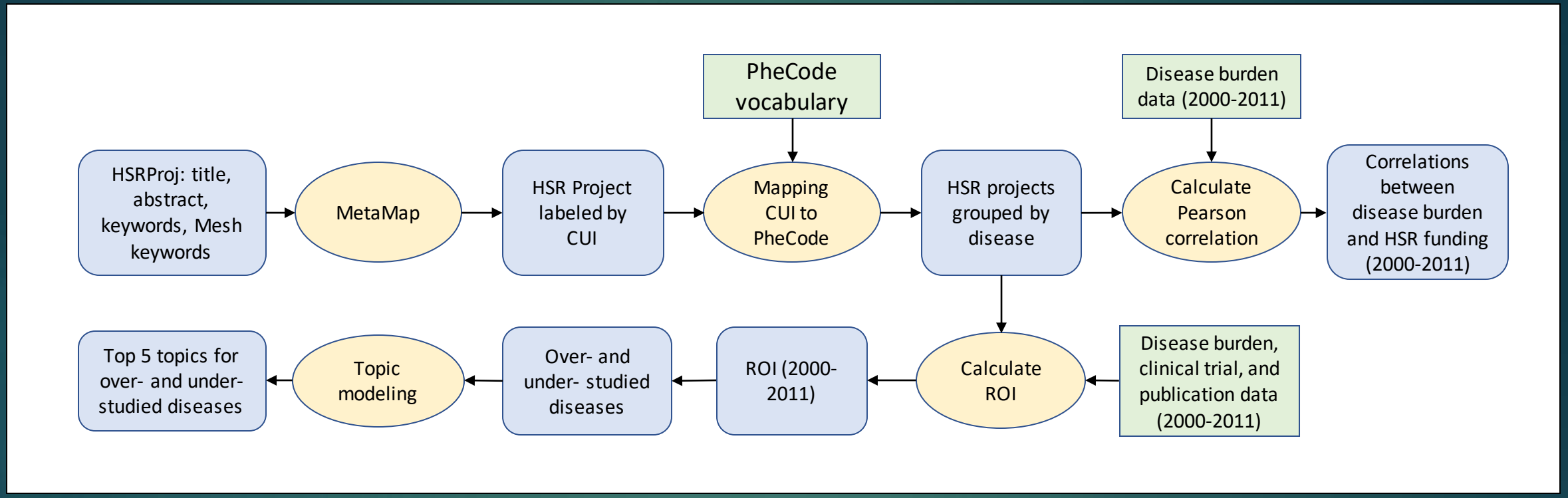


Figure 1 Flowchart of the analysis on HSRProj data. This figure exhibits the methods, tools, and additional datasets that we used to analyze the HSRProj data.
: HSRProj data : Additional data source : Method/Tool
Abbreviations: CUI – Concept Unique Identifier; **PheCode** – phenotype code vocabulary; **ROI** – Research Opportunity Index.

Correlations between HSRProj Funding and Disease Burden

Top 5 positively correlated diseases		
Disease	**Pearson Correlation**	**p value**
Acute pancreatitis	0.991	3.954e-10
Other intestinal obstruction	0.936	7.522e-06
Encephalitis	0.930	1.187e-05
Conduct disorders	0.893	9.331e-05
Other hemoglobinopathies	0.891	9.920e-05

Top 5 negatively correlated diseases		
Disease	**Pearson Correlation**	**p value**
Blindness and low vision	-0.923	1.814e-05
Renal failure NOS	-0.910	4.092e-05
Gonococcal infections	-0.845	5.374e-04
Chronic hepatitis	-0.837	6.766e-04
Substance addiction and disorders	-0.803	1.663e-03

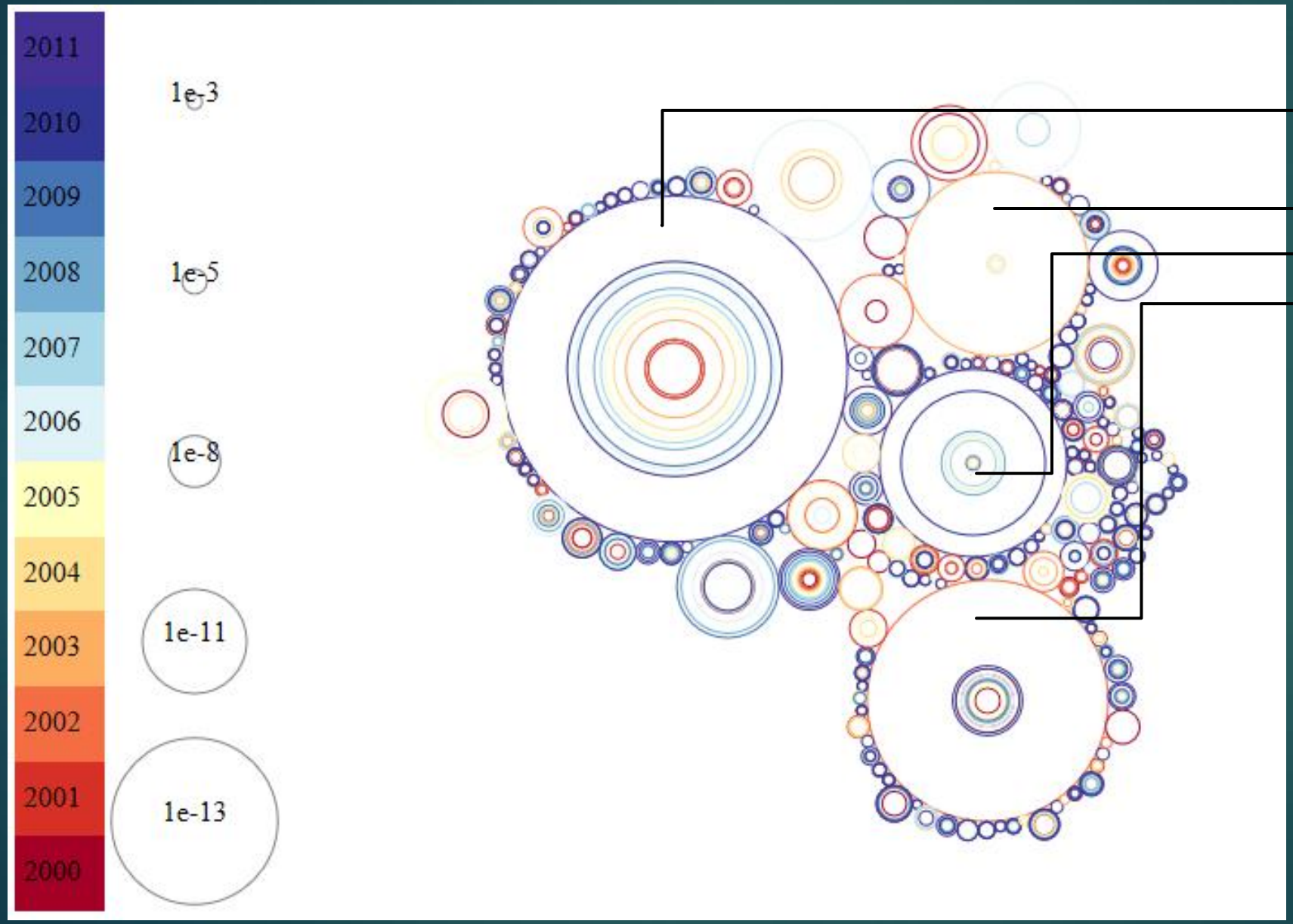
Update the ROI

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

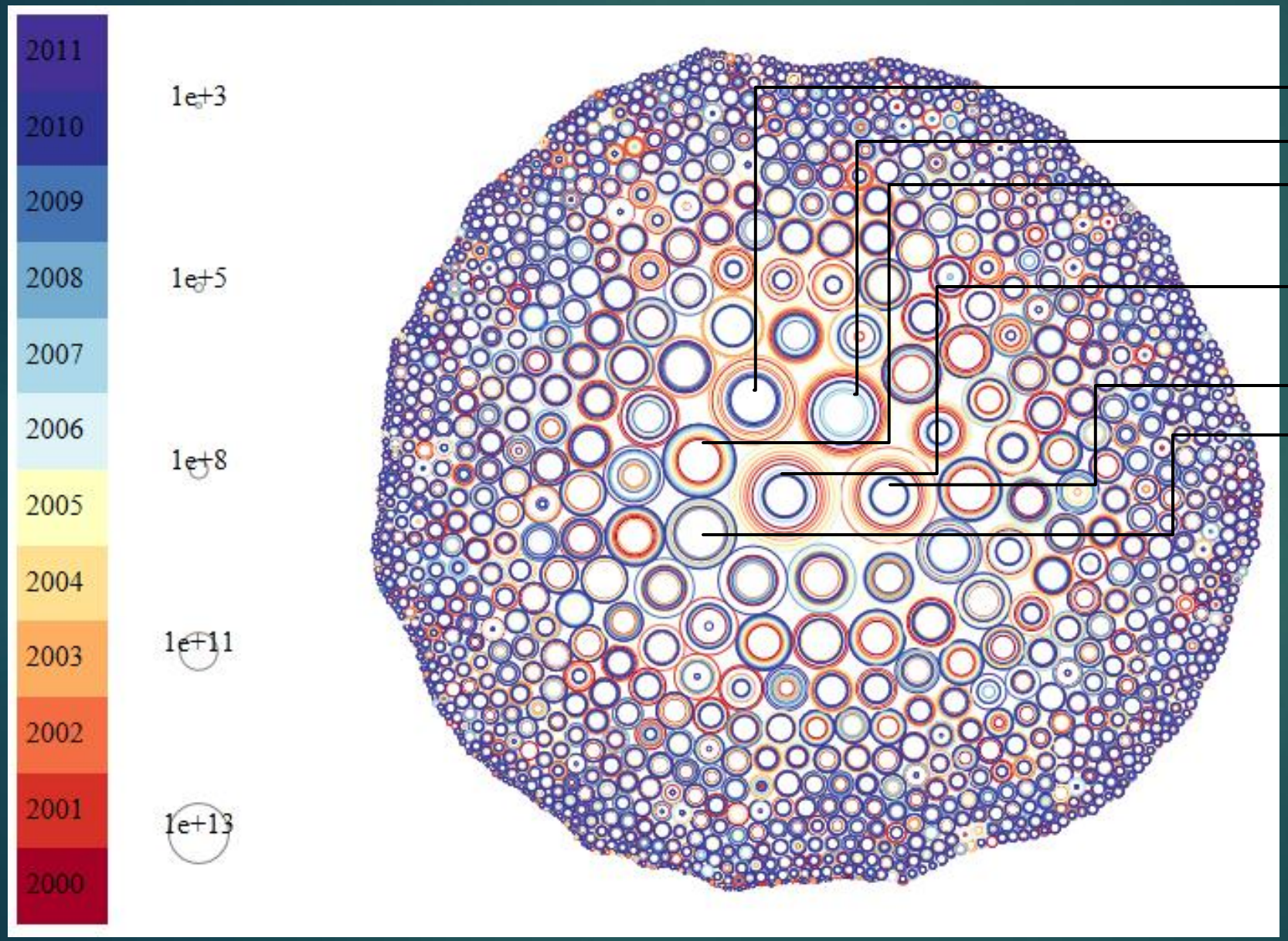
$$ROI = \log_{10}\left(\frac{b'}{p'} * \frac{b'}{t'} * \frac{b'}{f'}\right),$$

where f' is the normalized HSRProj funding

Updated ROI: Top 4 Over-studied Diseases

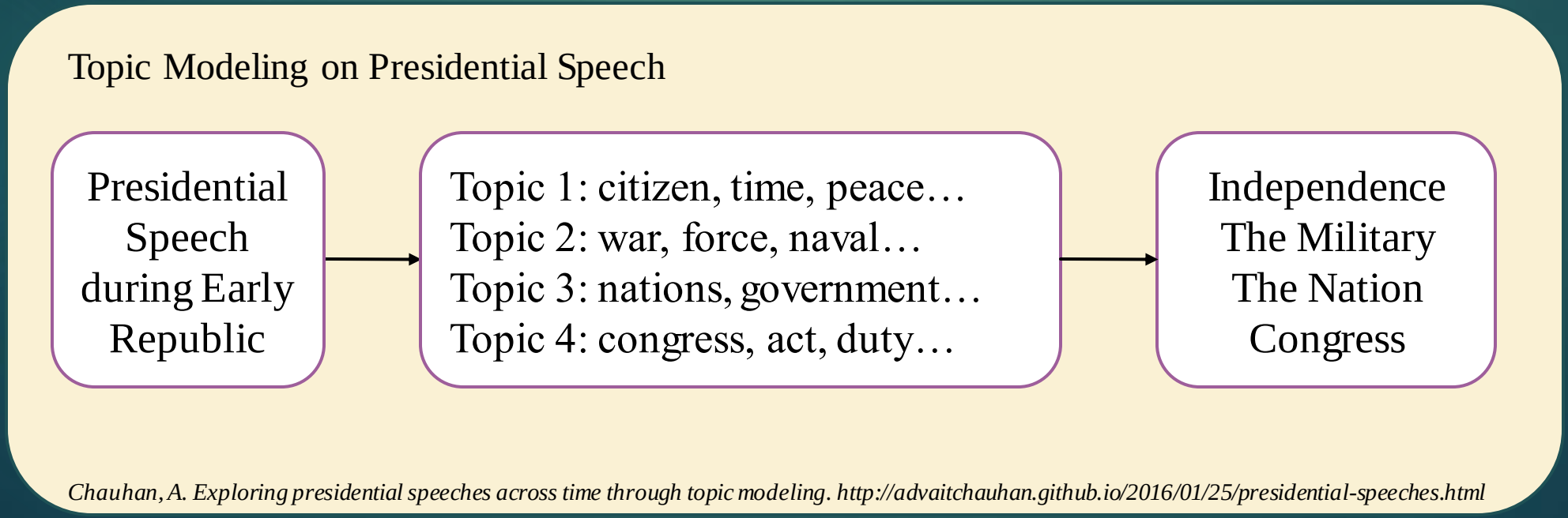


Updated ROI: Top 6 Under-studied Diseases



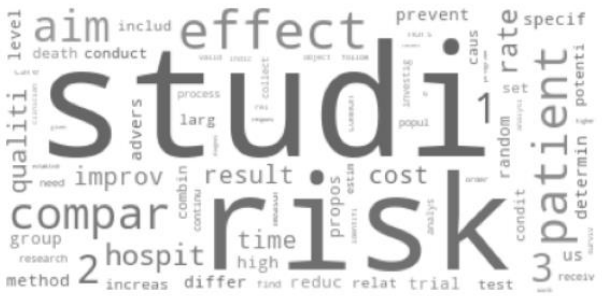
Topic Modeling

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



Chauhan, A. Exploring presidential speeches across time through topic modeling. http://advaitchauhan.github.io/2016/01/25/presidential-speeches.html

Results

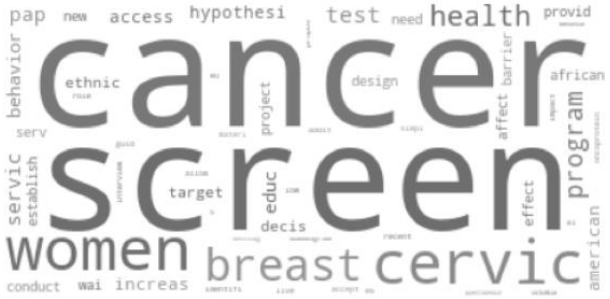


Topic 1: Risk factor

Topic 2: HPV Infection

Topic 3: Cancer control evaluation

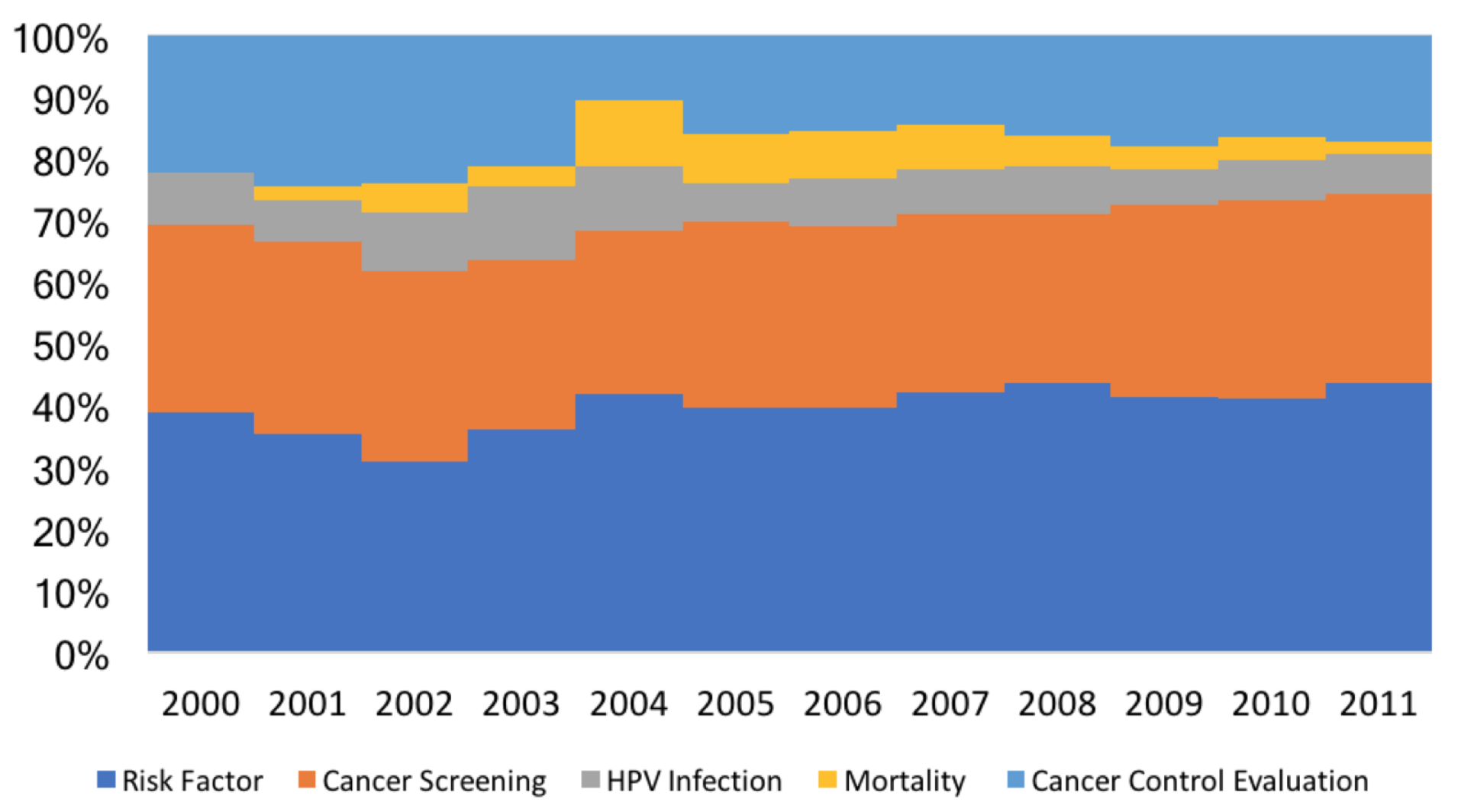
Topic 4: Mortality



Topic 5: Cancer screen

The top 5 topics among the projects related to cervical cancer

Results



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!