Structural Topic Models with Alzheimer's & Mild Cognitive Impairment Patients

Anahita Bahri ● Crystal Qin ● Qian Xu



2017 ALZHEIMER'S DISEASE FACTS AND FIGURES

ALZHEIMER'S DISEASE IS THE
6TH LEADING CAUSE
OF DEATH IN THE UNITED STATES

35% of caregivers for people with Alzheimer's or another dementia report that their health has gotten worse due to care responsibilities, compared to **19%** of caregivers for older people without dementia

1 IN 3
seniors dies with Alzheimer's or another dementia

IT KILLS MORE THAN
breast cancer and prostate cancer
COMBINED

Since 2000, deaths from heart disease have decreased by 14%

while deaths from Alzheimer's disease have increased by 89%

MORE THAN 5 MILLION AMERICANS ARE LIVING WITH ALZHEIMER'S BY 2050, THIS NUMBER COULD RISE AS HIGH AS 16 MILLION

EVERY 66 SECONDS
someone in the United States develops the disease

MORE THAN **15 MILLION AMERICANS** provide unpaid care for people with Alzheimer's or other dementias

IN 2016 these caregivers provided an estimated **18.2 BILLION HOURS** of care valued at over **$230 BILLION**

In 2017, Alzheimer's and other dementias will cost the nation $259 billion
By 2050, these costs could rise as high as
$1.1 TRILLION

Source: The Alzheimer's Association-- 2017 Alzheimer's Disease Facts and Figures

Can we uncover hidden "topics" among Alzheimer's and Mild Cognitive Impairment Patients?

Topic Models

A topic model is a statistical method used for discovering the main themes that occur in a collection of documents. A “topic” consists of a cluster of words that frequently occur together. Topic models have been used to:

- Find groups of products relevant to particular consumers
- Classify images based on small parts of the image (“words”)
- Identify and rank business competitors

What is a “topic” in our case? Since we’re trying to uncover groupings or themes of icd_codes, perhaps we can consider each “topic” as a cluster or group of common diseases.

Probabilistic Topic Models, David Blei

Topics

gene	0.04
dna	0.02
genetic	0.01
. . .	

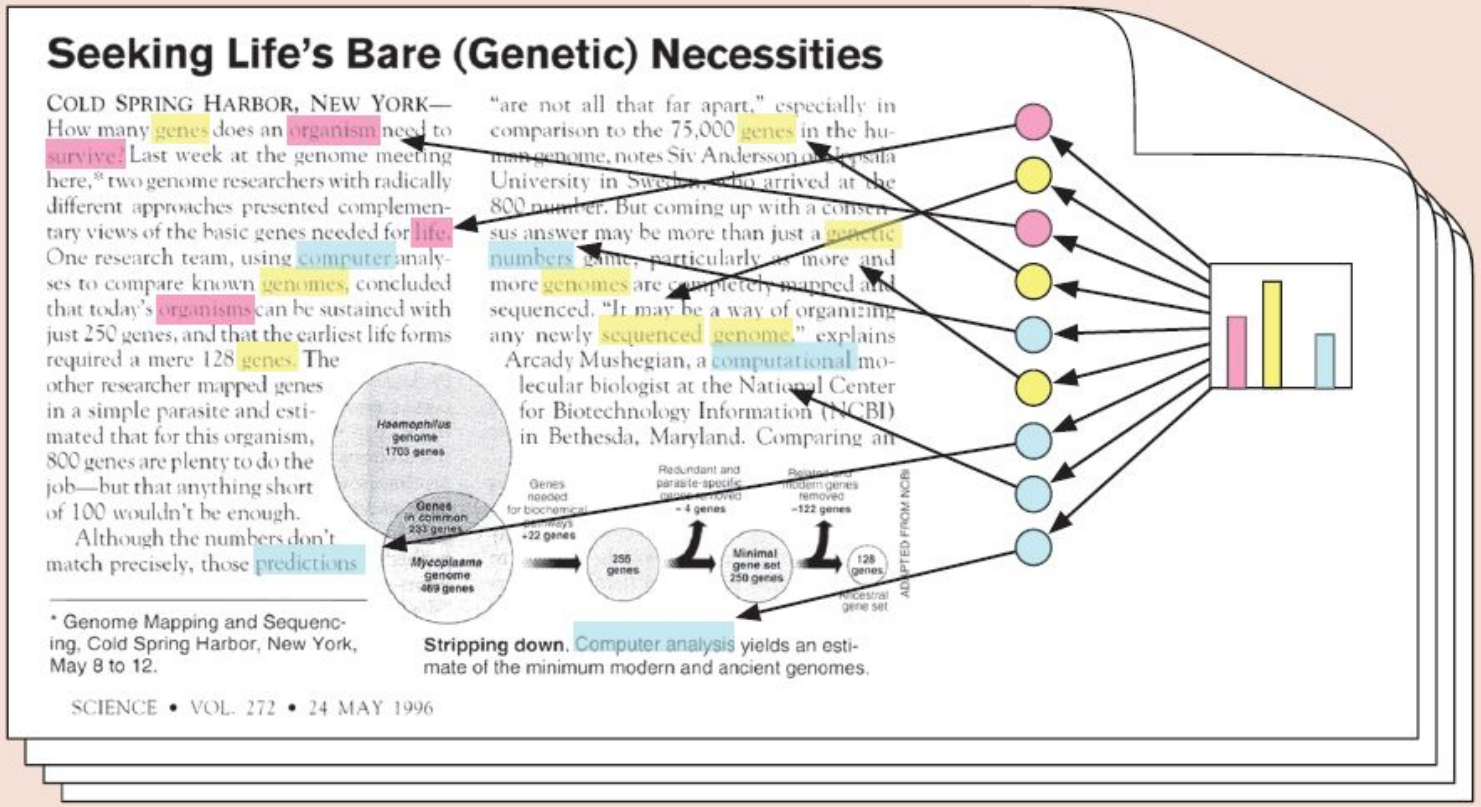
life	0.02
evolve	0.01
organism	0.01
. . .	

brain	0.04
neuron	0.02
nerve	0.01
. . .	

data	0.02
number	0.02
computer	0.01
. . .	

Documents

Topic proportions and assignments



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

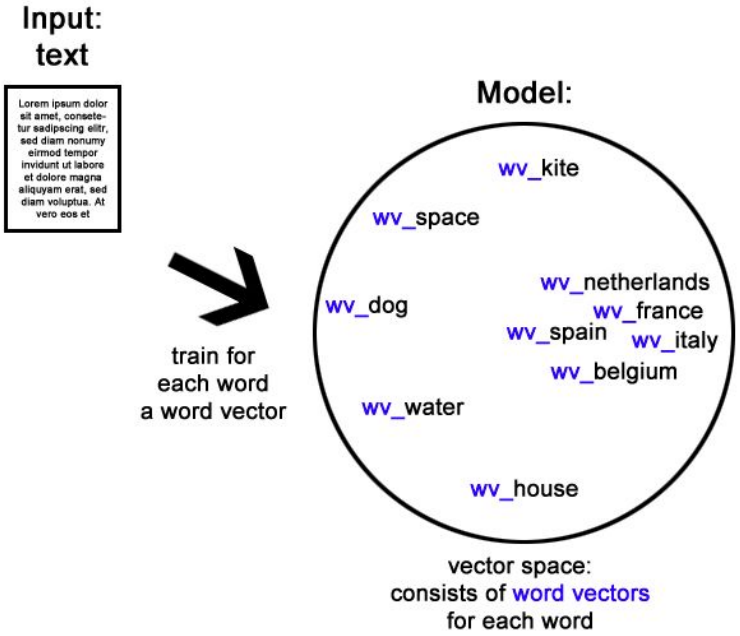
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

Latent Dirichlet Allocation

Exploratory Data Analysis: Word2Vec



most_similar('france'):

spain	0.678515
belgium	0.665923
netherlands	0.652428
italy	0.633130

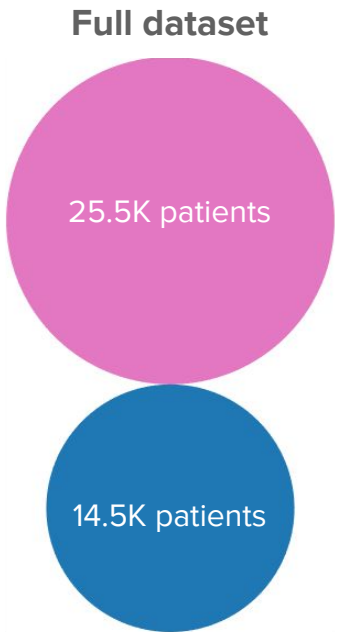
highest cosine distance values in vector space of the nearest words

- Originally released by Google and further improved at Stanford
- Models word-to-word relationships; used for learning vector representations of words, called "word embeddings"
- Similar words tend to be close to each other, but words can have multiple degrees of similarity
- Words returned by Word2Vec are words that co-occur in similar contexts:
  - Synonyms
  - Hyponyms
  - Hypernyms
  - Competitor Names
  - Antonyms

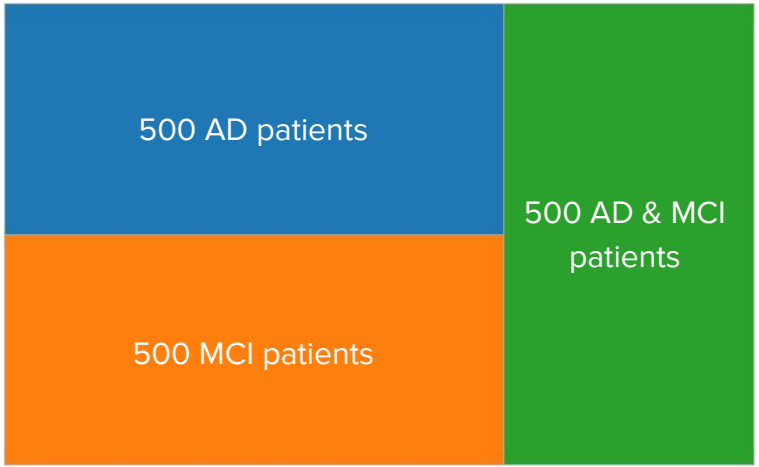
gensim: Deep Learning with Word2Vec

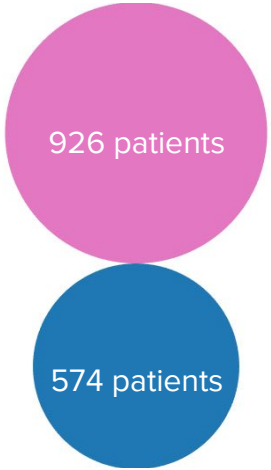
Exploratory Data Analysis: Word2Vec

Alzheimer's	Hypertension	Dementia	Memory	Female
Lyme	Penetration	Emphysema	Edema	Hypotension
Tietze's	Depression	Tremor	Palpitations	Osteoporosis
Pick's	Malnutrition	Sacroiliitis	Aphasia	Hypoglycemia
Cytomegaloviral	Obesity	Lymphedema	Dysuria	Atherosclerosis
Reiter's	Bronchitis/Emphysema	Morbidity/mortality	Nocturia	Neutropenia
Peyronie's	Gout	Encephalopathy	Dysphonia	Quadriplegia
Hirschsprung's	Glaucomatous	Psoriasis	Anorexia	Esophagitis
Schilder's	Measles	Mnfst	Hallucinations	Urticaria
Takayasu's	Multiphasic	Slip/trip/stmble	Wheezing	Hyperparathyroidism
Ritter's	Galactosemia	Myopathies	Heartburn	Orchitis/epididymitis



Data & Representation

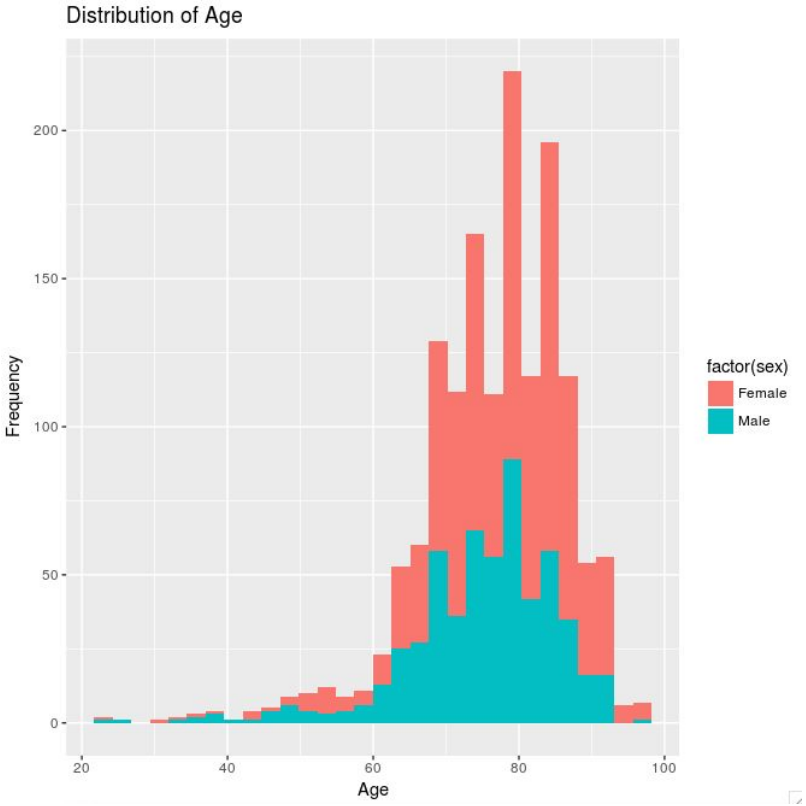




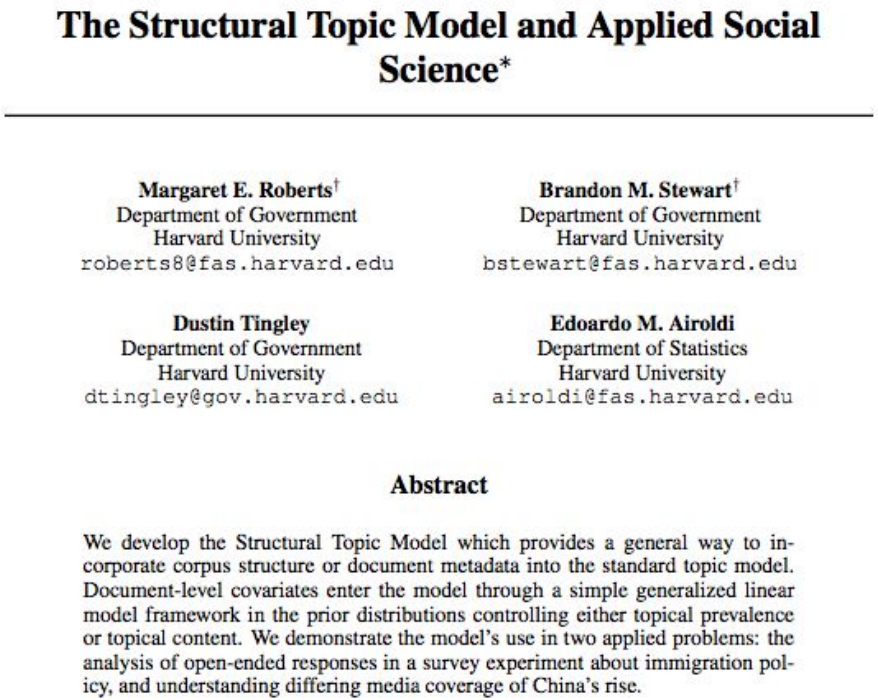
Data format: matrix with frequency of icd_code, in addition to gender and age information, with each row representing a unique patient. No sequence, but can take this into account.

1	2602	2637	2651	2652	2653	2675	2696	2697	2698	2700	2731	2732	2734	2742	2743	2751	2763	2795	2798
6	2	1	1	55	17	1	5	112	1	3	1	1	50	12	3	1	5	1	76

Patient Demographic Visualizations



Structural Topic Model



The Structural Topic Model and Applied Social Science*

Margaret E. Roberts†
Department of Government
Harvard University
roberts8@fas.harvard.edu

Brandon M. Stewart†
Department of Government
Harvard University
bstewart@fas.harvard.edu

Dustin Tingley
Department of Government
Harvard University
dtingley@gov.harvard.edu

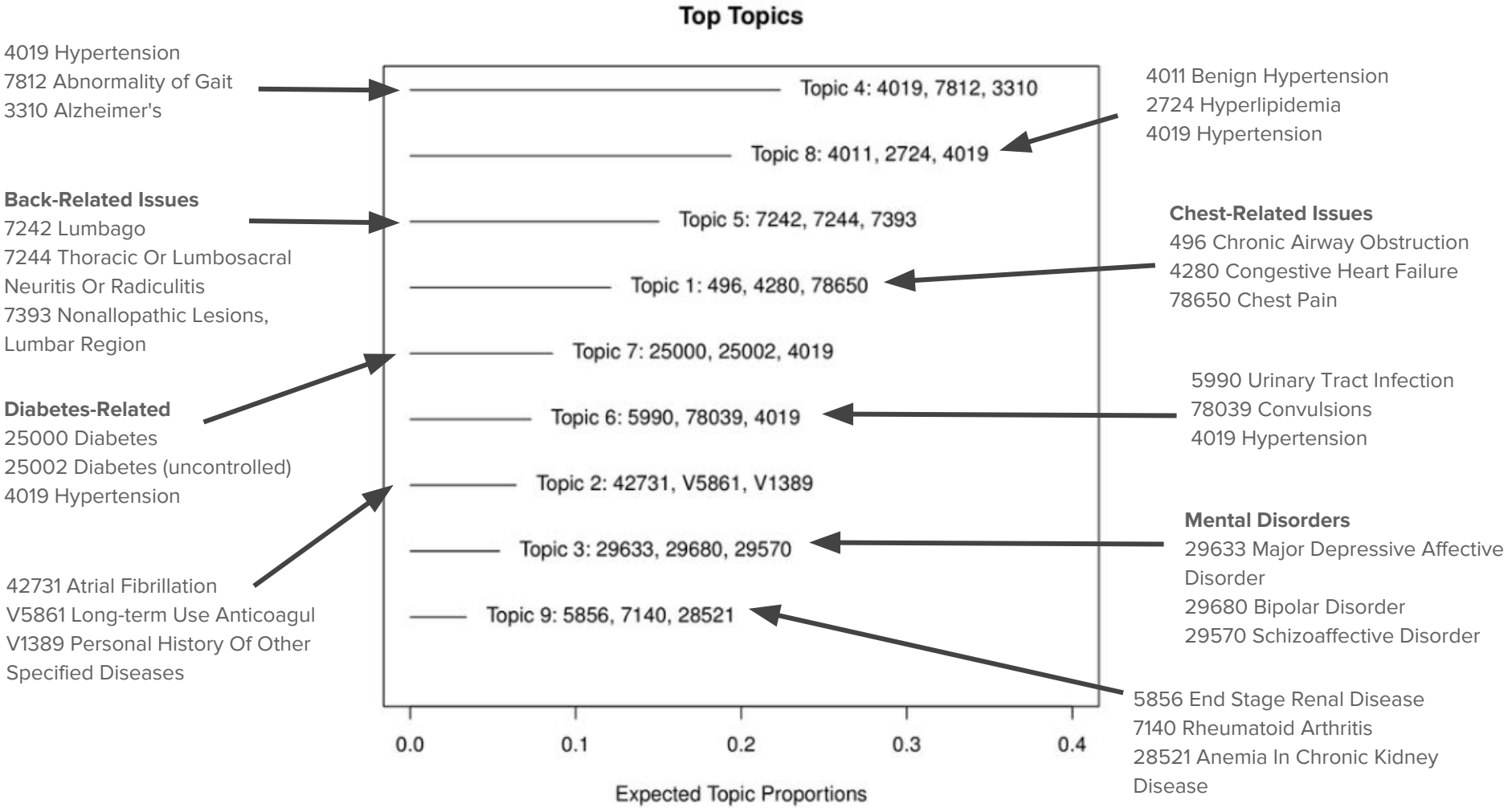
Edoardo M. Airoldi
Department of Statistics
Harvard University
airoldi@fas.harvard.edu

Abstract

We develop the Structural Topic Model which provides a general way to incorporate corpus structure or document metadata into the standard topic model. Document-level covariates enter the model through a simple generalized linear model framework in the prior distributions controlling either topical prevalence or topical content. We demonstrate the model's use in two applied problems: the analysis of open-ended responses in a survey experiment about immigration policy, and understanding differing media coverage of China's rise.

- Unlike other topic models (LDA), STM considers covariates, which can improve inference and qualitative interpretability
  - Simultaneously estimates prevalence of diseases controlling for covariates
- **STM**: General framework for topic modeling with document-level covariate information
  - We treated each patient icd_code history as a document, each icd_code as a word
  - Covariates included gender, age, diagnosis of Alzheimer's disease or not
  - searchK() function recommends the optimal number of topics for the data at hand

"The Structural Topic Model and Applied Social Science", Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, Edoardo M. Airoldi


Top Topics
4019 Hypertension
7812 Abnormality of Gait
3310 Alzheimer's
4011 Benign Hypertension
2724 Hyperlipidemia
4019 Hypertension
Back-Related Issues
7242 Lumbago
7244 Thoracic Or Lumbosacral Neuritis Or Radiculitis
7393 Nonallopathic Lesions, Lumbar Region
Chest-Related Issues
496 Chronic Airway Obstruction
4280 Congestive Heart Failure
78650 Chest Pain
Diabetes-Related
25000 Diabetes
25002 Diabetes (uncontrolled)
4019 Hypertension
5990 Urinary Tract Infection
78039 Convulsions
4019 Hypertension
42731 Atrial Fibrillation
V5861 Long-term Use Anticoagul
V1389 Personal History Of Other Specified Diseases
Mental Disorders
29633 Major Depressive Affective Disorder
29680 Bipolar Disorder
29570 Schizoaffective Disorder
5856 End Stage Renal Disease
7140 Rheumatoid Arthritis
28521 Anemia In Chronic Kidney Disease
Topic 4: 4019, 7812, 3310
Topic 8: 4011, 2724, 4019
Topic 5: 7242, 7244, 7393
Topic 1: 496, 4280, 78650
Topic 7: 25000, 25002, 4019
Topic 6: 5990, 78039, 4019
Topic 2: 42731, V5861, V1389
Topic 3: 29633, 29680, 29570
Topic 9: 5856, 7140, 28521
0.0
0.1
0.2
0.3
0.4
Expected Topic Proportions

STM: Top Topics Among Different Patient Groups

AD Only

Mental Disorders: Major Depressive Affective Disorder, Bipolar Disorder, Anxiety State

Cerebrovascular Diseases: Acute But Ill-Defined, Other and Ill-Defined

Kidney-Related Diseases: End Stage Renal Disease, Encounter For Extracorporeal Dialysis

MCI Only

Brain & Vision Related: Cerebral Artery Occlusion, Exudative Senile Macular Degeneration

Chest-Related Issues: Chest Pain, Chronic Airway Obstruction, Obstructive Chronic Bronchitis

Cancer: Other Malignant Lymphomas--Extranodal And Solid Organ Sites, Multiple Myeloma (no mention of remission)

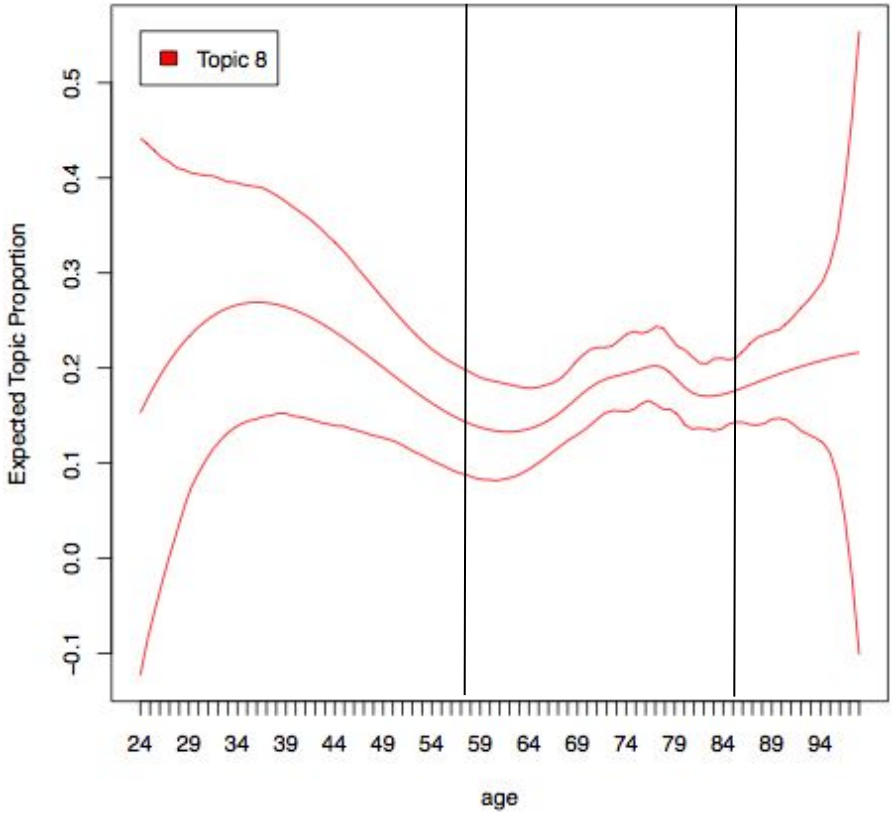
AD & MCI

Chest-Related Issues: Chronic Airway Obstruction, Shortness of Breath

Coronary Atherosclerosis: Native Coronary Artery, Unspecified Type Of Vessel, Native Or Graft

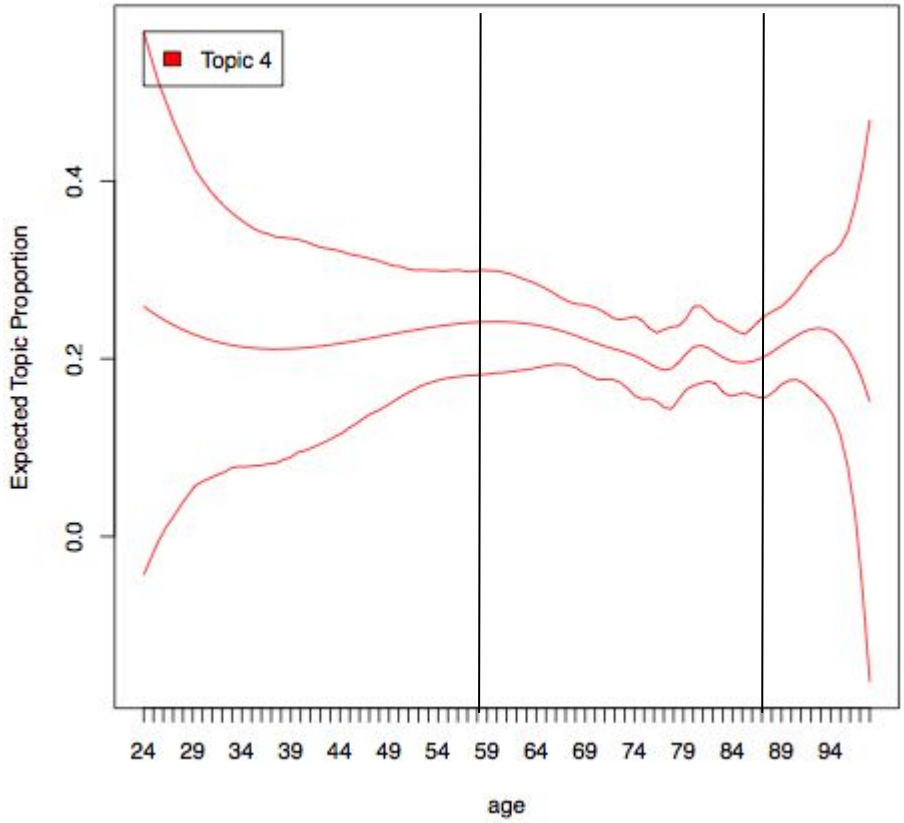
Movement-Related Issues: Abnormality of Gait, Paralysis Agitans (Parkinson's)

Age Trend of Topic 8: HT, HLD, HCL



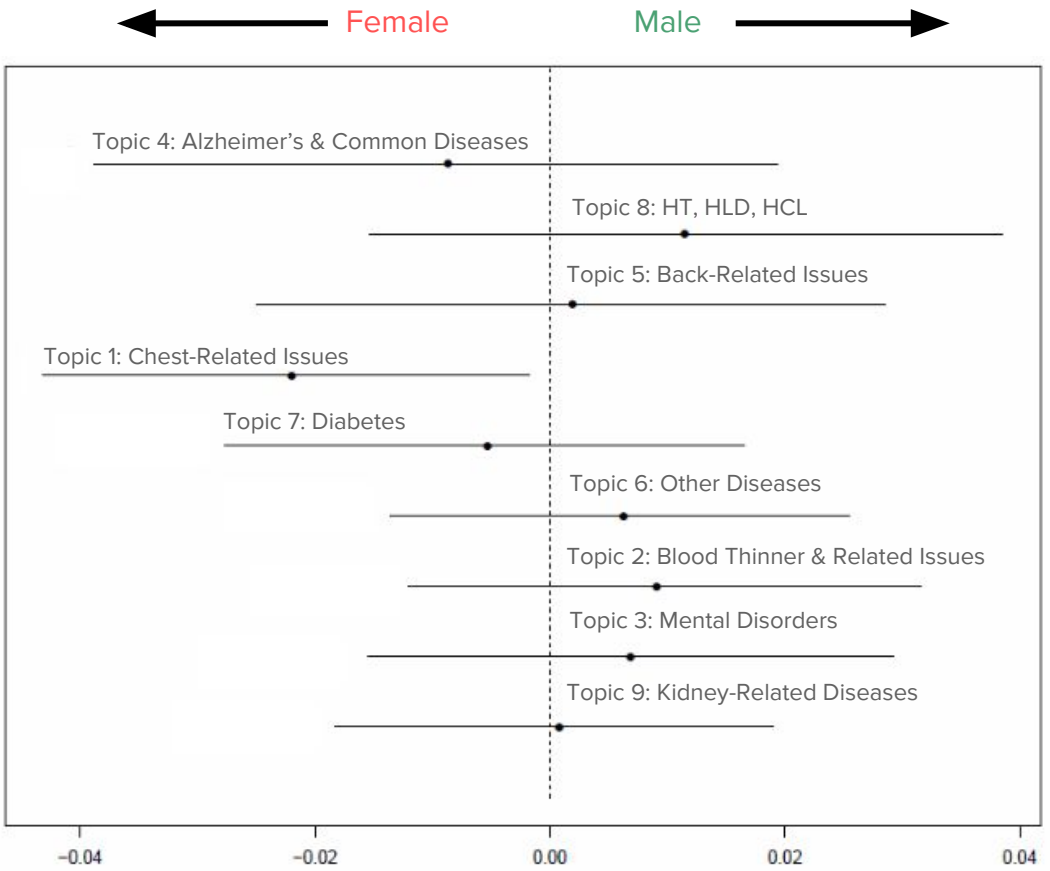
- 4011 Benign Hypertension
- 4019 Hypertension
- 2724 Hyperlipidemia
- 7020 Actinic Keratosis
- 2720 Pure Hypercholesterolemia
- V0481 Prophylactic Vaccination and Inoculation Against Influenza
- 36252 Exudative Senile Macular Degeneration

Age Trend of Topic 4: Alzheimer’s & Common Diseases



- 4019 Hypertension
- 7812 Abnormality of Gait
- 3310 Alzheimer's
- 1110 Pityriasis Versicolor
- 5990 Urinary Tract Infection
- 4011 Benign Hypertension
- 2449 Unspecified Acquired Hypothyroidism

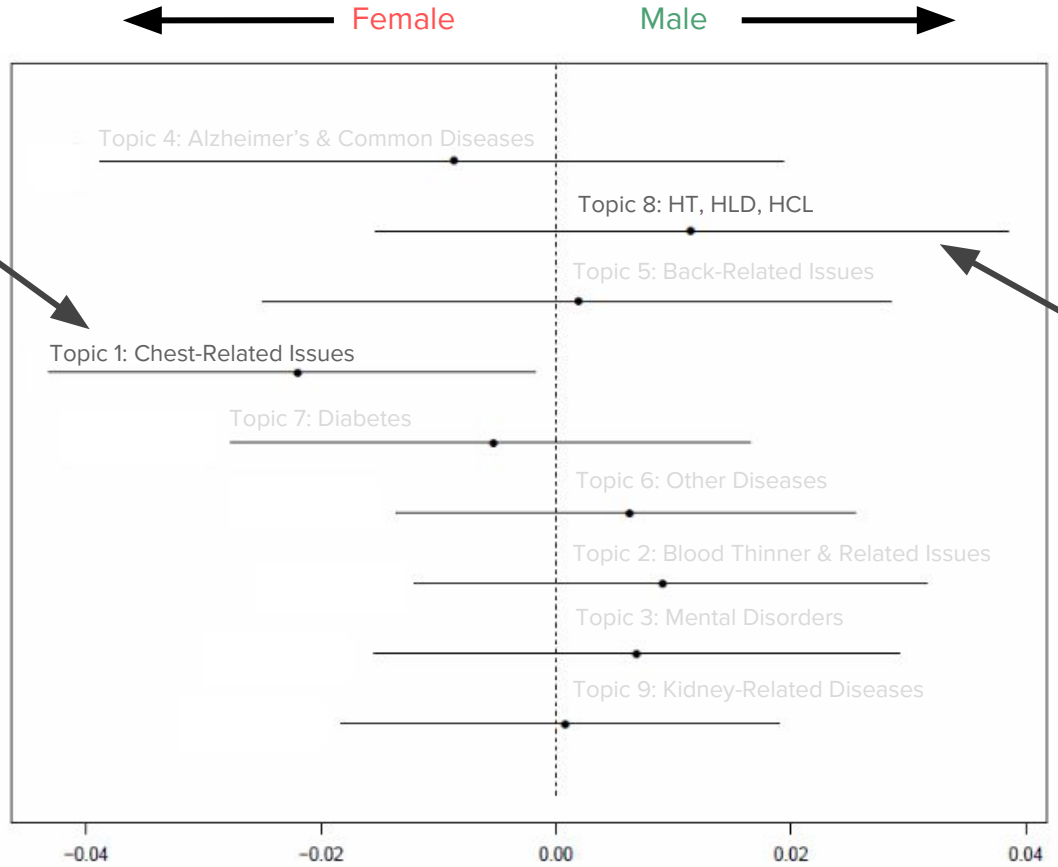
Topic Proportion Comparison: Gender



Topic Proportion Comparison: Gender

Chest-Related Issues

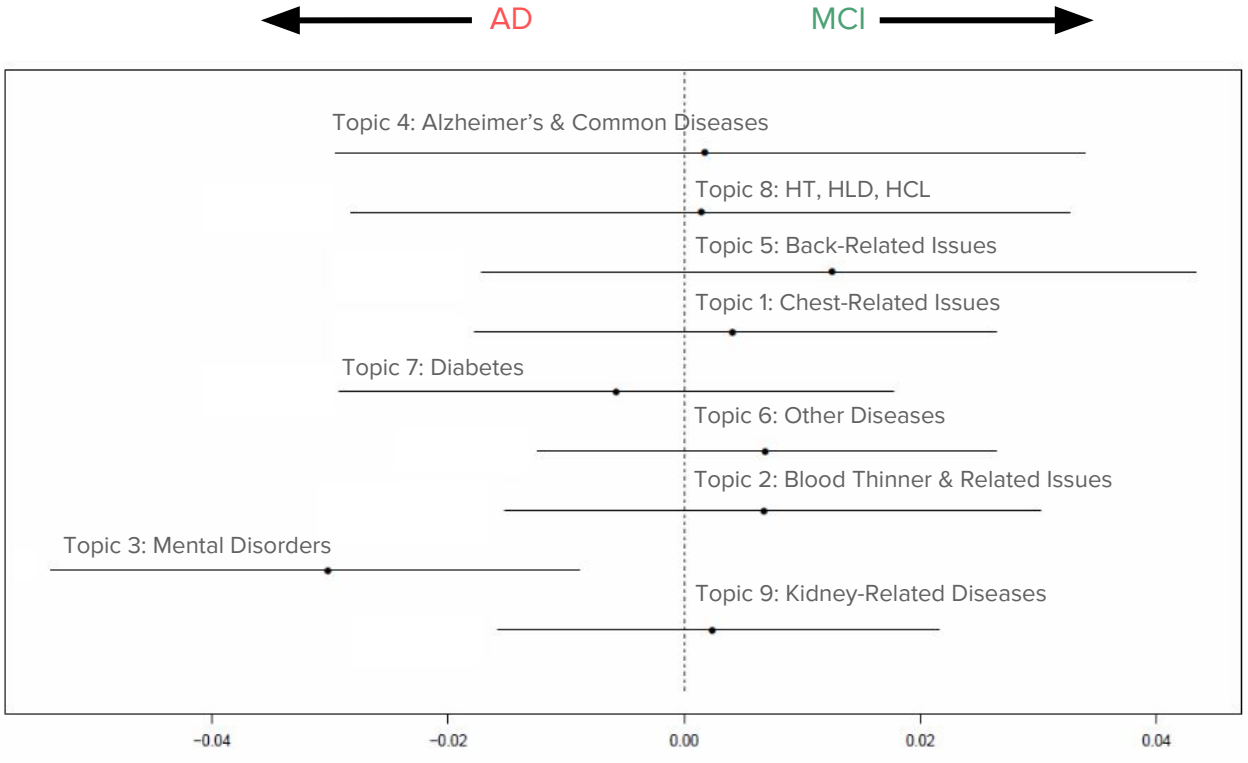
- Chronic Airway Obstruction
- Congestive Heart Failure
- Chest Pain
- Shortness Of Breath
- Hypertension
- Coronary Atherosclerosis Of Native Coronary Artery
- Coronary Atherosclerosis Of Unspecified Type Of Vessel, Native Or Graft



HT, HLD, HCL

- Benign Hypertension
- Hyperlipidemia
- Hypertension
- Actinic Keratosis
- Pure Hypercholesterolemia
- Prophylactic Vaccination and Inoculation Against Influenza
- Exudative Senile Macular Degeneration

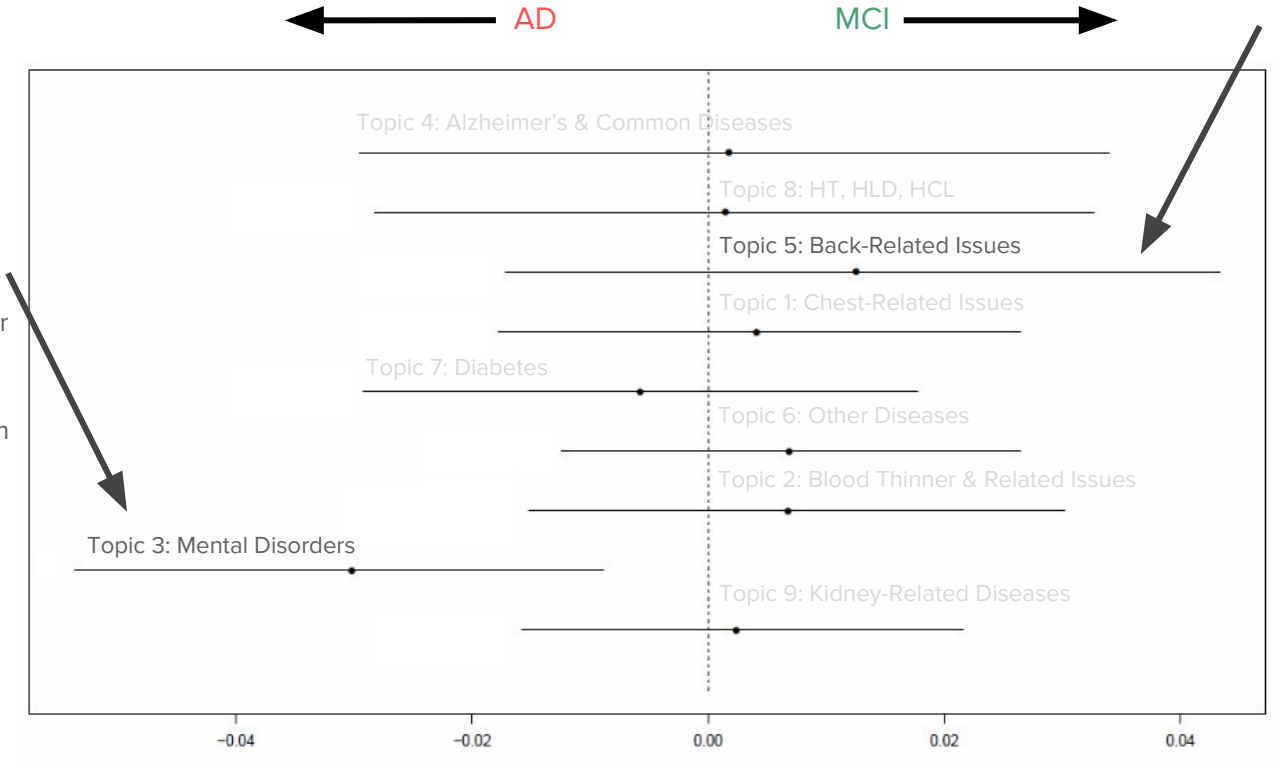
Topic Proportion Comparison: AD vs. MCI Patients



Topic Proportion Comparison: AD vs. MCI Patients

(?) Mental Disorders

- Major Depressive Affective Disorder
- Bipolar Disorder
- Schizoaffective Disorder
- Depressive Disorder Not Elsewhere Classified
- Malignant Neoplasm Of Breast (Female) Unspecified Site
- Long-term (Current) Use Of Other Medications
- Other Malignant Lymphomas Unspecified Site



Back-Related Issues

- Lumbago
- Thoracic Or Lumbosacral Neuritis Or Radiculitis
- Nonallopathic Lesions, Lumbar Region
- Degeneration Of Lumbar
- Spinal Stenosis, Lumbar Region
- Nonallopathic Lesions, Cervical Region
- Lumbosacral Spondylosis Without Myelopathy

Conclusion

Using structural topic models, we uncovered hidden topics, or groups of common diseases. Assuming that our data subset is somewhat representative of the population, we can speculate that the following could be true:

- **Mental disorders are more common in AD patients vs. MCI patients**
- Cancer, particularly Malignant Lymphomas and Multiple Myeloma, is more common in MCI patients
- Coronary Atherosclerosis is common for those patients diagnosed with both AD and MCI

Structural topic model is an innovative and relatively new technique in the statistics and natural language processing realms. If implemented on the whole dataset, which is doable, one may find more interesting results using this technique, particularly if sequences of icd_code is taken into account.

Further Study: Why were diagnoses related to lung diseases, sacroiliac joint inflammation, lymphatic system blockage, and skin rashes considered *similar* to “Dementia” using Word2Vec? Is there a connection there? 😉

Thank you!

Appendix

Structural Topic Model Resources

Structural Topic Model, like other topic models, uncovers the underlying topics, concepts, or themes that occur in a collection of documents. What sets this particular model apart, however, is the fact that covariates can be entered into the model, which can improve inference and qualitative interpretability. In our project's case, this means that we can simultaneously estimate the prevalence of diseases among Alzheimer's and MCI patients, controlling for covariates.

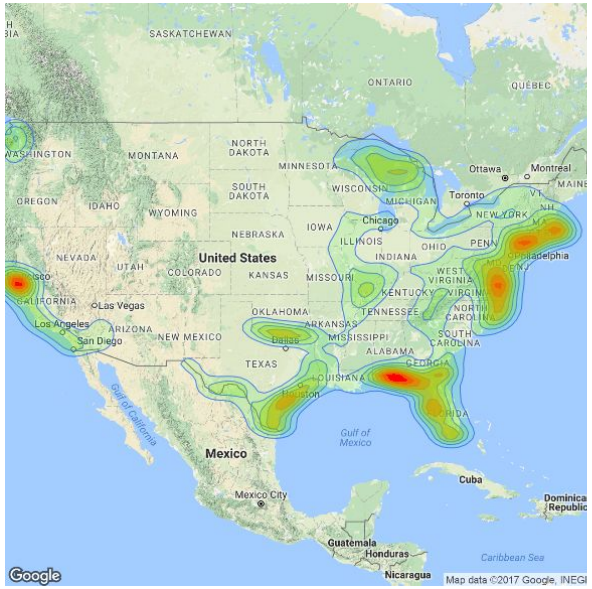
"The Structural Topic Model and Applied Social Science", Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, Edoardo M. Airoldi

"stm: An R Package for the Structural Topic Model", Authors: Molly Roberts, Brandon M. Stewart, Dustin Tingley

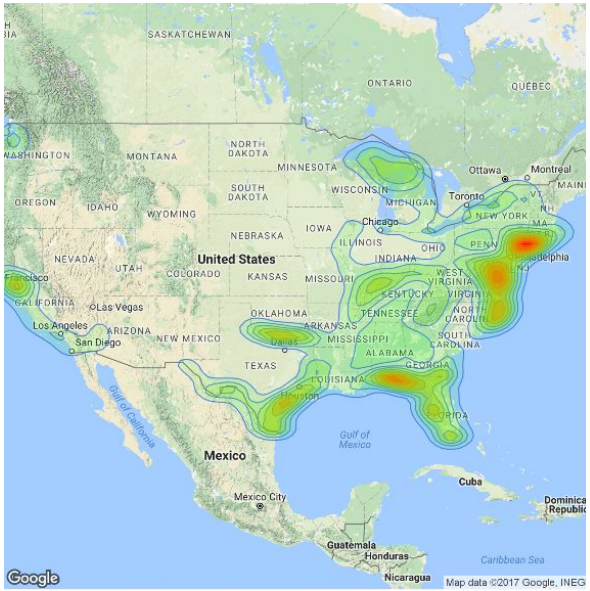
- *Includes vignette, multiple methods papers, supporting packages (visualization options using D3), and published applications*

Our Data Subset

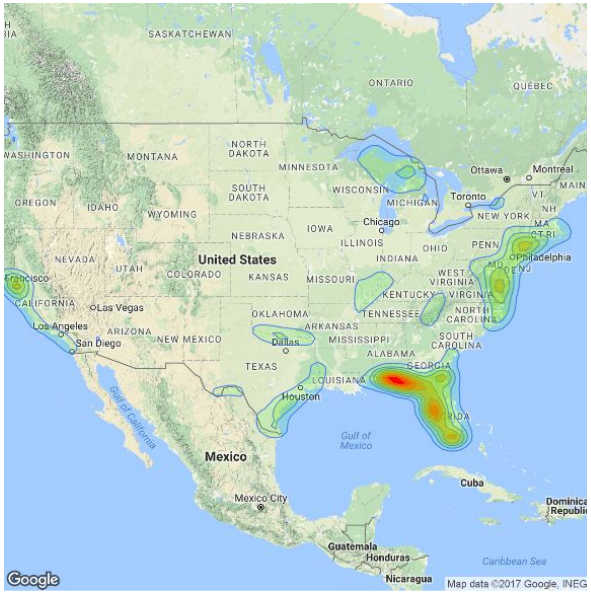
MCI Patients (500)



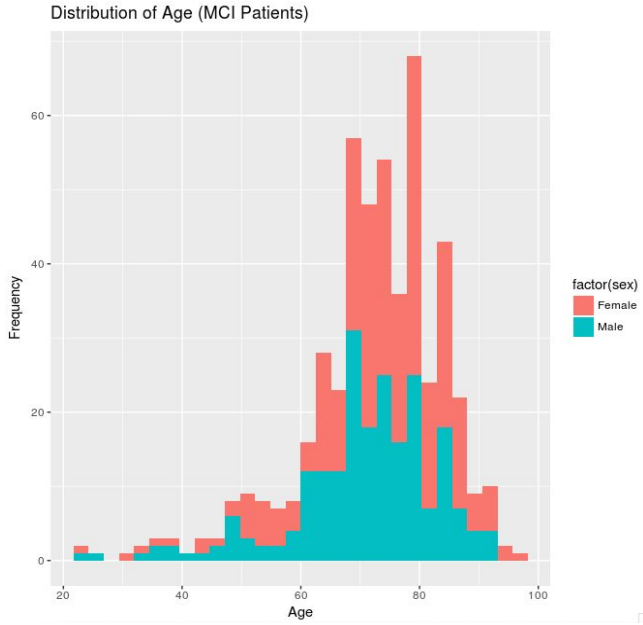
AD Patients (500)



MCI & AD Patients (500)

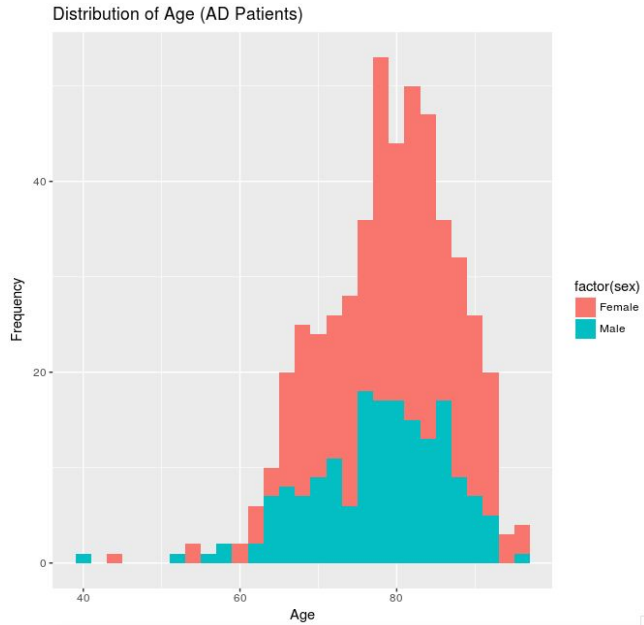


Our Data Subset



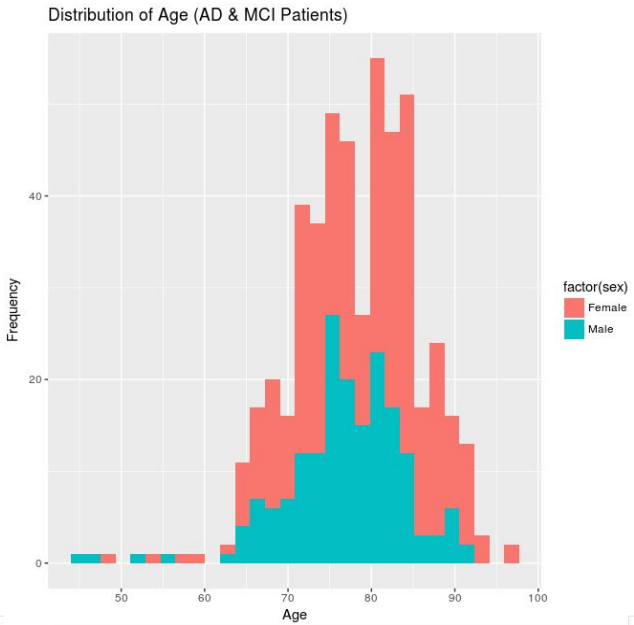
Female: 281

Male: 219



Female: 326

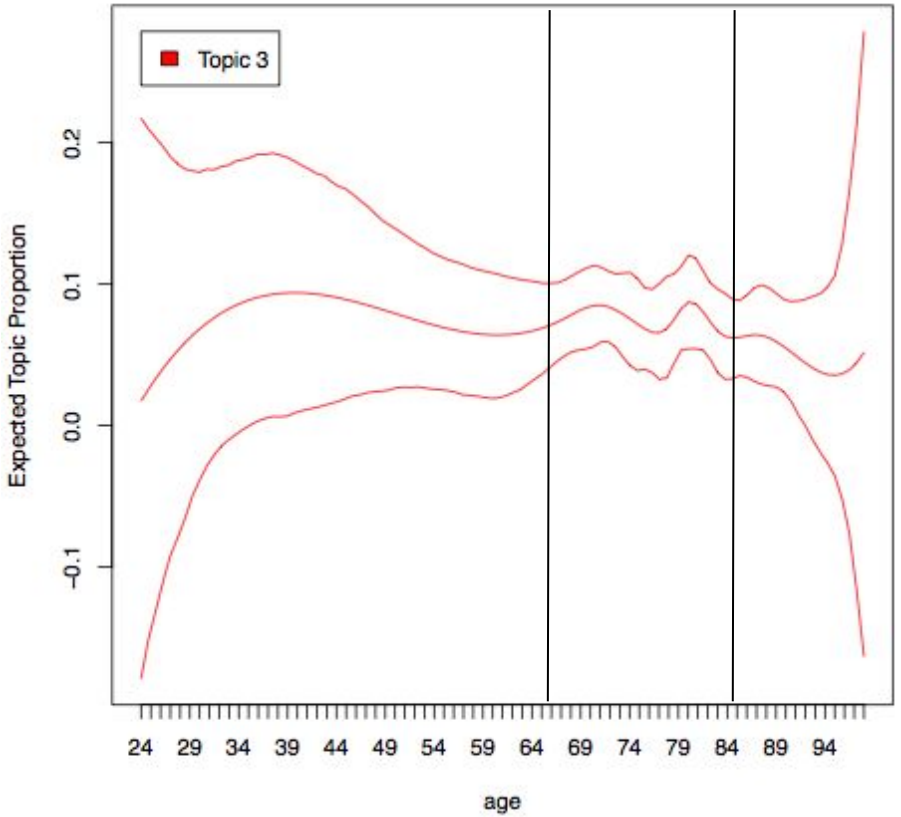
Male: 174



Female: 319

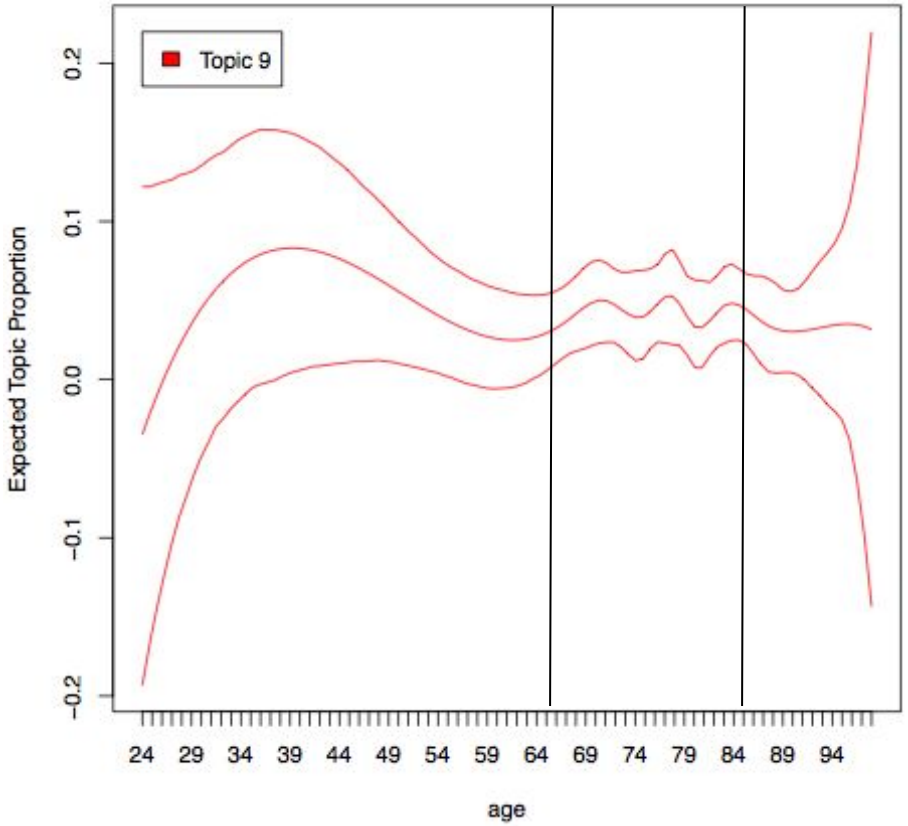
Male: 181

Age Trend of Topic 3: Mental Disorders



- 29633 Major Depressive Affective Disorder
- 29680 Bipolar Disorder
- 29570 Schizoaffective Disorder
- 311 Depressive Disorder Not Elsewhere Classified
- 1749 Malignant Neoplasm Of Breast (Female) Unspecified Site
- V5869 Long-term (Current) Use Of Other Medications
- 20280 Other Malignant Lymphomas Unspecified Site

Age Trend of Topic 9: Kidney-Related Diseases



- 5856 End Stage Renal Disease
- 7140 Rheumatoid Arthritis
- 28521 Anemia In Chronic Kidney Disease
- 5853 Chronic Kidney Disease, Stage III
- 5859 Chronic Kidney Disease, Unspecified
- 5854 Chronic Kidney Disease, Unspecified, Stage II