Text Mining for Associations in Cardiovascular Case Reports

Presented by Kitu Komya
On behalf of J. Harry Caufield

August 4, 2017

Background

- A **case report** is a detailed report written by a clinician about a patient's demographics, symptoms, diagnosis, and treatments
- **MeSH terms** **(medical subject headings)** are standardized, summarized phrases a clinician uses to describe the patient, such as demographics, diseases, and treatments
- **RN terms** **(registry number)** are drugs/substances mentioned in the case report
- **PubMed** is a free search engine that contains case reports and other scientific literature
- **Text mining** is a method to analyze data stored in text format (in our case, case reports)

background - **overview** - goals - workflow - clean - analyze - visualize - summary - next steps

Overview

Problems

The number one cause of death is heart disease, which if better understood, could lower death rate

Hundreds of thousands of cardiovascular case reports are gold mines of knowledge, yet unanalyzed

Solutions

Find the most occurring MeSH or RN terms within cardiovascular case reports

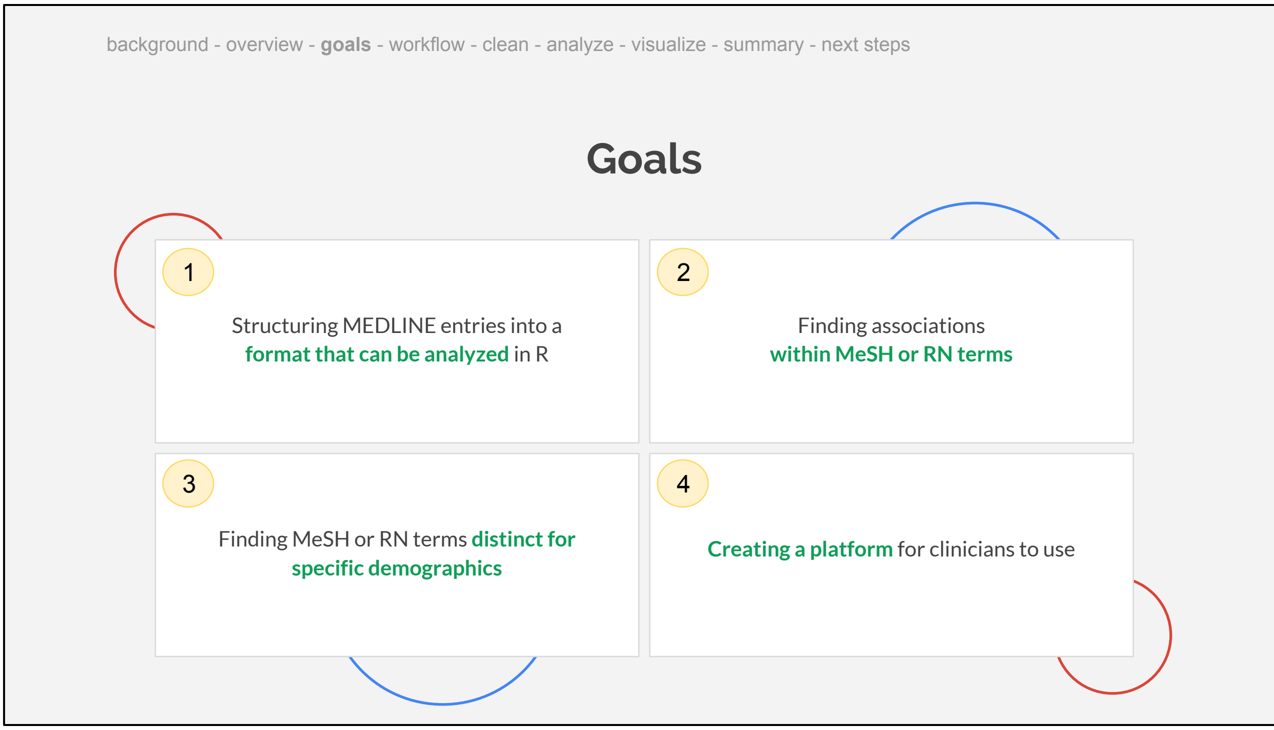
Find associations between MeSH or RN terms and specific demographics of people

Impacts

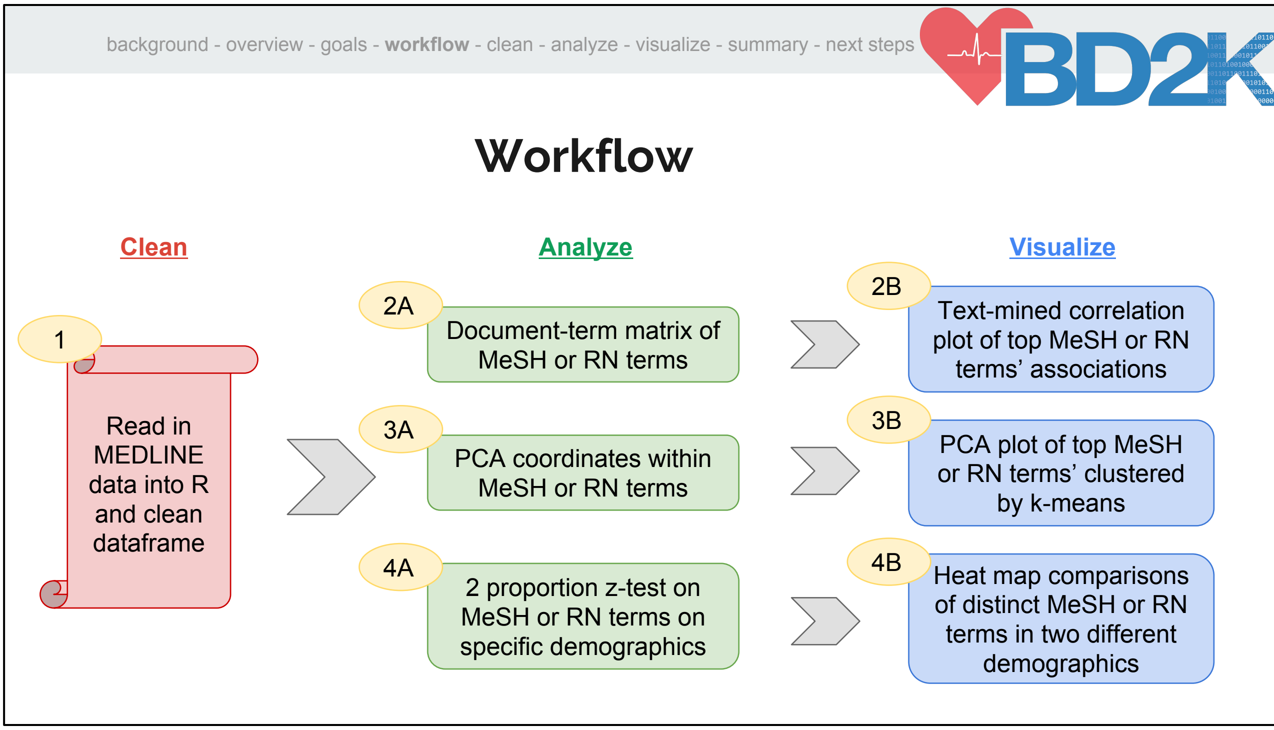
Will determine which MeSH or RN terms, for instance, are most associated with a female aged patient with a stroke

Will allow clinicians to make informed, customized decisions based on data

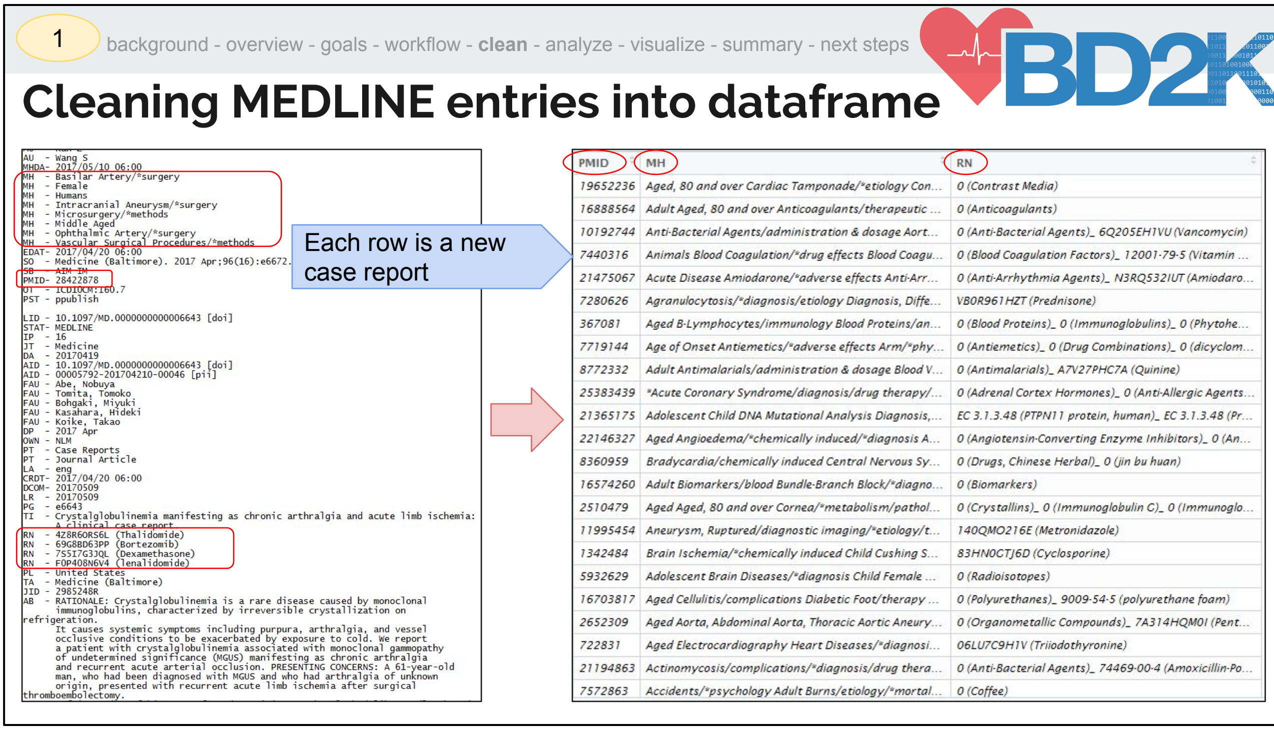
Through this project, we hope to uncover patterns within cardiovascular case reports. Given that heart disease is the number one cause of death, this project is important to pursue in order to better understand heart diseases. This project will look specifically into MeSH and RN terms, which are standardized phrases, and explore what associations exist among them each. Then, the project will look into specific demographics of people to see which MeSH terms and which RN terms are most distinct to that group of people. The impact of this project is huge: it can determine exactly which MeSH and RN terms are most distinct for a particular population group, which will immensely aid in clinical diagnostics and making more informed decisions on behalf of the clinicians.



The structure of the raw data is quite messy. The project hopes to clean it into a readable format that can be analyzed. Then, we will explore associations within MeSH and RN terms. Afterward, we'll look into which MeSH and RN terms are distinct for certain groups of populations. Finally, we hope to create a platform that enables clinicians to input their own demographics to have an output of MeSH and RN terms specific to their inputted demographics.



There are three main steps: Cleaning, analyzing, and visualizing the data. In order to clean the data, I loaded the data into RStudio and cleaned it into a neat dataframe, where each row represents a distinct case report. Afterward, I analyzed the data using three different features, each of which lead to its own visualization. I created a document-term matrix of the MeSH and RN terms which lead to the correlation plot of the top MeSH and RN terms' association. This plot shows relationships among multiple MeSH and RN terms. In addition, I applied the PCA technique onto the MeSH and RN terms and created PCA coordinates clustered by k-means. Its plot shows spatial relationship among all the MeSH or RN terms. Finally, I have currently been working on 2 proportion z-tests on MeSH and RN terms on specific demographics which show which MeSH or RN term is most distinct to each demographic. From these numbers, I created a heat map that shows the differences between two different demographics.



On the left you see what the raw MEDLINE data format looks like. I have circled the three most important attributes: PMID (an ID for each case report), MH (MeSH terms), and RN (RN terms). As you can see, a case report can have multiple MeSH or RN terms. However, this format is unusable in data science, so I converted it into a neat dataframe in RStudio. As you can see, I have only preserved the 3 variables that I had mentioned in this dataframe. Each row represents a new case report. This data is much more easy to work with.

BD2K

Further cleaning dataframe

PMID	MH	RN
19652236	Aged, 80 and over Cardiac Tamponade/*etiology Con…	0 (Contrast Media)
16888564	Adult Aged, 80 and over Anticoagulants/therapeutic …	0 (Anticoagulants)
10192744	Anti-Bacterial Agents/administration & dosage Aort…	0 (Anti-Bacterial Agents)_ 6Q205EH1VU (Vancomycin)
7440316	Animals Blood Coagulation/*drug effects Blood Coagu…	0 (Blood Coagulation Factors)_ 12001-79-5 (Vitamin …
21475067	Acute Disease Amiodarone/*adverse effects Anti-Arr…	0 (Anti-Arrhythmia Agents)_ N3RQ532IUT (Amiodaro…
7280626	Agranulocytosis/*diagnosis/etiology Diagnosis, Diffe…	VB0R961HZT (Prednisone)
367081	Aged B-Lymphocytes/immunology Blood Proteins/an…	0 (Blood Proteins)_ 0 (Immunoglobulins)_ 0 (Phytohe…
7719144	Age of Onset Antiemetics/*adverse effects Arm/*phy…	0 (Antiemetics)_ 0 (Drug Combinations)_ 0 (dicyclom…
8772332	Adult Antimalarials/administration & dosage Blood V…	0 (Antimalarials)_ A7V27PHC7A (Quinine)
25383439	*Acute Coronary Syndrome/diagnosis/drug therapy/…	0 (Adrenal Cortex Hormones)_ 0 (Anti-Allergic Agents…
21365175	Adolescent Child DNA Mutational Analysis Diagnosis,…	EC 3.1.3.48 (PTPN11 protein, human)_ EC 3.1.3.48 (
22146327	Aged Angioedema/*chemically induced/*diagnosis A…	0 (Angiotensin-Converting Enzyme Inhibitors)_ 0 (
8360959	Bradycardia/chemically induced Central Nervous Sy…	0 (Drugs, Chinese Herbal)_ 0 (jin bu huan)
16574260	Adult Biomarkers/blood Bundle-Branch Block/*diagno…	0 (Biomarkers)
2510479	Aged Aged, 80 and over Cornea/*metabolism/pathol…	0 (Crystallins)_ 0 (Immunoglobulin C)_ 0 (Immunoglo…
11995454	Aneurysm, Ruptured/diagnostic imaging/*etiology/t…	140QMO216E (Metronidazole)
1342484	Brain Ischemia/*chemically induced Child Cushing S…	83HN0CTJ6D (Cyclosporine)
5932629	Adolescent Brain Diseases/*diagnosis Child Female …	0 (Radioisotopes)
16703817	Aged Cellulitis/complications Diabetic Foot/therapy …	0 (Polyurethanes)_ 9009-54-5 (polyurethane foam)
2652309	Aged Aorta, Abdominal Aorta, Thoracic Aortic Aneury…	0 (Organometallic Compounds)_ 7A314HQM0I (Pent…
722831	Aged Electrocardiography Heart Diseases/*diagnosi…	06LU7C9H1V (Triiodothyronine)
21194863	Actinomycosis/complications/*diagnosis/drug thera…	0 (Anti-Bacterial Agents)_ 74469-00-4 (Amoxicillin-Po…
7572863	Accidents/*psychology Adult Burns/etiology/*mortal…	0 (Coffee)

PMID	MH	named_RN
19652236	Aged 80 and over, Cardiac Tamponade etiology, Cont…	
16888564	Adult, Aged 80 and over, Anticoagulants therapeutic …	
10192744	Anti Bacterial Agents administration and dosage, Ao…	Vancomycin
7440316	Animals, Blood Coagulation drug effects, Blood Coagu…	Vitamin K, Warfarin
21475067	Acute Disease, Amiodarone adverse effects, Anti Arrh…	Amiodarone
7280626	Agranulocytosis diagnosis etiology, Diagnosis Differe…	Prednisone
367081	Aged, B Lymphocytes immunology, Blood Proteins an…	
7719144	Age of Onset, Antiemetics adverse effects, Arm phys…	Dicyclomine, Nitrofurantoin, Doxylamine, Pyridoxine
8772332	Adult, Antimalarials administration and dosage, Bloo…	Quinine
25383439	Acute Coronary Syndrome diagnosis drug therapy et…	clopidogrel, Ticlopidine, Aspirin
21365175	Adolescent, Child, DNA Mutational Analysis, Diagnosi…	PTPN11 protein, human, Protein Tyrosine Phosphatas…
22146327	Aged, Angioedema chemically induced diagnosis, An…	
8360959	Bradycardia chemically induced, Central Nervous Sy…	
16574260	Adult, Biomarkers blood, Bundle Branch Block diagno…	
2510479	Aged, Aged 80 and over, Cornea metabolism patholo…	
11995454	Aneurysm Ruptured diagnostic imaging etiology ther…	Metronidazole
1342484	Brain Ischemia chemically induced, Child, Cushing Sy…	Cyclosporine
5932629	Adolescent, Brain Diseases diagnosis, Child, Female, …	
16703817	Aged, Cellulitis complications, Diabetic Foot therapy, …	polyurethane foam
2652309	Aged, Aorta Abdominal, Aorta Thoracic, Aortic Aneur…	Pentetic Acid, Technetium Tc 99m Pentetate
722831	Aged, Electrocardiography, Heart Diseases diagnosi…	Triiodothyronine
21194863	Actinomycosis complications diagnosis drug therap…	Amoxicillin-Potassium Clavulanate Combination
7572863	Accidents psychology, Adult, Burns etiology mortalit…	

Removed general RN terms with a “0” code

However, although this format is preferred, there was more cleaning to do. MeSH terms contain special characters, such as apostrophes and slashes and parantheses, and many of the techniques in text mining do not recognize these special characters and simply use it as a way to split up phrases. I cleaned up the MeSH terms so that each entry was simply separated by a comma and had no special characters. I also immensely reduced the size of the RN terms. On the left, you can see that all of the RN terms have codes preceeding it. RN terms with a “0” code are more general drug terms, whereas those with actual numbers are a specific, standardized drug number. To keep drug terms consistent and named, I removed all the RN terms with the “0” code, which you can cross verify by seeing that only the named RN terms are included in the dataframe in the right.

BD2K

Analyzing the data: document-term matrix

Definition: A **document-term matrix (dtm)** is a matrix that describes the frequency of terms that occur in a set of documents. Its rows correspond to documents (in our case, case reports), and the columns are terms (in our case, MeSH or RN terms)

Purpose: A dtm can find associations among terms

Example:

Columns correspond to MeSH or RN terms

Rows correspond to case reports

	administration	adverse	aged	and	cardiac	computed	contrast	diagnostic	dosage	effects	etiology	fatal	female	heart
1	1	1	1	2	1	1	1	1	1	1	2	1	1	1
2	0	0	1	4	0	0	0	5	0	0	3	0	1	0
3	2	0	1	2	0	0	0	0	2	0	0	0	0	0
4	0	1	0	0	0	0	0	0	0	2	1	0	1	0
5	0	3	1	1	0	0	0	0	0	3	3	0	0	0
6	0	0	1	0	0	0	0	0	0	0	2	0	1	0
7	0	0	1	1	0	0	0	2	0	0	0	0	0	0
8	0	4	0	0	0	0	0	0	0	4	0	0	1	0
9	2	0	0	2	0	0	0	0	2	0	0	0	0	0
10	7	0	1	8	0	0	0	0	7	0	2	0	1	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	2	2	1	2	0	0	0	0	2	2	0	0	1	0
13	0	0	0	0	0	0	0	0	0	0	0	0	1	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	1
15	0	0	2	1	0	0	0	0	0	0	1	0	1	0

After cleaning the dataframe, the first step was to analyze the associations of MeSH terms with each other, as well as RN terms with each other. A technique using document-term matrix does exactly that. This matrix essentially shows the counts of each MeSH or RN term (which is a column name) to each case report (which is a row name). This aids in finding associations within MeSH terms. For instance, a MeSH term need not be contained in the same case report as another MeSH term to have high association. If both MeSH terms appear across similar case reports or across similar MeSH terms, then they will be highly correlated.

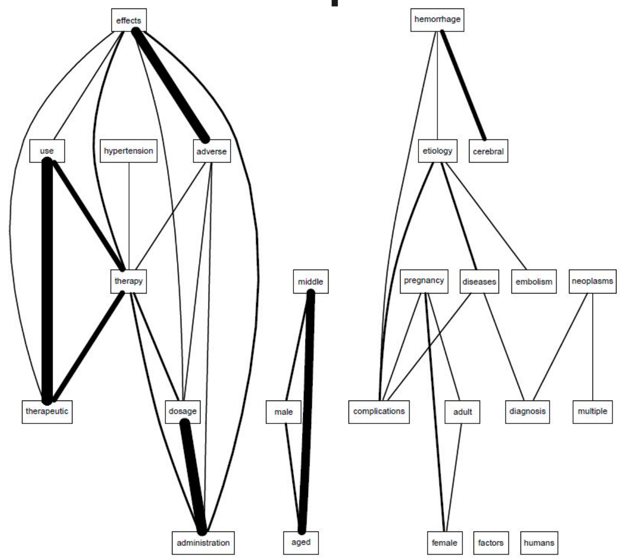
Visualizing the data: dtm in a correlation plot

Advantages:

- Plot shows associations among top 25 MeSH terms
- The darker the lines, the stronger the correlation

Disadvantages:

- Does not take in phrases of MeSH terms, only singular words (middle aged broken down into "middle" and "aged")
- No spatial relationship among clusters shown



The document term matrix technique leads us to this correlation plot, which contains the top 25 MeSH terms from a collection of cardiovascular case reports. The darker the lines, the stronger the association between the words. However, although this technique is useful in showing overarching relations among multiple MeSH terms, it splits phrases into words. For instance, "middle aged" has been split into "middle" and "aged", so although useful, it's not exactly accurate in portraying real relationships among meaningful phrases. Another problem is that this plot doesn't show spatial relationships among terms. It simply shows, in random space, where each "cluster" belongs. A more useful plot would show the spatial relationship of each cluster. However, this plot is not absolutely meaningless, as we can see that the relationships between "pregnancy" and "female" are accurately portrayed, as are other clusters that make sense.

BD2K

Analyzing the data: PCA coordinates

Using PCA to show associations within MeSH terms in only 2 dimensions

Clustering by k-means to group MeSH terms into similar categories

K-means is essentially a technique to cluster observations that are close to each other spatially

ID	PCA1	PCA2	Cluster
etiology	0.31955191	0.952384672	6
heart	-1.07218194	3.545831521	4
physiopathology	-1.49130028	0.518461947	9
valve	-1.42583232	3.268753899	4
aged	0.21139784	0.663125198	6
artery	1.59014325	-0.008467525	2
adult	0.10536255	0.583295980	7
diseases	0.49128025	0.654443187	6
middle	0.19827658	0.629678619	6
aneurysm	1.17912151	0.838794266	6
female	-0.01411499	0.393900254	7
aortic	-0.23814353	1.406392677	6
child	-0.05902981	0.244917624	7
neoplasms	0.15936839	1.357988426	6
male	-0.09333556	0.204639022	7
abnormalities	-0.10967766	0.162379323	7
electrocardiography	-1.18755650	0.325481043	9

To solve the problem of enabling spatial association among terms, I next tried the PCA technique so that it could cluster all similar words together onto 2 dimensions only. I clustered it by k-means so that on the plot, it's grouped aesthetically by categories that make sense. K-means is essentially a technique that groups words that are close to each other together. As you can see from this dataframe itself, the clusters already make sense: "heart" and "valve" are anatomical terms which are in the same cluster, and "female" and "male" are also in the same cluster.

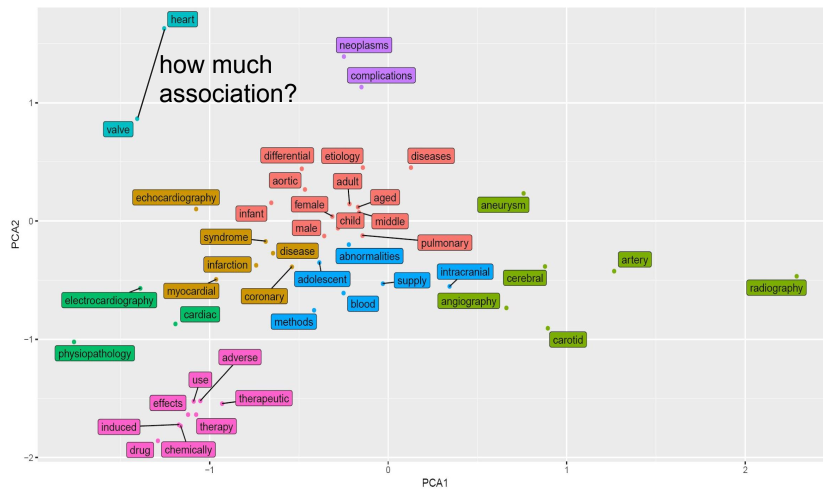
Visualizing the data: PCA plot clustered by k-means

Advantages:

- ➢ Plot shows associations within top 50 MeSH terms
- ➢ Can see clusters that make sense
- ➢ This particular plot only takes in singular words, but can take in phrases

Disadvantage:

- ➢ Doesn't quantitatively assess associations



The resulting PCA plot looks like this. This plot shows the top 50 MeSH terms from a collection of cardiovascular case reports. You can see that these clusters make sense. In the red cluster, you have mostly demographics, while in the pink cluster, you have very general MeSH terms. This plot allows us to see spatial associations among MeSH terms as well as seeing clusters within MeSH terms. Note that since this was a prior project, this plot only contains singular words. After I had moved on from this project, I had finally learned how to read in MeSH terms with multiple words in, so I may possibly go back to this project to see what the plot looks like with multiple worded phrases. As of now though, I have moved on to work on a different project.

Analyzing the data: 2 proportion z-test

Definition: This test determines statistical significance between two proportions, taking into account different sample sizes.

Example: Can show if proportion of male case reports containing “stroke” is statistically different from the proportion of female case reports containing “stroke”

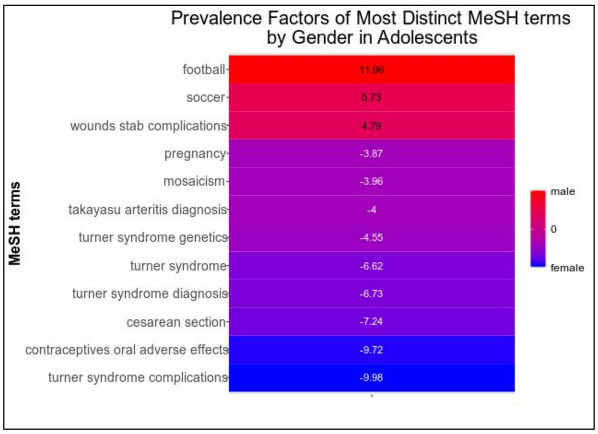
Data:

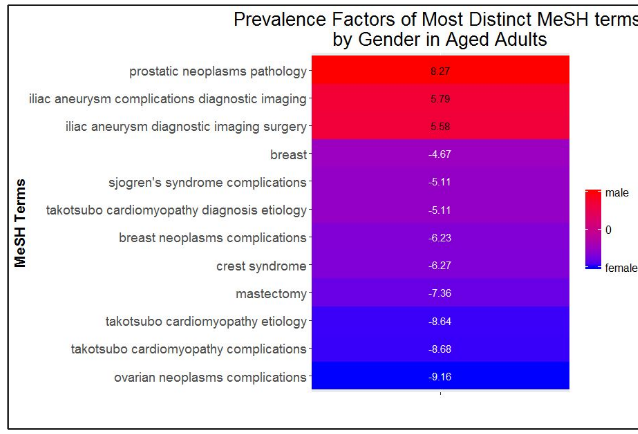
word	male_aged	female_aged	prop_male	prop_female	p-value	factor
ovarian neoplasms complications	7	53	0.0001034386	0.0009472744	4.857683e-11	-9.16
takotsubo cardiomyopathy complications	23	165	0.0003398697	0.0029490617	2.433173e-31	-8.68
takotsubo cardiomyopathy etiology	7	50	0.0001034386	0.0008936550	2.811081e-10	-8.64
prostatic neoplasms pathology	50	5	0.0007388471	0.0000893655	1.484001e-07	8.27
mastectomy	12	73	0.0001773233	0.0013047364	1.187596e-13	-7.36
crest syndrome	11	57	0.0001625464	0.0010187668	3.633843e-10	-6.27
breast neoplasms complications	20	103	0.0002955388	0.0018409294	2.105032e-17	-6.23
iliac aneurysm complications diagnostic imaging	56	8	0.0008275088	0.0001429848	2.734801e-07	5.79
iliac aneurysm diagnostic imaging surgery	54	8	0.0007979549	0.0001429848	5.973657e-07	5.58
takotsubo cardiomyopathy diagnosis etiology	18	76	0.0002659850	0.0013583557	8.381359e-12	-5.11
sjogren's syndrome complications	18	76	0.0002659850	0.0013583557	8.381359e-12	-5.11
breast	131	506	0.0019357794	0.0090437891	2.586728e-67	-4.67

Will be using values from “factor” to create plots

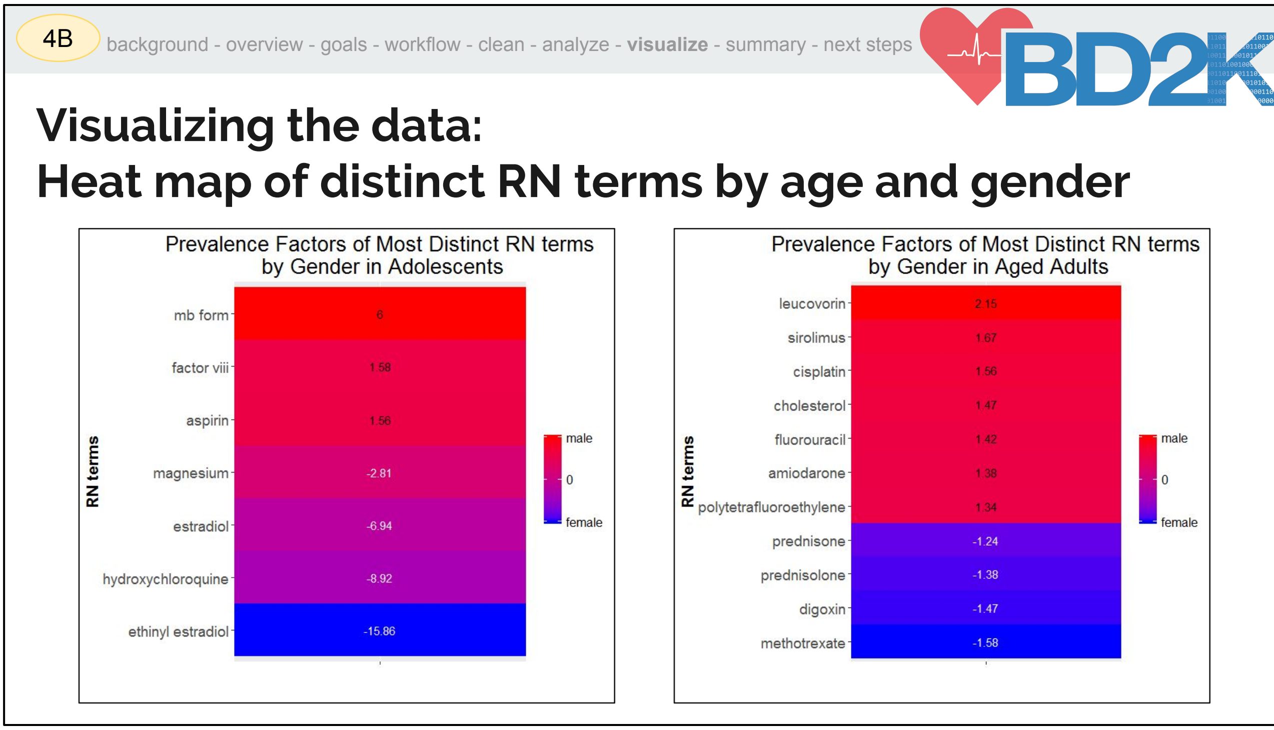
Analyzing the data via 2 proportion z-tests has been my latest project. This method will allow me to find statistical significance in the proportion of case reports across two different demographics. For instance, if I want to know whether the proportion of male case reports containing “stroke” is statistically significant from the proportion of female case reports containing “stroke,” this test will be of use, since it also takes into account different sample sizes. As you can see from the data, the “word” refers to the MeSH terms, the “male_aged” refers to the raw count of number of case reports within “aged male” population that contains each “word,” as does the “female_aged” variable. “prop_male” and “prop_female” refer to the proportion of case reports that include each word. Since the majority of case reports are male, it’s important to compare proportions instead of raw counts. The “p-value” shows the results of the test. All of the words in this dataframe are significant, with a p-value less than 0.05. The “factor” variable refers to the factor by which the higher proportion is to the lower proportion between “prop_male” and “prop_female.” We will actually be using the “factor” to display the visualizations on the next few slides.

Visualizing the data:
Heat map of distinct MeSH terms by age and gender



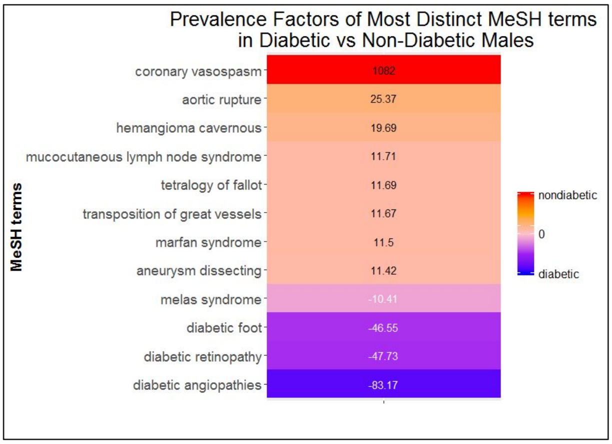


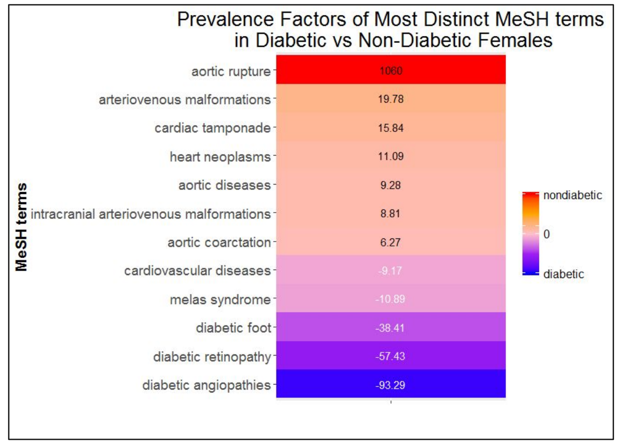
These heat maps show distinct MeSH terms across age and gender. On the left, you have a heat map that compares MeSH terms between female adolescents (bottom and in white text) and male adolescents (top and in black text). Some of these MeSH terms make sense, such as “football” and “soccer” being associated with male adolescents, and “pregnancy” and “cesarean section” associated with female adolescents. The numbers on each MeSH term represent the factor. So for instance, to interpret football’s “11.06,” we could say that “football is a MeSH term that is prevalent in male adolescent cardiovascular case reports 11.06 times more than in female adolescent cardiovascular case reports.” Similarly, the heat map on the right shows the distinct differences between female aged adults and male aged adults. Interestingly, none of the MeSH terms across the adolescent and aged adults demographics match, which leads us to believe that both age and gender play a role in MeSH terms.



Similar to the previous plots, these two heat maps show the most distinct RN, or drug, terms across the same age groups. On the left, you see the distinct drugs used between female adolescents versus male adolescents. On the right, you see the distinct drugs used between female aged adults and male aged adults. Again, it's interesting to note that none of the drug terms are common across the two age demographics. I have conducted similar analyses for 7 age groups in regards to MeSH and RN term differences across gender: infant, child, adolescent, young adult, adult, middle aged, and aged adults. From those, I have only shown 2 age demographics.

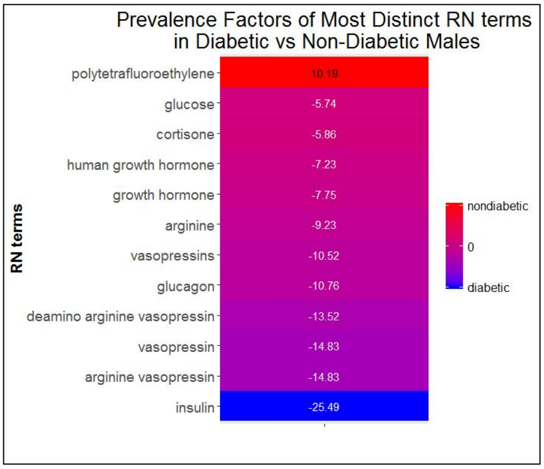
Visualizing the data:
Heat map of diabetics’ distinct MeSH terms by gender

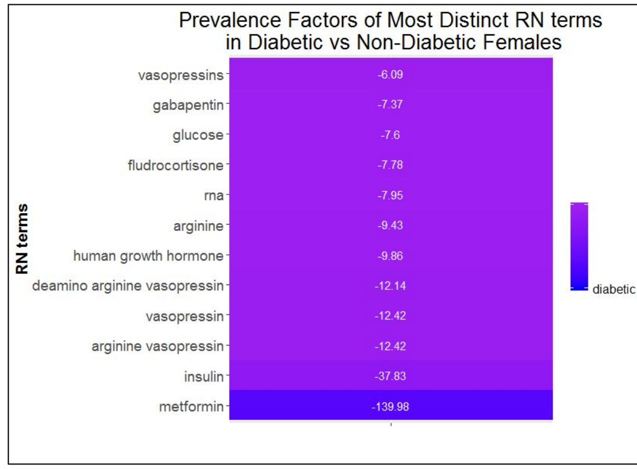




Similar to the previous plots, the plots above show the distinct MeSH terms across diabetic patients. On the left, you have a heat map of distinct MeSH terms of diabetic male patients versus nondiabetic male patients. On the right, you have a heat map of distinct MeSH terms of diabetic female patients versus nondiabetic female patients. Very interestingly, both heat maps have a MeSH term that is strikingly present in nondiabetic patients. After conferring with Dr. David Liem, we came to the conclusion that the nondiabetic terms are not necessarily “anti-diabetic” but rather “less present in diabetic patients.” It’s still quite alarming to see that there is such a high correlation against diabetic patients. However, the plots are still valid as we can confirm that the MeSH terms listed for the diabetic patients are, in fact, very true to diabetic patients, such as “melas syndrome, diabetic foot, diabetic retinopathy, and diabetic angiopathies.”

Visualizing the data:
Heat map of diabetics' distinct RN terms by gender





Here are two heat maps comparing distinct RN terms across diabetic male versus nondiabetic male patients on the left as well as diabetic female versus nondiabetic female patients on the right. Very interestingly, most of these RN terms are strictly for diabetic patients, which makes sense since nondiabetic patients won't necessarily be needing medication. Across both genders, there are a few common RN terms, but also some striking differences, such as "metformin" in females and "cortisone" in males.

Summary

- Text mining to find associations **within MeSH terms or within RN terms**
- Text mining to find associations of MeSH and RN terms **in specific demographics**
- Results will **find trends** that would have otherwise been lost in the data
- Further analyses will lead to **improved, personalized clinical diagnostics**

After that brief overview of what my project has been about, I would like to conclude by restating what the project's main purposes are. This project mainly revolves around text mining techniques to find associations among MeSH and RN terms as well as finding associations within specific demographics. These results will allow us to find patterns that otherwise would not have been discovered. These findings can lead to more tuned, improved clinical diagnostics as well, since a clinician can use a patient's specific demographics to understand their conditions better.

Next Steps

- **Create a platform** (website) using R Shiny for clinicians so that they can input demographics (“female”, “aged”, “diabetes mellitus”) and receive an **output of distinct MeSH and RN terms**
- Website will include **correlation plot**, **PCA plot**, as well as **heat map**

However, the project is far from done. I hope to build an accessible website that will allow clinicians to input their own demographics instead of me manually choosing demographics and analyzing its data. The website will include all three visualizations you have seen in this presentation and will primarily aim to aid clinicians.

Acknowledgements

Thank you for your time in listening to my presentation, Dr. Ping.

I would also like to acknowledge BD2K's lab resources and J. Harry Caufield for supporting me in my endeavors and for giving me the opportunity to explore data science.

Finally, I would like to thank you, Dr. Ping for your time in listening to my presentation. I very much appreciate your time. I am also very grateful for the opportunity to work in this lab and to conduct data science under the supervision of Harry. Thank you again, for allowing me to pursue this endeavor.