Exploring Geographic and Drug Data from Annotated Case Reports

Presented by
Clement Feyt, Kitu Komya, and Joel Perez

December 1, 2017

Background(1)

- Millions of medical case reports published on-line
- Aggregation of textual knowledge using Natural Language Processing: incredible potential of new discoveries and improving evidence based medicine
- Necessity of constituting a training set of labeled case reports
    - Manual annotation of 3000 case reports (thanks to David, Sanjana, Jessica, John, Travis, Anders, Joshua, Harry, Sarah, Clement, Dibakar and Michaela), of which 200 are still under quality control process

Background(2)

- Multiple disease systems to explore
- Promising avenues to analyze first are geographic and drug data:
  - How does **drug usage** compare between **gender and age**? (Joel)
  - How does **location** affect **frequency of case reports** within a certain disease system? (Kitu)

Goals

1. Text-mine from annotated data to extract relevant information
2. Analyze drug usage between Gender and Age(Joel)
3. Explore relationship between location and frequency of case reports within a certain disease system (Kitu)
4. Create a usable and interactive platform to present data (R Shiny)

Joel’s Workflow: Pharmacological drug Data

Clean

1

Extracting Drugs from Case Reports

Analyze

2

Top Drugs between groups

Visualize

3

Plot based on Gender and Age Groups

Extraction Process

- Started with the drugs list dataset from National Library of Medicine
- Cleaning process:
  - Removed duplicates
  - Subsets:
    - e.g. “aspirin”, “aspirins”
    - e.g. “Mesna”, “product containing mesna”, “product containing mesna medical product”
- Compared list to the case reports’ Pharmacological Therapy column to extract the drug names

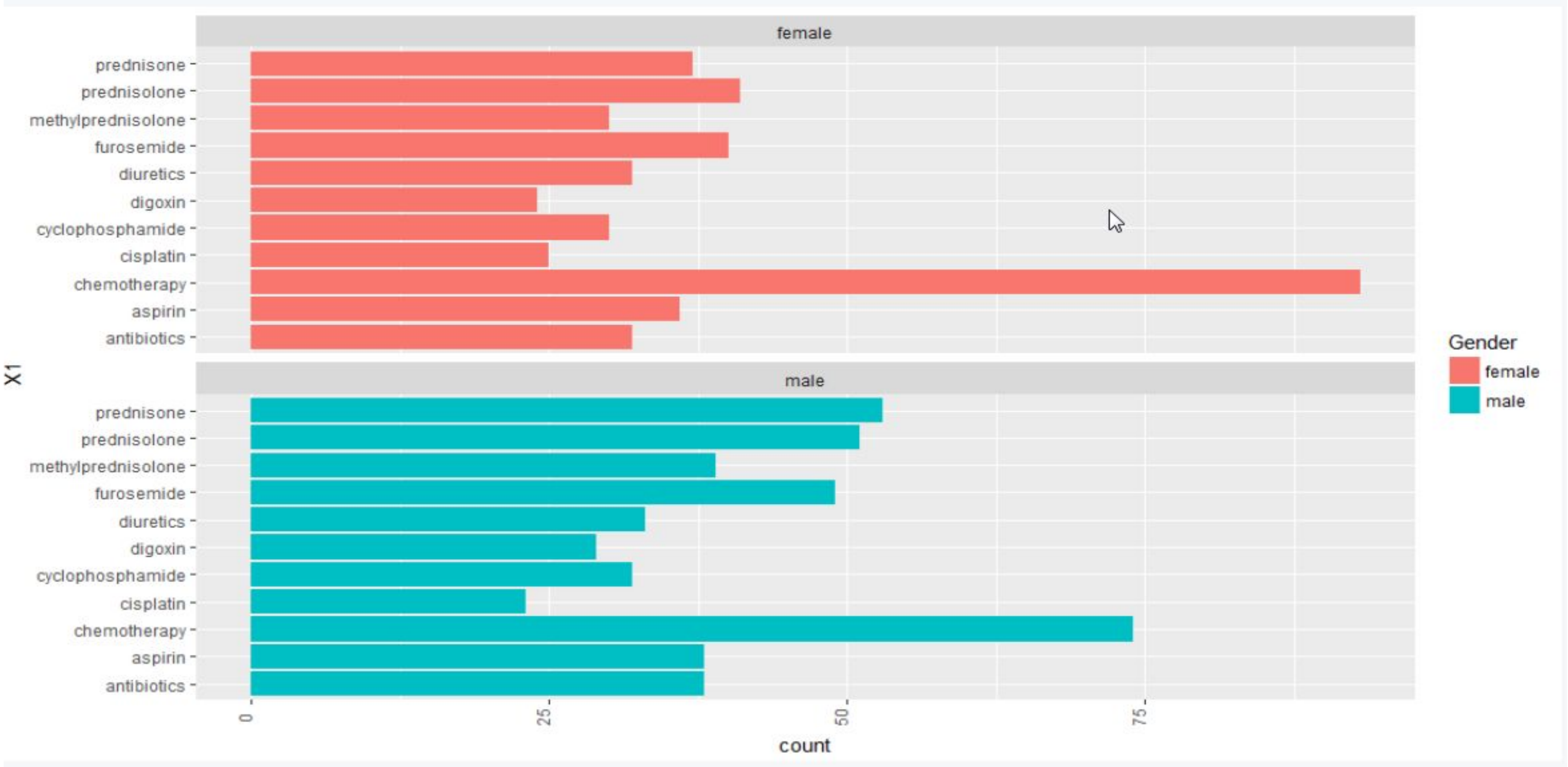
Analysis

Drug Usage:

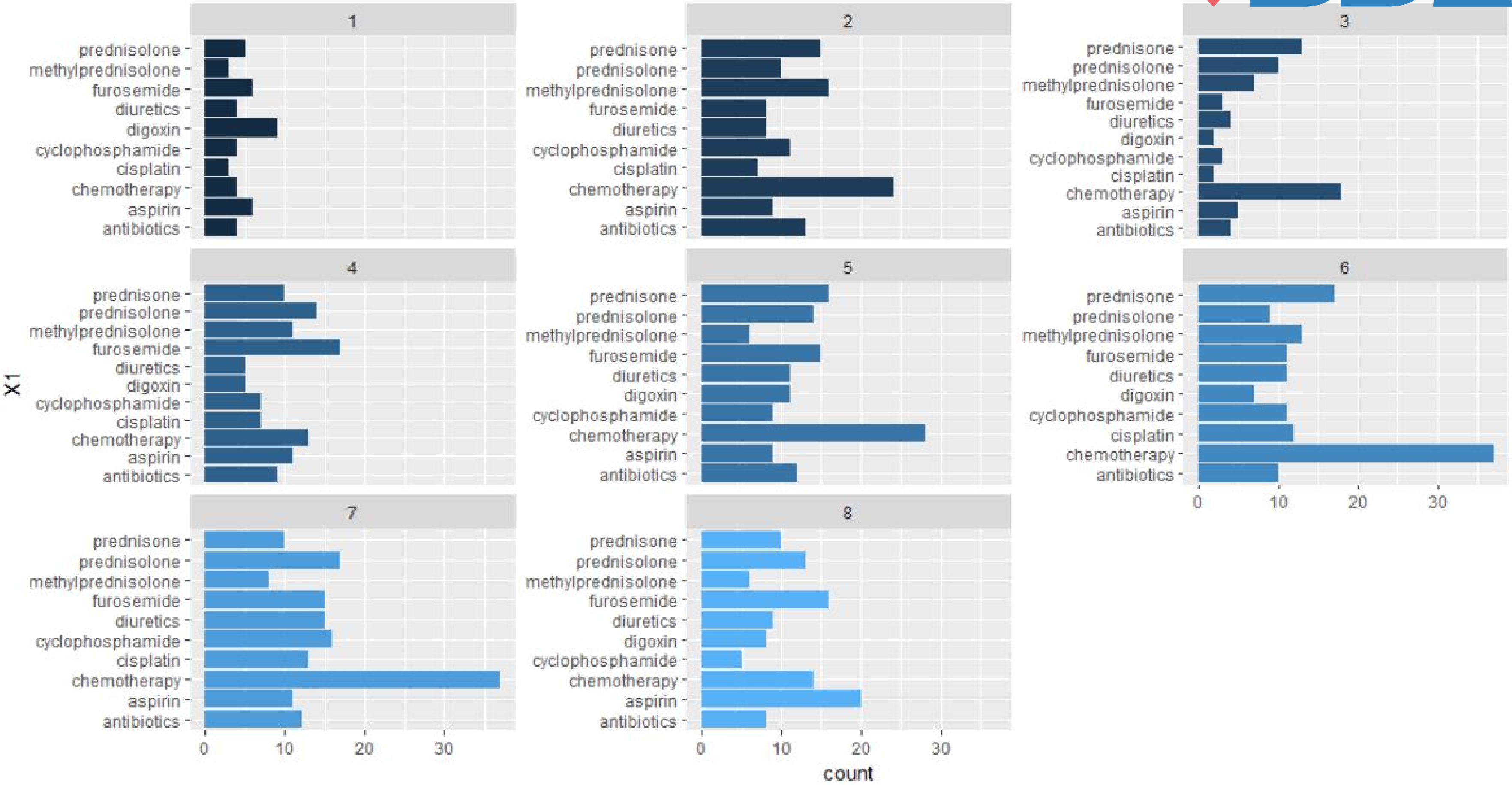
- What are the most frequently used drugs based on gender?
- What are the most frequently used drugs based on age group?

Age Groups:	1	Age: 0 - 2
	2	Age: 2 - 10
	3	Age: 10-20
	...	Age: 20~

Top drugs per gender



Top Drugs per Age Group



Kitu's Workflow: Geographic Data

Clean

1

Extract only relevant information

Analyze

2

Text-mine to extract location

Visualize

3

Plot world choropleth map as well as demographic plots

Extracting only relevant information

CR Number	Institution
CCR1476	edmonton general hospital; university of alberta
CCR579	division of surgical oncology, department of surgery, n...
CCR1633	department of pathology, university of arkansas for m...
CCR1000	department of otolaryngology, head and neck surgery,...
CCR1203	section of neuroradiology, clinic of neurosurgery, chris...
CCR2976	department of pediatrics, yokohama city university sch...
CCR2685	department of radiology, ohshima clinic, sakurada nishi...
CCR1032	retina service, department of ophthalmology, massach...
CCR606	from the departments of general and endocrine surger...
CCR1237	department of ophthalmology, university of arkansas f...
CCR1555	department of medicine, division of infectious diseases, ...
CCR2740	unità di pneumologia e terapia semi-intensiva respiratori...
HCR001	divisions of cardiology and cardiac surgery, university ...
CCR292	cardiology department, complejo hospitalario universita...

Hospital Institution_1	University Institution_2
All	All
edmonton general hospital	university of alberta
division of surgical oncology, department of surgery, n...	NA
department of pathology, university of arkansas for m...	NA
0	NA
section of neuroradiology, clinic of neurosurgery, chris...	NA
retina service, department of ophthalmology, massach...	NA
from the departments of general and endocrine surger...	NA
department of ophthalmology, university of arkansas f...	NA
1	NA
1	NA
service de maladies infectieuses aigues, pôle maladies i...	service de rhumatologie, assistance publique hôpitaux d...
section of oncopathology and regenerative biology, de...	NA
department of orthopaedics, university of illinois at chic...	NA
cardiovascular institute, hospital clínico san carlos, c/o ...	NA

Text-mining to extract location

- Ran into a lot of problems in extracting location
  - Dataframe had both cities and countries
  - Used a dataset of all cities first...3 mil+, so very inefficient
- Re-looked at dataset to find trends
  - Extracted all Country names and US state names
  - Extracted all Country abbreviations
  - Extracted major US cities
- Some still left
  - Misspellings of cities or countries (ex: ilinois, or pekingchina)
  - Manually annotated rest
  - In the future, will extract from a hospital dataset (which is majority of data)

Visualizing the geographic data + other demographics

Let's demo it!

Summary

- Location plays an important part in frequency of case reports
    - Japan!
- Demographics change over country as well as disease group
- This interactive, geographic tool proves useful in finding new insights

Next Steps

- Search “signs and symptoms” within diseases
  - Will require Natural Language Processing
- Explore how case reports are related to each other
- Perhaps standardize frequency of case reports to region’s population