# Bayesian NLP in R on Clinical Text: Predictions from Electronic Health Records

Edward P Flinchem
DaVita Medical Group

Cascadia R Conference
June 8, 2019

edflinchem@gmail.com
Edward.Flinchem@healthcarepartners.com
linkedin.com/in/epf00
425-876-8872

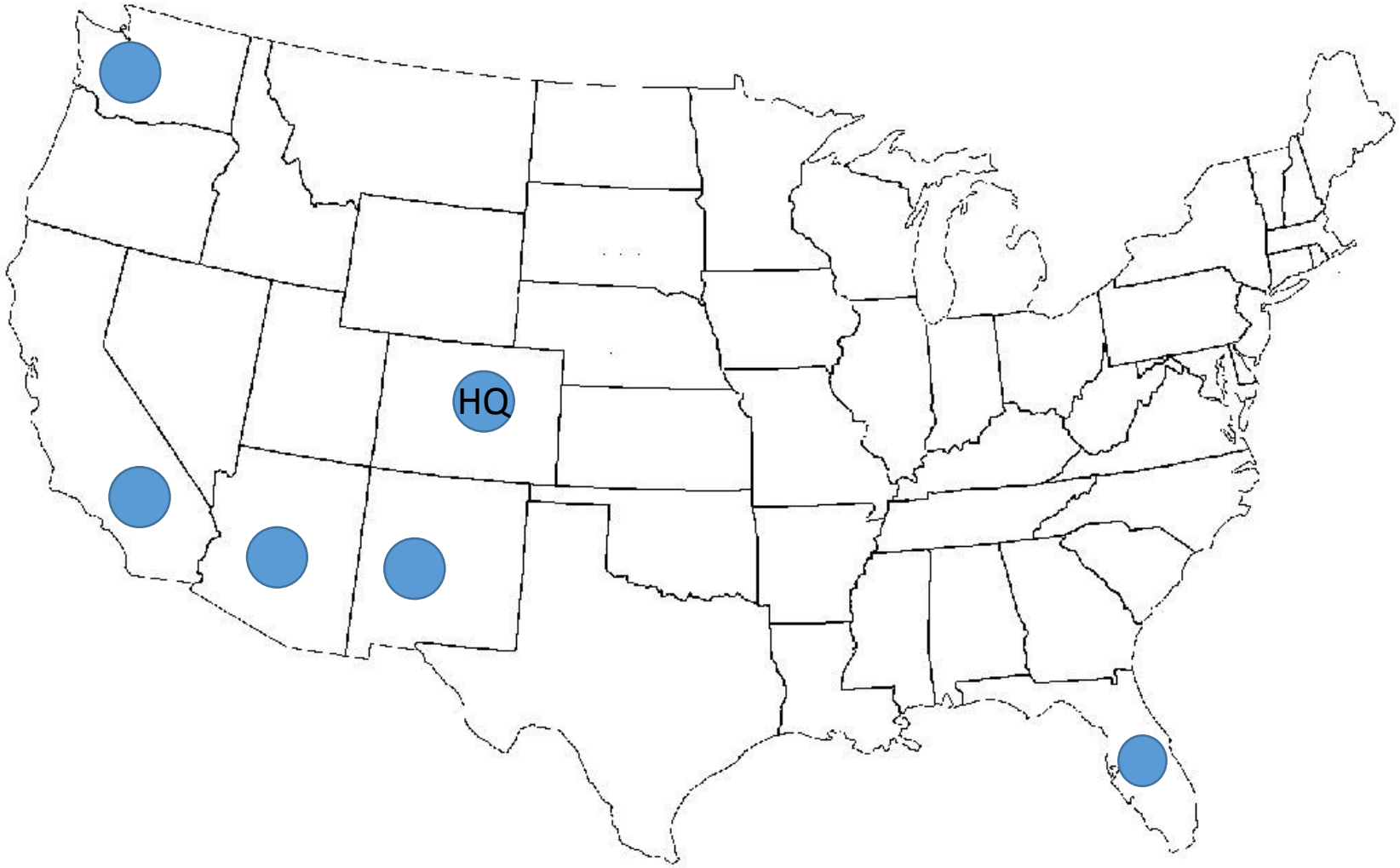


DaVita Medical Group

# What is the DaVita Medical Group?

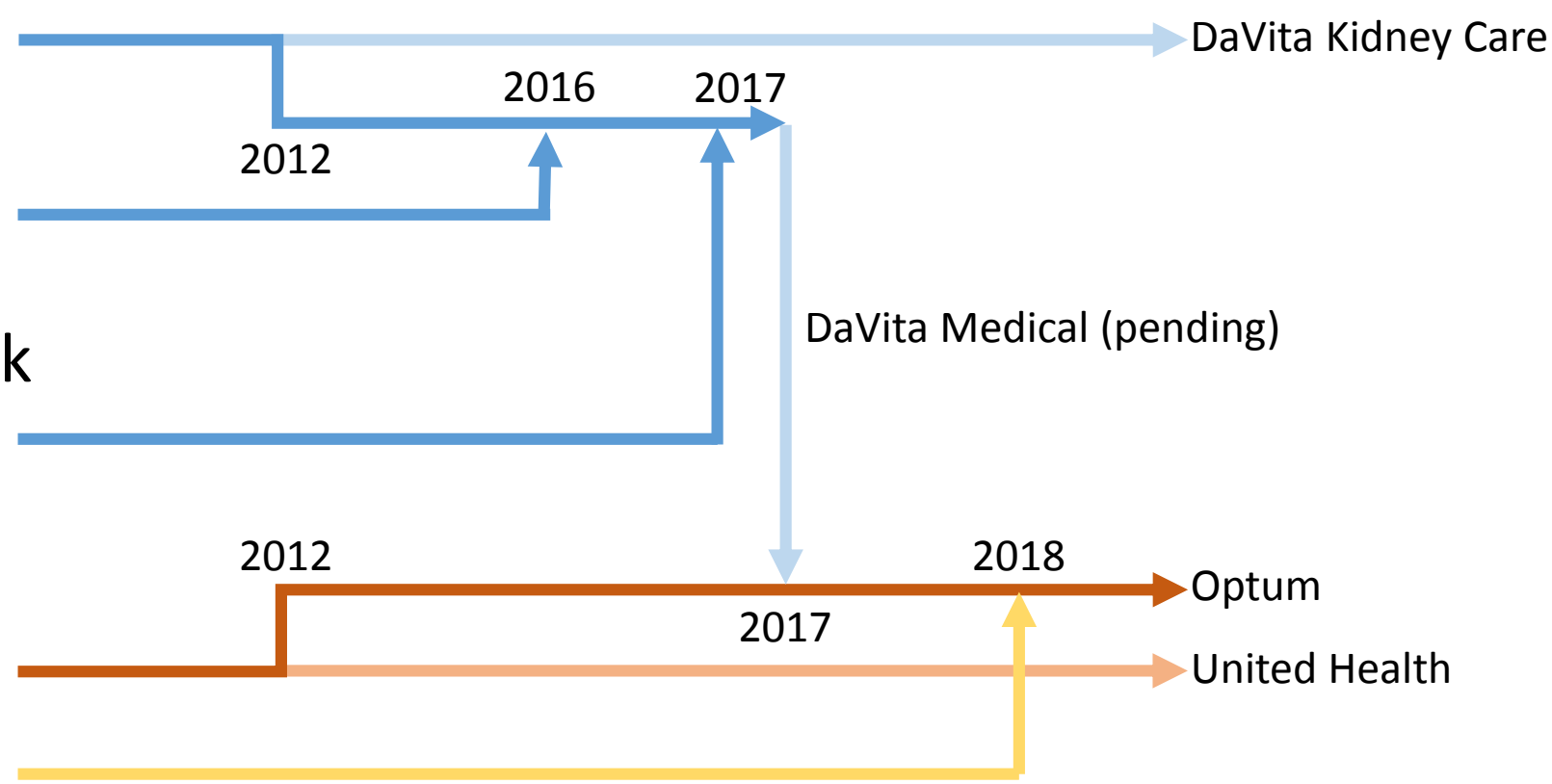
| | |
|---|---|
| Employees | 8-9 Thousand |
| Patients | 1.5 Million |
| Office Visits | Millions/year |

| |
|---|
| Primary Care |
| Specialty Care |
| Ambulatory Surgery |



# Healthcare is consolidating

- DaVita Kidney Care
  - DaVita Medical Group
- Everett Clinic
  - 600 clinicians, 95yo
- NW Physicians Network
  - 1000 clinicians
  - Pierce County
- **United Health** (Optum)
- Seattle Polyclinic
  - 210 clinicians, 101yo



# What? / Why? / How?

- Apply machine learning to predict mortality 90 days in the future
  - In the US, 60% of deaths occur in hospitals
  - But, 70-80% of Americans would prefer to die at home
  - Also, less than 50% of patients needing palliative care receive it
- Synthesize structured and unstructured patient data into a score
  - ~ mortality risk in 90 days
  - Naïve Bayes approach

# The reveal

- Healthcare is a domain where text > numeric data by volume
- Clinical text contains strong signals
  - Ex: 90 day mortality is highly predictable
  - Models outperform humans (by a wide margin)
- R is a great tool to work with text
- Naïve Bayes is highly interpretable w.r.t. text
- Naïve Bayes is on par w/deep learning *for some health problems*

# Machine Learning Sequence

- SQL – query/join tables from a data warehouse
- Wrangle/restructure data (base R)
- Train the model
- Visualize (ggplot + ggrepel)
  - Entire model
  - Populations of patients
  - Single patient, point in time
  - Single patient, change in time

# What is a visit/encounter?

A patient

- Age, Sex

A practitioner

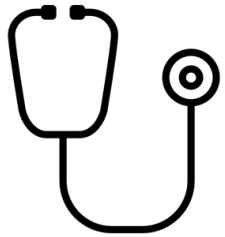
- Specialty

Location

- Office
- Hospital
- Nursing facility
- Patient’s home

1..N diagnoses

- ICD-9
- ICD-10



0..N procedures

- CPT codes

Structured

Unstructured

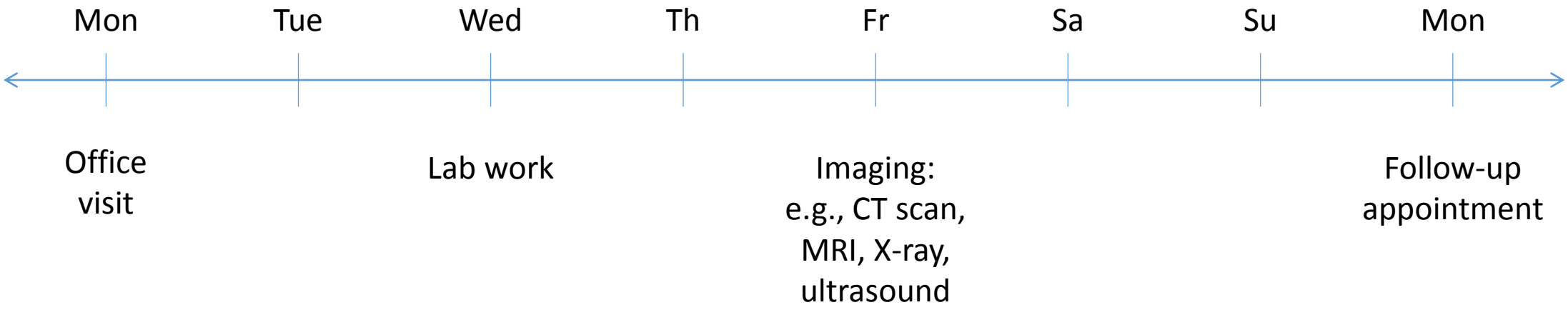
# ICD-What???

1. “International Classification of Diseases”
2. Project of the W.H.O., the World Health Organization
3. ICD-9 was in use 1979-2015 (in the US), ICD-10 thereafter
4. ICD-9 ~ 14,000 diagnostic codes. ICD-10 ~ 70,000

Ex: ICD-9: 488.82 = “Influenza due to identified novel influenza A virus with other respiratory manifestations”

Note: 17 different ICD-9 codes mention “influenza”

# All visits in an 8 day window -- Why 8 days?



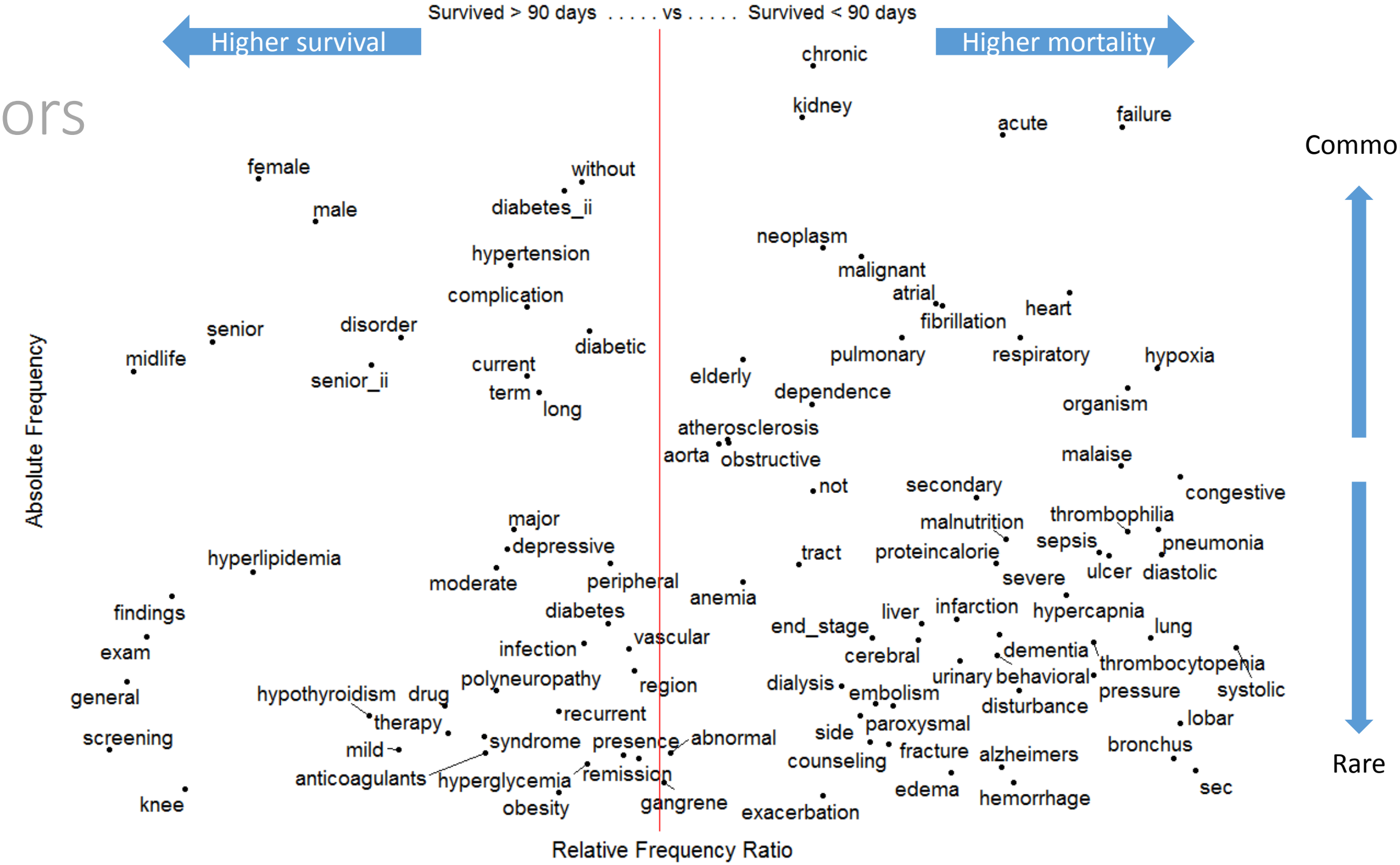
# Naïve Bayesian NLP

Word Frequency Analysis + Information Theory

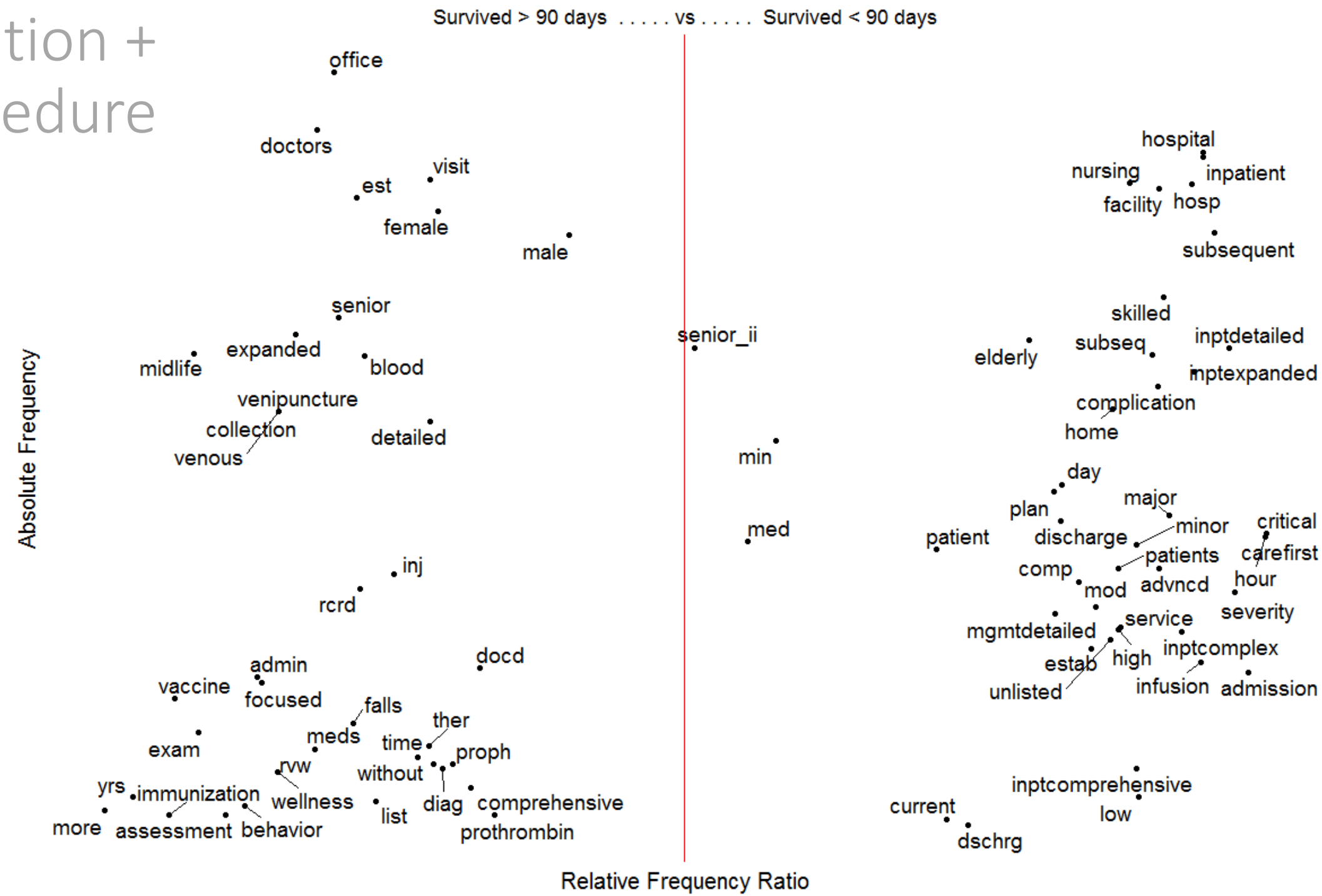
# Word Counts w/Binary Outcome (T vs F)

1. Data are labeled: True/False (e.g., Mortality in 90 days or not)
2. Find top 100 words *regardless* of outcome (T or F)
3. Recount those words *by outcome*: pT[word], pF[word]
4. Visualize
   a. Vertical = Log(probability overall)
   b. Horizontal = Log(probability *ratio*: pT/pF)
   c. So, Left vs Right dispersion → information in the words

# DX Factors


Survived > 90 days . . . . . vs . . . . . Survived < 90 days
Higher survival
Higher mortality
Common
Rare
Absolute Frequency
Relative Frequency Ratio
female
male
without
diabetes_ii
hypertension
complication
disorder
senior
midlife
senior_ii
diabetic
current
term
long
chronic
kidney
acute
failure
neoplasm
malignant
atrial
fibrillation
heart
pulmonary
respiratory
hypoxia
elderly
dependence
organism
atherosclerosis
aorta
obstructive
malaise
not
secondary
congestive
malnutrition
thrombophilia
sepsis
pneumonia
tract
proteincalorie
ulcer
diastolic
severe
major
depressive
hyperlipidemia
moderate
peripheral
anemia
findings
diabetes
liver
infarction
hypercapnia
lung
exam
infection
vascular
end_stage
cerebral
dementia
thrombocytopenia
general
polyneuropathy
region
dialysis
urinary
behavioral
pressure
systolic
hypothyroidism
drug
therapy
recurrent
embolism
disturbance
lobar
screening
syndrome
presence
abnormal
paroxysmal
side
bronchus
mild
fracture
alzheimers
counseling
anticoagulants
hyperglycemia
remission
edema
hemorrhage
sec
knee
obesity
gangrene
exacerbation

# Location + Procedure



# R rocks text! (5 useful patterns)

1. Combining
2. Splitting
3. Cleaning
4. Counting
5. Operating on sets

# Combining -- via “aggregate()”

| Encounter_id | ICD |
|---|---|
| 96894926 | UNSPECIFIED ATRIAL FIBRILLATION |
| 96894926 | LONG TERM (CURRENT) USE OF ANTICOAGULANTS |
| 96895042 | ENCNTR FOR GENERAL ADULT MEDICAL EXAM W/O ABNORMAL FINDINGS |
| 96895044 | ACTINIC KERATOSIS. |
| 96895044 | DISORDER OF THE SKIN AND SUBCUTANEOUS TISSUE, UNSPECIFIED |

```
> aggregate(d$ICD, by = list(d$Encounter_id), FUN = paste)
 Group.1                                                                    x
1 96894926    UNSPECIFIED ATRIAL FIBRILLATION LONG TERM (CURRENT) USE OF ANTICOAGULANTS
2 96895042                      ENCNTR FOR GENERAL ADULT MEDICAL EXAM W/O ABNORMAL FINDINGS
3 96895044 ACTINIC KERATOSIS. DISORDER OF THE SKIN AND SUBCUTANEOUS TISSUE, UNSPECIFIED
```

# Splitting (aka, “tokenize”) -- via “strsplit()”

```
> strsplit("ACTINIC KERATOSIS DISORDER OF THE SKIN AND SUBCUTANEOUS TISSUE UNSPECIFIED", " ")[[1]]
 [1] "ACTINIC"      "KERATOSIS"    "DISORDER"     "OF"           "THE"          "SKIN"
 [7] "AND"          "SUBCUTANEOUS" "TISSUE"       "UNSPECIFIED"
```

# Cleaning -- via “gsub()”

gsub is short for “global substitute”

```
> s <- "ACTINIC KERATOSIS. DISORDER OF THE SKIN AND SUBCUTANEOUS TISSUE, UNSPECIFIED"
> gsub("[^a-z]", " ", tolower(s))                    # restrict character set to a..z
[1] "actinic keratosis  disorder of the skin and subcutaneous tissue  unspecified"
```

# Counting -- via “table()”

```
> words <- strsplit("Mary had a little lamb little lamb little lamb Mary had a little lamb whose fleece was white as snow", " ")[[1]]
> words
 [1] "Mary"   "had"    "a"      "little" "lamb"   "little" "lamb"   "little" "lamb"   "Mary"   "had"
[12] "a"      "little" "lamb"   "whose"  "fleece" "was"    "white"  "as"     "snow"
> table(words)
words
     a     as fleece    had   lamb little   Mary   snow    was  white  whose
     2      1      1      2      4      4      2      1      1      1      1
```

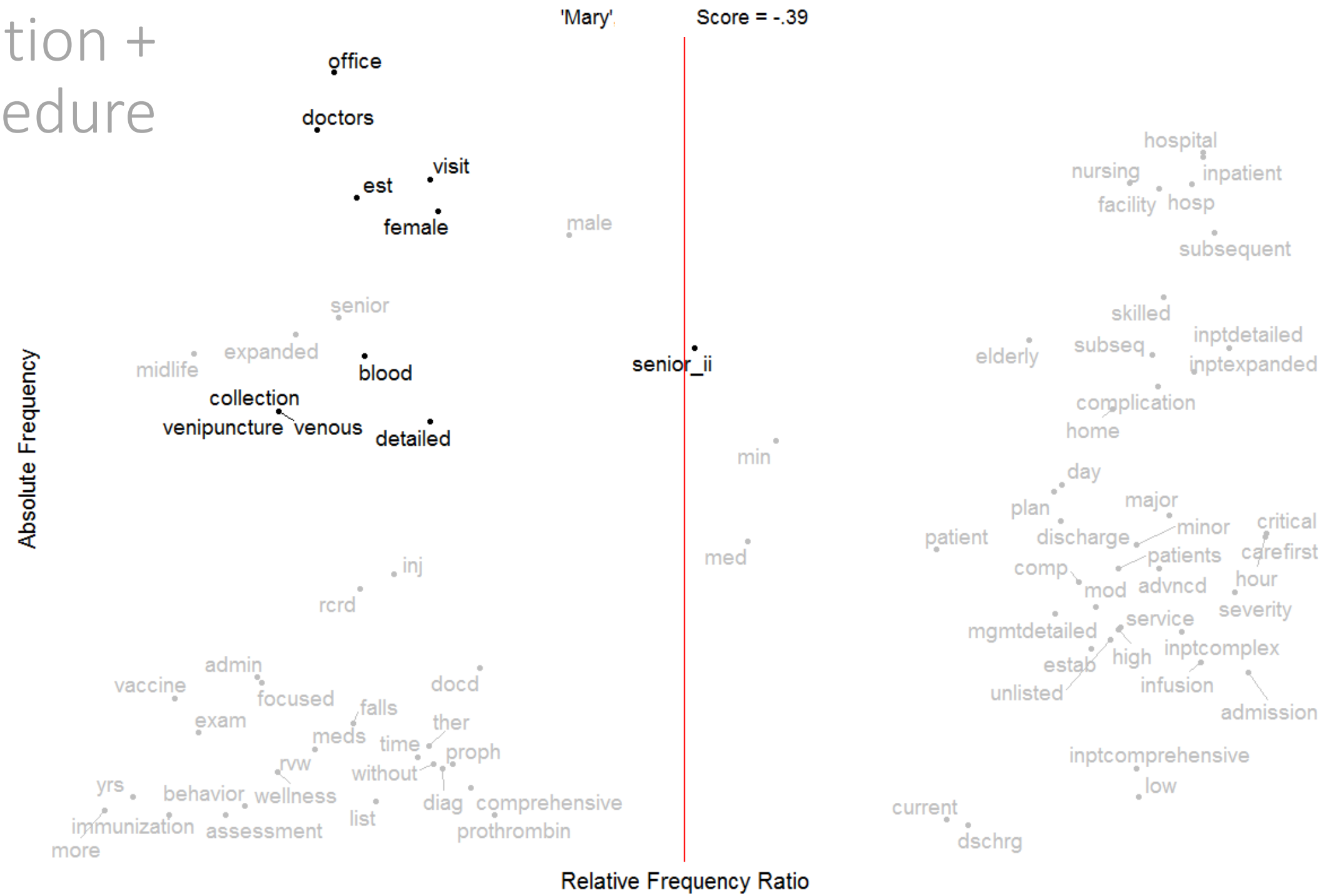
# Operating on sets -- via “%in%”

```
> s
 [1] "ACTINIC"      "KERATOSIS"    "DISORDER"     "OF"           "THE"          "SKIN"
 [7] "AND"          "SUBCUTANEOUS" "TISSUE"       "UNSPECIFIED"
>
> stops <- c("OF", "THE", "AND")
>
> s %in% stops                    # which things on the left are members of the set on the right?
 [1] FALSE FALSE FALSE  TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE
>
> s[ !(s %in% stops) ]
[1] "ACTINIC"      "KERATOSIS"    "DISORDER"     "SKIN"         "SUBCUTANEOUS" "TISSUE"
[7] "UNSPECIFIED"
```

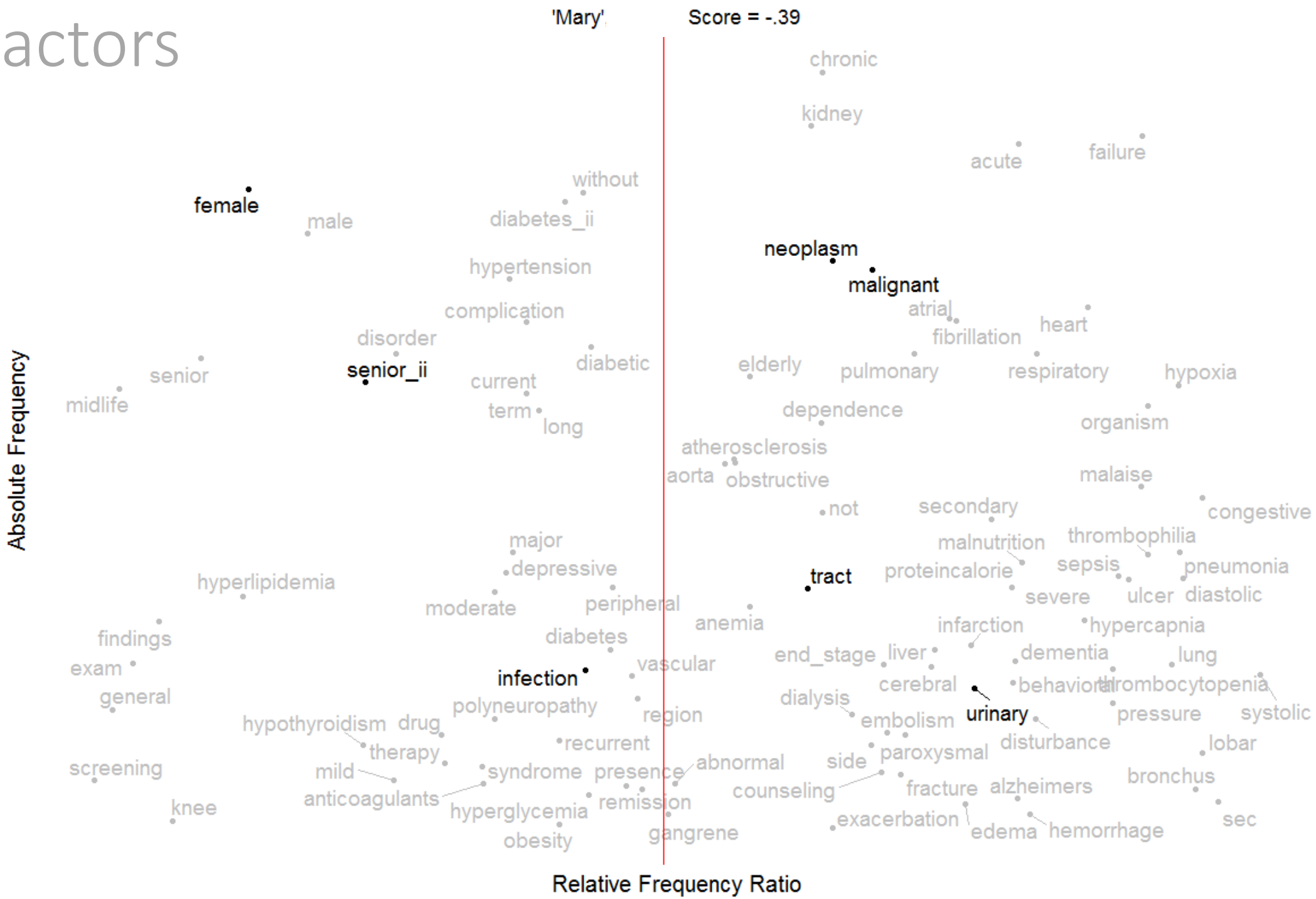
# How to score 1 patient?

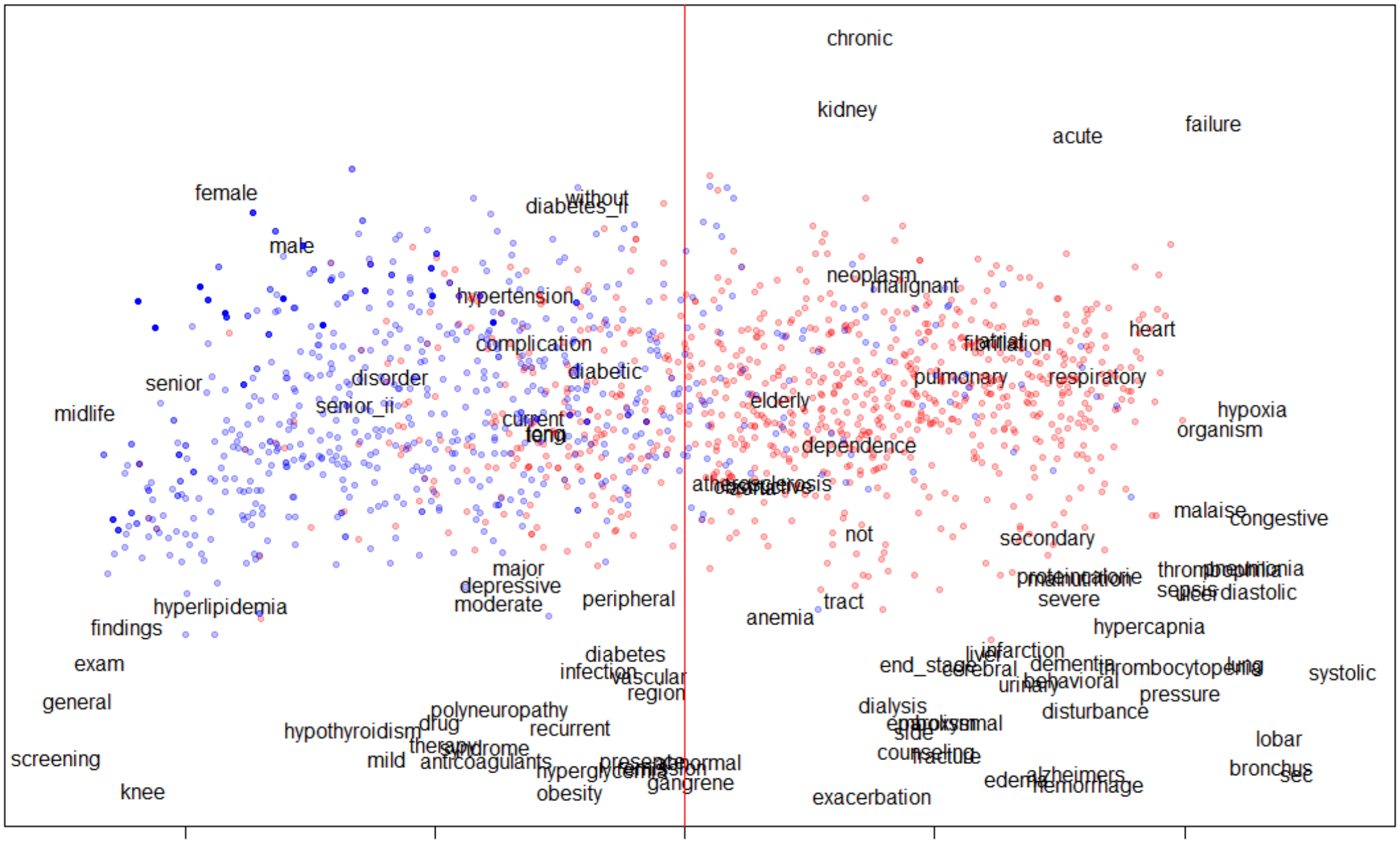
An example

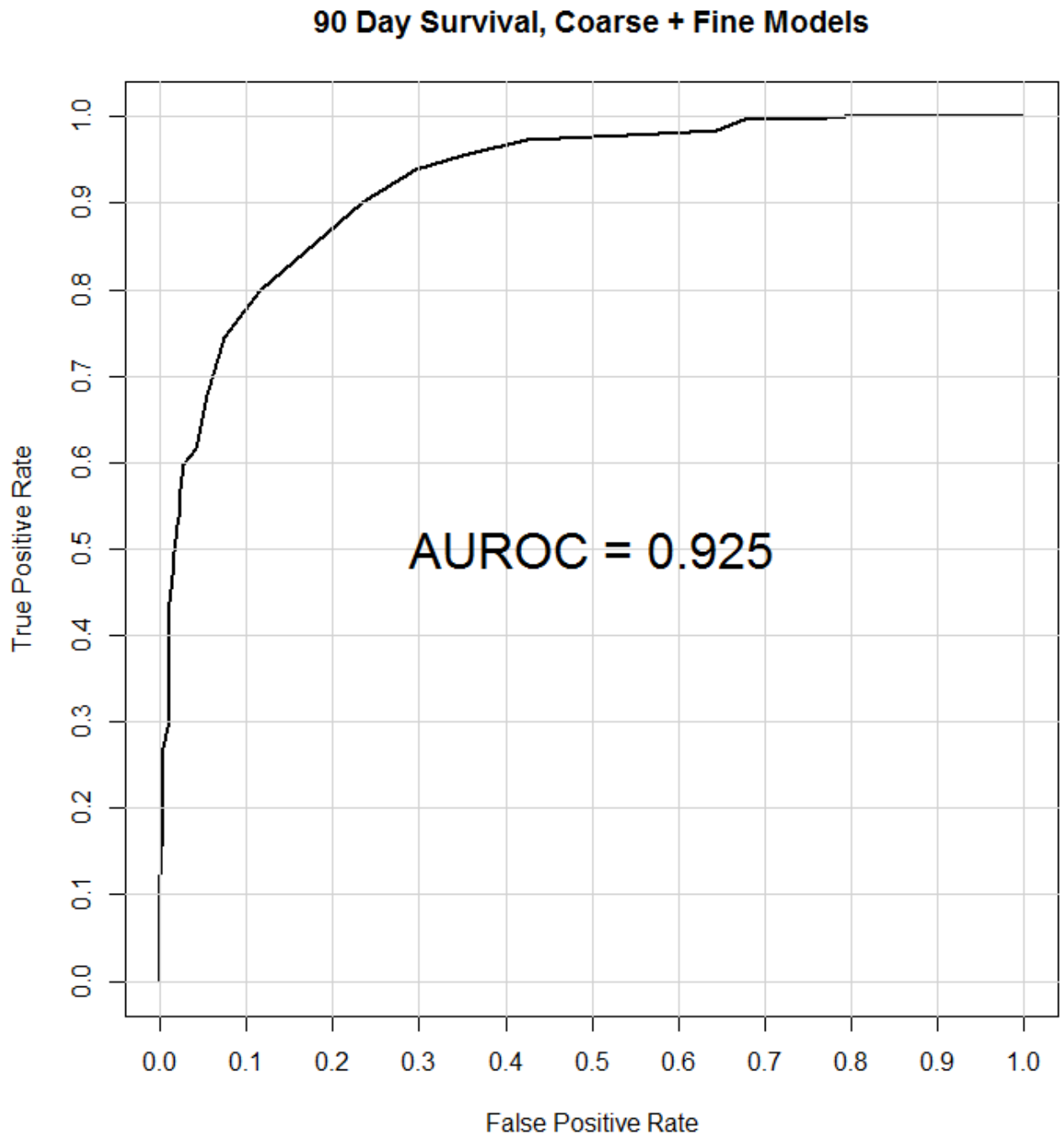
# Location + Procedure



# DX Factors


'Mary'
Score = -.39
Absolute Frequency
Relative Frequency Ratio
female
male
without
diabetes_ii
hypertension
complication
disorder
senior_ii
diabetic
senior
current
midlife
term
long
chronic
kidney
acute
failure
neoplasm
malignant
atrial
heart
fibrillation
elderly
pulmonary
respiratory
hypoxia
dependence
organism
atherosclerosis
aorta
obstructive
malaise
not
secondary
congestive
malnutrition
thrombophilia
tract
proteincalorie
sepsis
pneumonia
severe
ulcer
diastolic
major
depressive
hyperlipidemia
moderate
peripheral
anemia
infarction
hypercapnia
findings
diabetes
exam
vascular
infection
end_stage
liver
dementia
lung
general
cerebral
behavioral
thrombocytopenia
dialysis
urinary
polyneuropathy
region
embolism
pressure
systolic
hypothyroidism
drug
recurrent
disturbance
lobar
therapy
paroxysmal
screening
mild
syndrome
presence
abnormal
side
bronchus
counseling
fracture
alzheimers
anticoagulants
remission
knee
hyperglycemia
exacerbation
sec
edema
hemorrhage
obesity
gangrene


Survived > 90 days . . . . . vs . . . . . Survived < 90 days
chronic
kidney
failure
acute
female
without
diabetes_ii
male
neoplasm
malignant
hypertension
heart
complication
fibrillation
disorder
diabetic
pulmonary
respiratory
senior
elderly
hypoxia
midlife
senior_ii
current
organism
dependence
malaise
congestive
not
secondary
major
depressive
moderate
peripheral
severe
hyperlipidemia
tract
anemia
findings
hypercapnia
diabetes
infection
vascular
region
exam
general
dialysis
pressure
disturbance
systolic
polyneuropathy
drug
recurrent
hypothyroidism
lobar
screening
mild
anticoagulants
counseling
fracture
bronchus
knee
obesity
gangrene
exacerbation
hemorrhage


90 Day Survival, Coarse + Fine Models
True Positive Rate
False Positive Rate
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
AUROC = 0.925

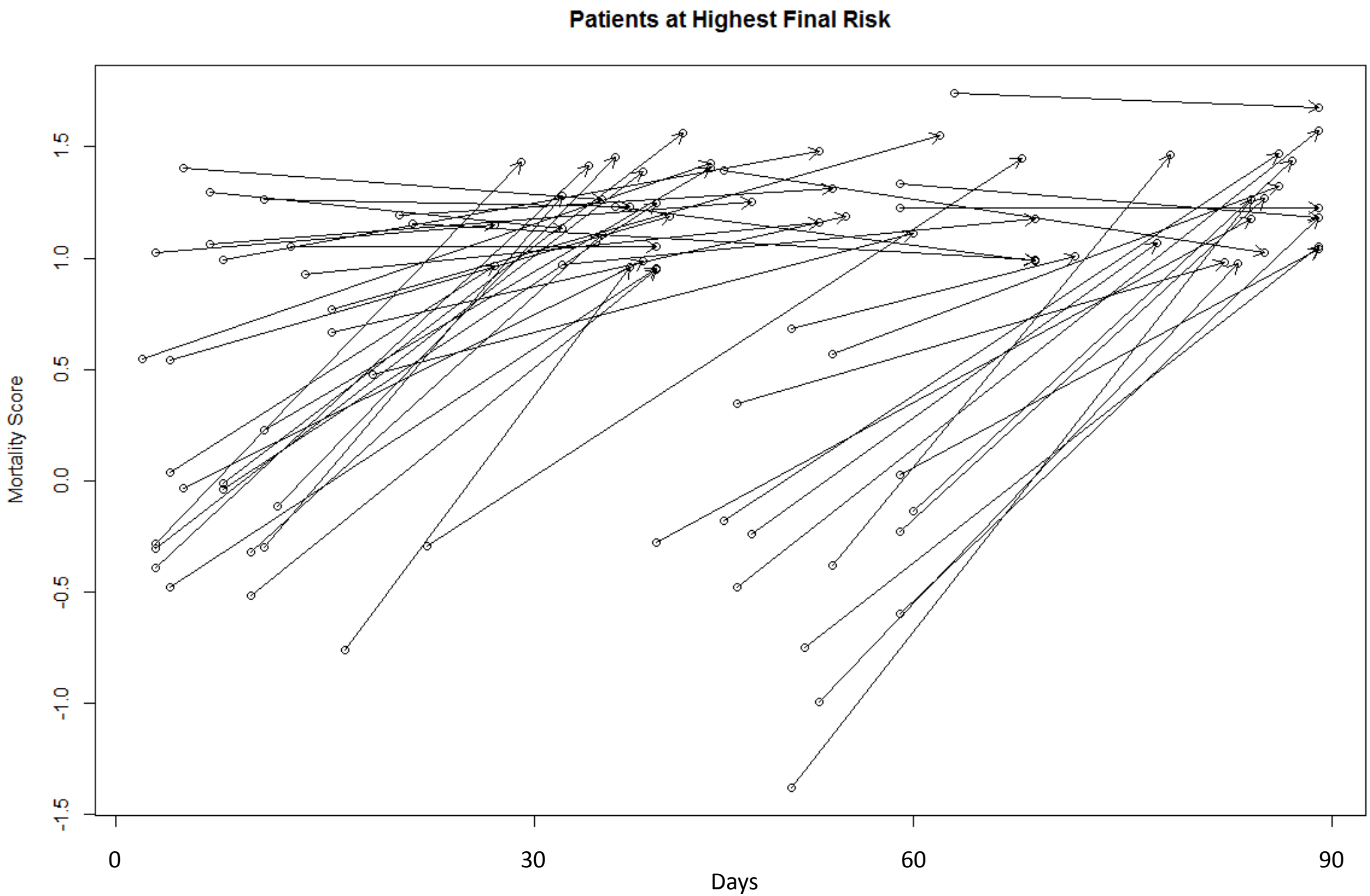
# Mortality Prediction Literature

| | Reference | Target (days) | Population | Method | AUROC |
|---|---|---|---|---|---|
| 1.1 | Elfilky, et al., 2018 | 30 | Chemotherapy patients | xgboost | .94 |
| 1.2 | | 180 | Chemotherapy patients | xgboost | .87 |
| 2.1 | Ahmad, et al., 2018 | 180 | All | xgboost | .79 |
| 2.2 | | 180 | All | Clinicians | .71 |
| 2.3 | | 90 | Hospital patients w/heart failure | xgboost | .766 |
| 2.4 | | 120 | Hospital patients w/heart failure | xgboost | .879 |
| 3.1 | Avati, et al., 2017 | 90-365 | All | Deep learning | .93 |
| 3.2 | | | Hospital patients | Deep learning | .87 |
| 4 | White, et al., 2016 | Various | Various (a review article) | Clinicians | .74 - .78 |
| 5 | Goldstein, et al., 2016 | 90 | Kidney Dialysis patients | Logistic regression | .72 - . 76 |
| 6 | Current work | 90 | All | Naïve Bayes/NLP | .925 |

# Shifts Over Time

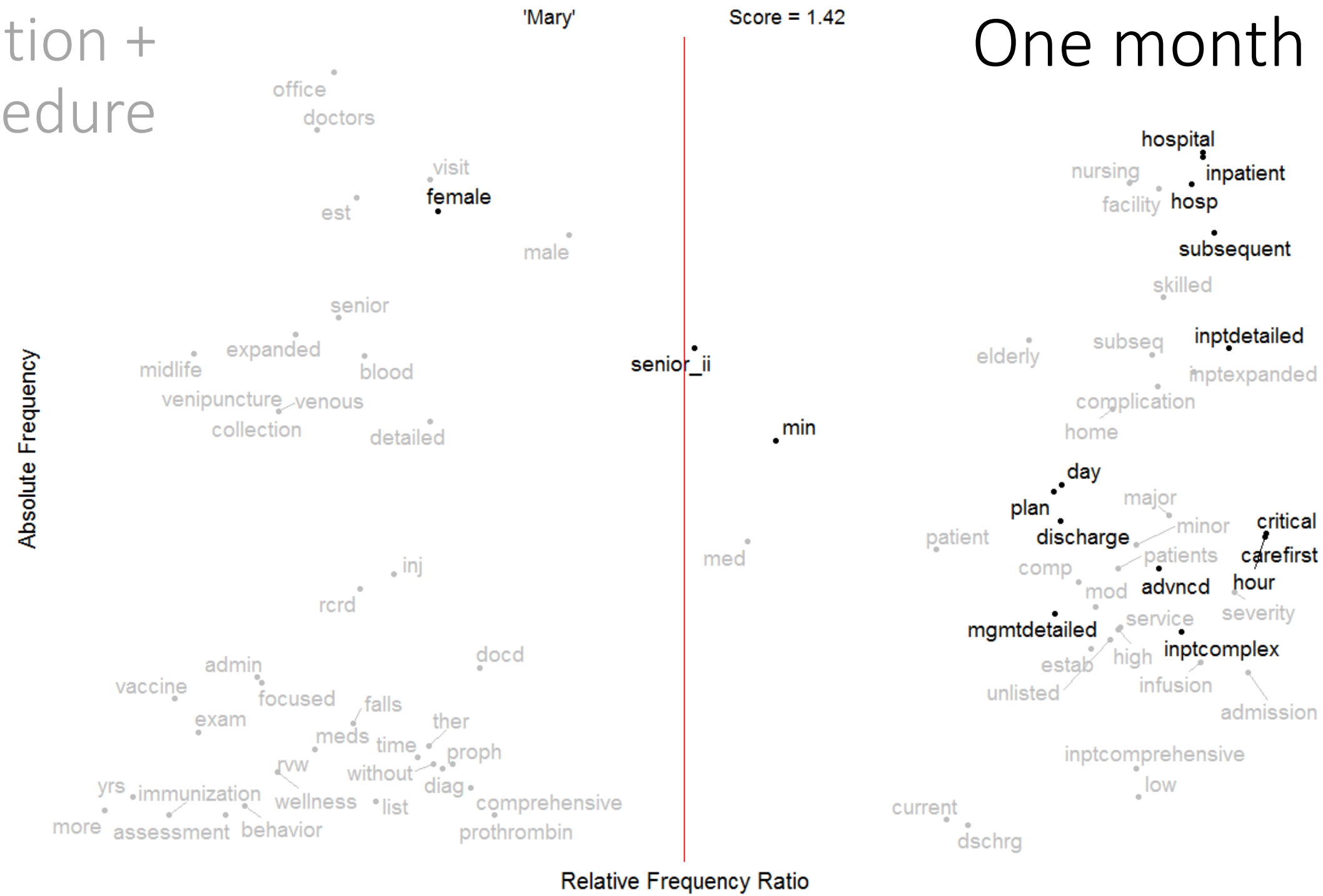
## Of patients w/2 episodes of care:

1. Max increase in risk
2. Max decline in risk
3. Who joined the acute population
4. Who joined the “well” (lower risk) population
5. Who left the “well” population
6. Who left the acute population


Patients at Highest Final Risk
Mortality Score
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
0
30
60
90
Days

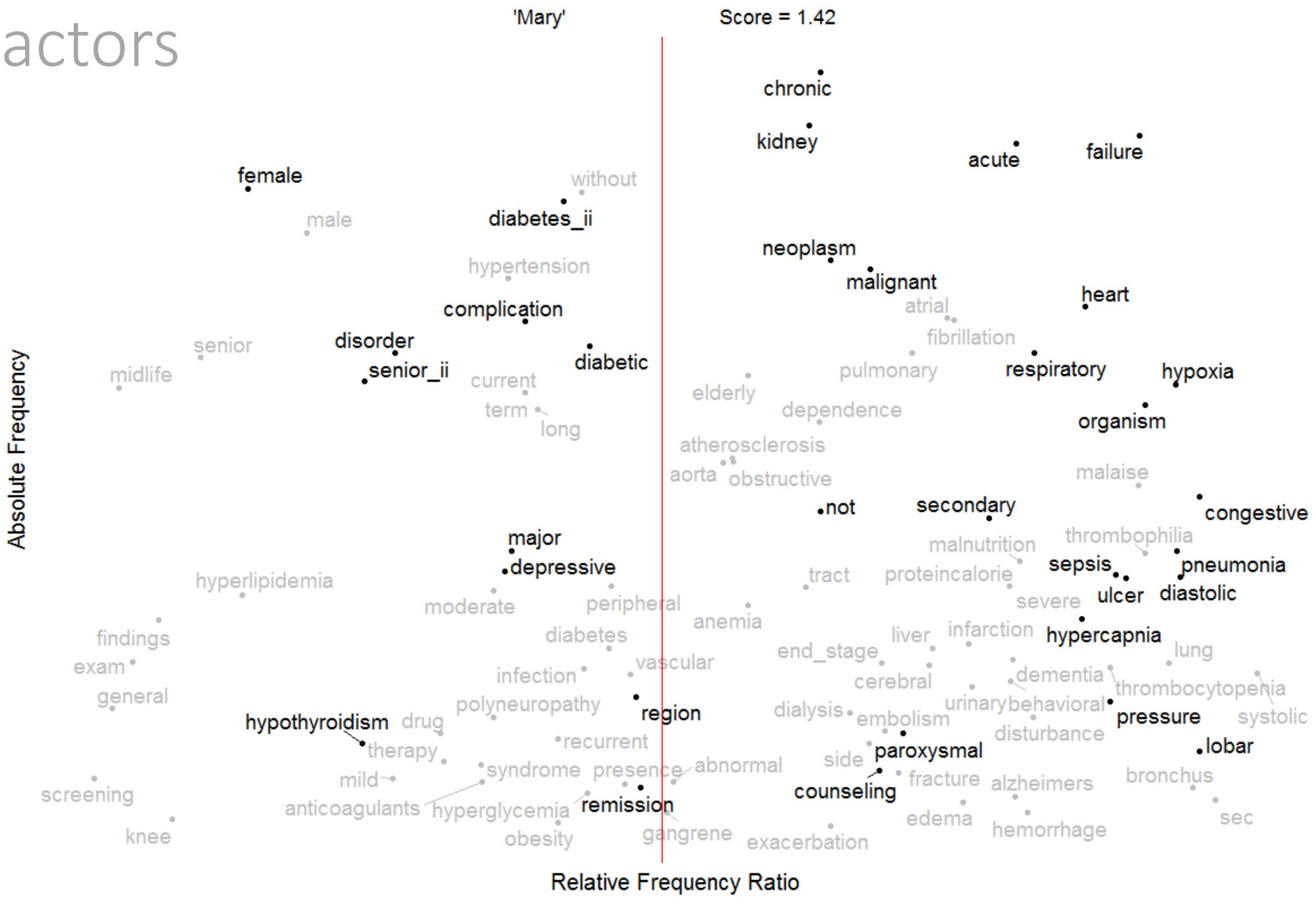
# Single Patient Summaries

# Location + Procedure

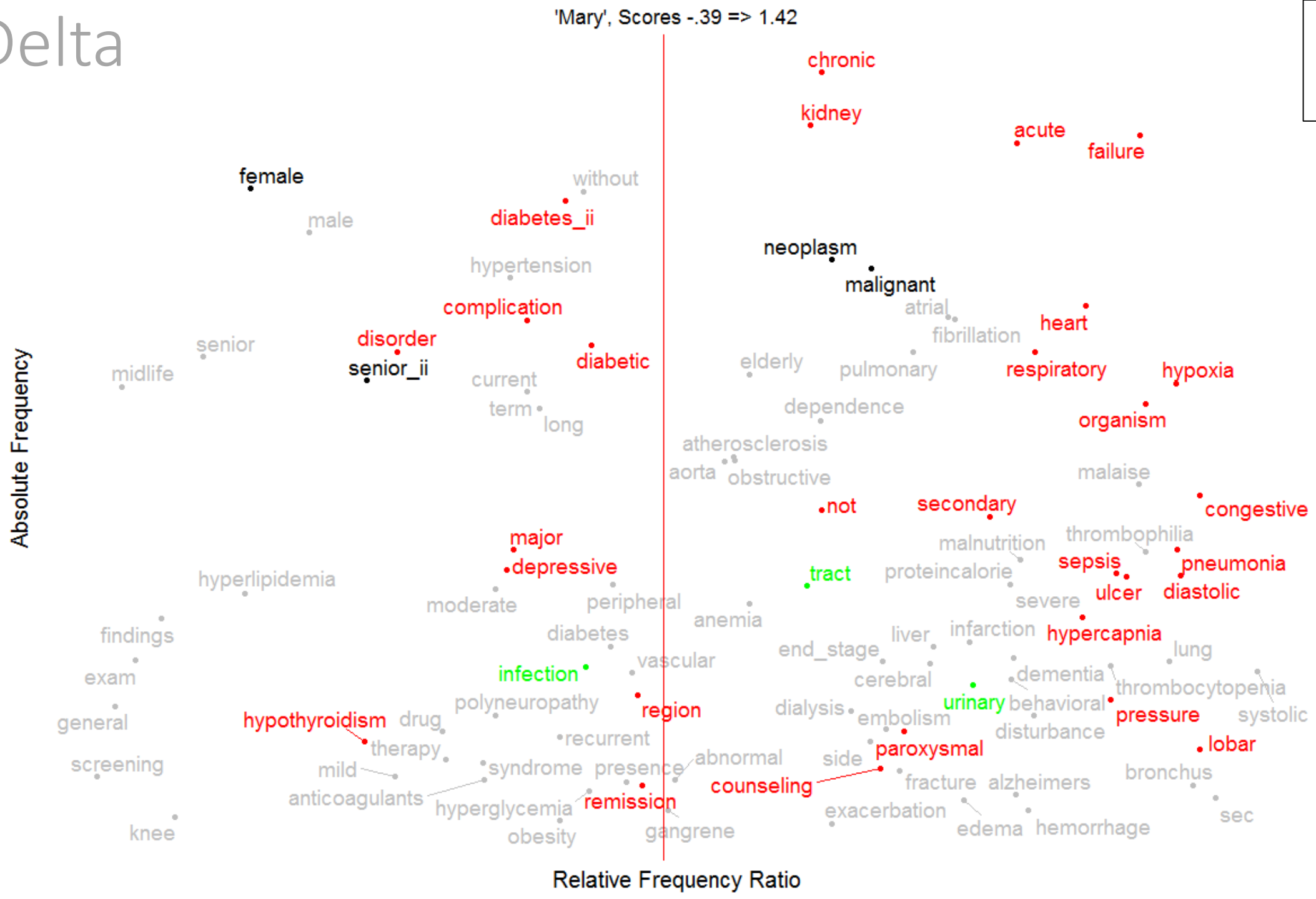



# One month later

# DX Factors


'Mary'
Score = 1.42
chronic
kidney
acute
failure
female
without
male
diabetes_ii
neoplasm
hypertension
malignant
heart
complication
atrial
fibrillation
senior
disorder
diabetic
pulmonary
respiratory
hypoxia
midlife
senior_ii
current
elderly
dependence
term
long
organism
atherosclerosis
aorta
obstructive
malaise
not
secondary
congestive
major
malnutrition
thrombophilia
depressive
sepsis
pneumonia
hyperlipidemia
tract
proteincalorie
ulcer
diastolic
moderate
peripheral
severe
anemia
hypercapnia
findings
diabetes
liver
infarction
lung
exam
vascular
end_stage
infection
cerebral
dementia
thrombocytopenia
general
polyneuropathy
region
dialysis
urinary
behavioral
pressure
systolic
hypothyroidism
drug
embolism
disturbance
therapy
recurrent
paroxysmal
lobar
side
abnormal
syndrome
presence
fracture
alzheimers
bronchus
mild
screening
counseling
remission
edema
sec
anticoagulants
hyperglycemia
hemorrhage
knee
obesity
gangrene
exacerbation
Absolute Frequency
Relative Frequency Ratio

# DX Delta


'Mary', Scores -.39 => 1.42
Resolved
Continuing
New
Absolute Frequency
Relative Frequency Ratio
chronic
kidney
acute
failure
female
without
male
diabetes_ii
neoplasm
hypertension
malignant
complication
atrial
heart
fibrillation
disorder
senior
diabetic
elderly
pulmonary
respiratory
hypoxia
senior_ii
midlife
current
dependence
term
long
organism
atherosclerosis
aorta
obstructive
malaise
not
secondary
congestive
major
malnutrition
thrombophilia
depressive
tract
proteincalorie
sepsis
pneumonia
hyperlipidemia
ulcer
diastolic
moderate
peripheral
severe
anemia
findings
diabetes
liver
infarction
hypercapnia
end_stage
lung
exam
vascular
infection
cerebral
dementia
thrombocytopenia
polyneuropathy
urinary
behavioral
general
region
dialysis
embolism
pressure
systolic
hypothyroidism
drug
disturbance
recurrent
therapy
paroxysmal
lobar
screening
abnormal
side
mild
syndrome
presence
fracture
alzheimers
bronchus
counseling
anticoagulants
remission
exacerbation
sec
hyperglycemia
knee
edema
hemorrhage
obesity
gangrene

# Unique Characteristics of NB/NLP

- Leverages unstructured data, *the most widely available kind*
- Predictive skill can compete with neural networks in some domains
- Not compute intensive → agility
- Fully “transparent” + interpretable model “reasoning”
  - 2D Word frequency visualization
  - optional before/after temporal change
- Framework for multiple use cases
  - Hospital admissions, when to discharge, future total cost of care, ...

# Future directions

- Socio-demographics
  - Food security, transportation access, isolation, mobility, etc.
- Medications
- Lab results
- Radiology/imaging reports
- Pathology reports
- Medical devices and equipment orders
- Historical time-at-risk
- Optimizing the prediction window for intervention

# END

Questions?

Please reach out to me at:

linkedin.com/in/epf00 edflinchem@gmail.com 425-876-8872