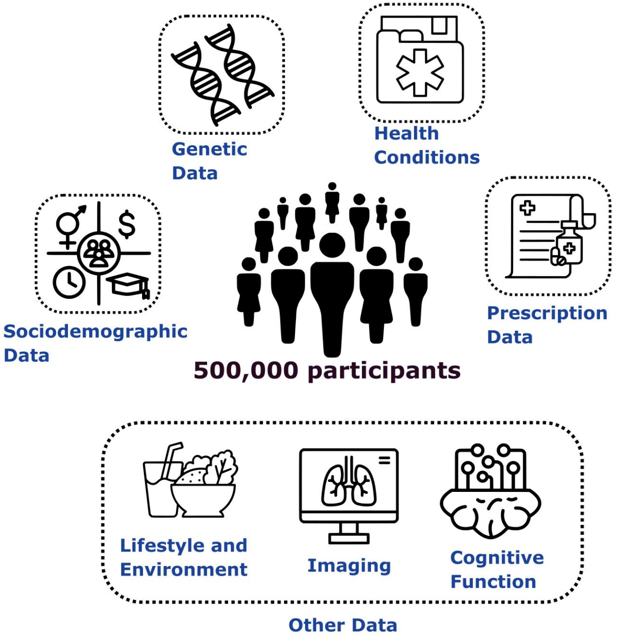
An exploratory data science approach: Understanding the OMOP data for use as phenotype in genotype-phenotype association studies

by **Xinlu Shi**

supervisor: **Alexander Hauser, Jakob Madsen**

UK Biobank



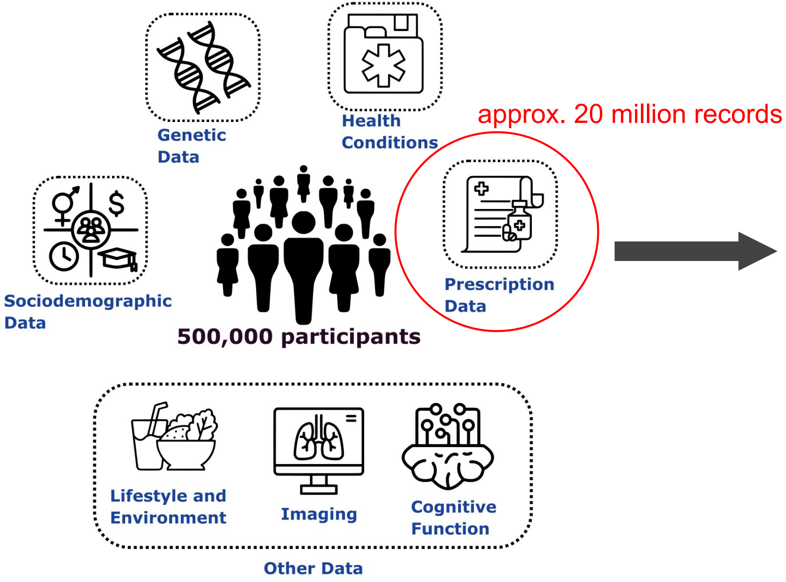
Motivation: drug use pattern → phenotype → genotype

Related Works

- Troels Siggaard et al. "**Disease trajectory browser for exploring temporal, population-wide disease progression patterns in 7.2 million Danish patients**". In: Nature Communications 11.1 (2020), p. 4952.
  merge linear trajectories into disease trajectory networks; exploring patterns of disease progression
- Tuomo Kiiskinen, Pyry Helkkula, Kristi Krebs, et al. " **Genetic predictors of lifelong medication-use patterns in cardiometabolic diseases**". In: Nature Medicine 29.1 (2023), pp. 209–218.
  genetic associations with drug adherence and switching patterns in cardiometabolic diseases
- Bjarni V. Halldorsson, Hannes P. Eggertsson, Kristjan H. S. Moore, et al. "**The sequences of 150,119 genomes in the UK Biobank**". In: Nature 607.732 (2022), pp. 732–740.
  rare and common genetic variants influencing drug response → refined models of pharmacogenomics and personalized medicine

Motivation: drug use pattern → phenotype → genotype

UK Biobank



Genetic Data

Health Conditions

approx. 20 million records

Sociodemographic Data

500,000 participants

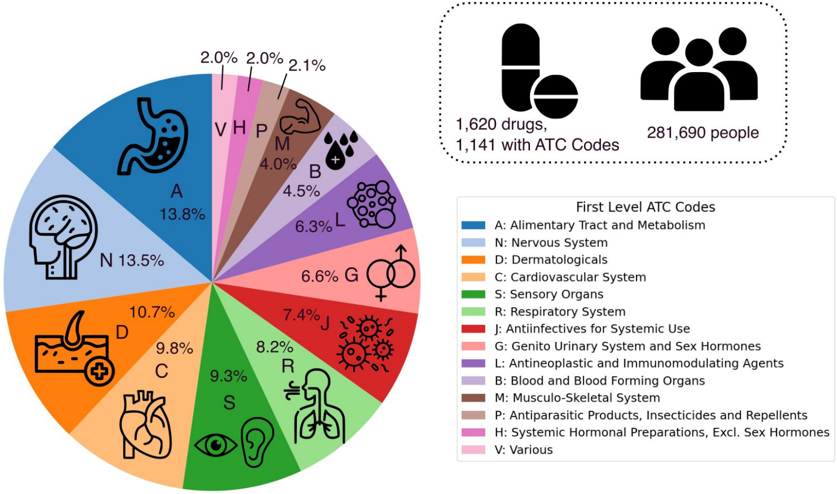
Prescription Data

Lifestyle and Environment

Imaging

Cognitive Function

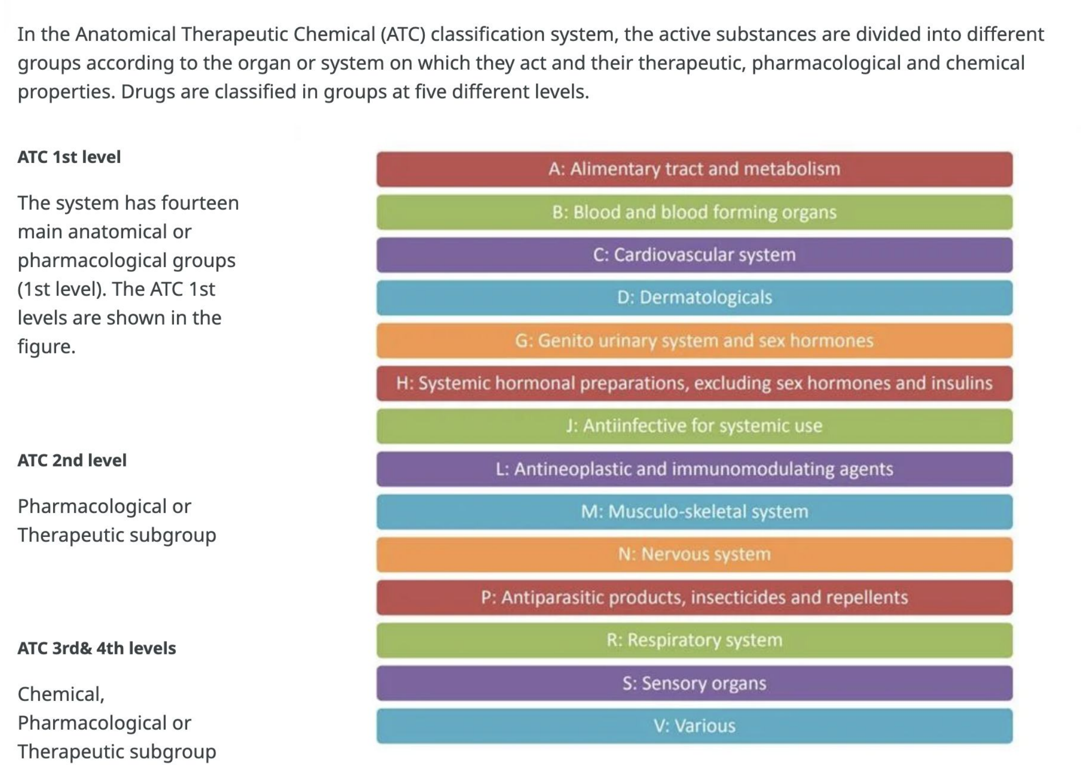
Other Data



1,620 drugs,
1,141 with ATC Codes

281,690 people

ATC code



In the Anatomical Therapeutic Chemical (ATC) classification system, the active substances are divided into different groups according to the organ or system on which they act and their therapeutic, pharmacological and chemical properties. Drugs are classified in groups at five different levels.

ATC 1st level

The system has fourteen main anatomical or pharmacological groups (1st level). The ATC 1st levels are shown in the figure.

A: Alimentary tract and metabolism
B: Blood and blood forming organs
C: Cardiovascular system
D: Dermatologicals
G: Genito urinary system and sex hormones
H: Systemic hormonal preparations, excluding sex hormones and insulins
J: Antiinfective for systemic use
L: Antineoplastic and immunomodulating agents
M: Musculo-skeletal system
N: Nervous system
P: Antiparasitic products, insecticides and repellents
R: Respiratory system
S: Sensory organs
V: Various

ATC 2nd level

Pharmacological or Therapeutic subgroup

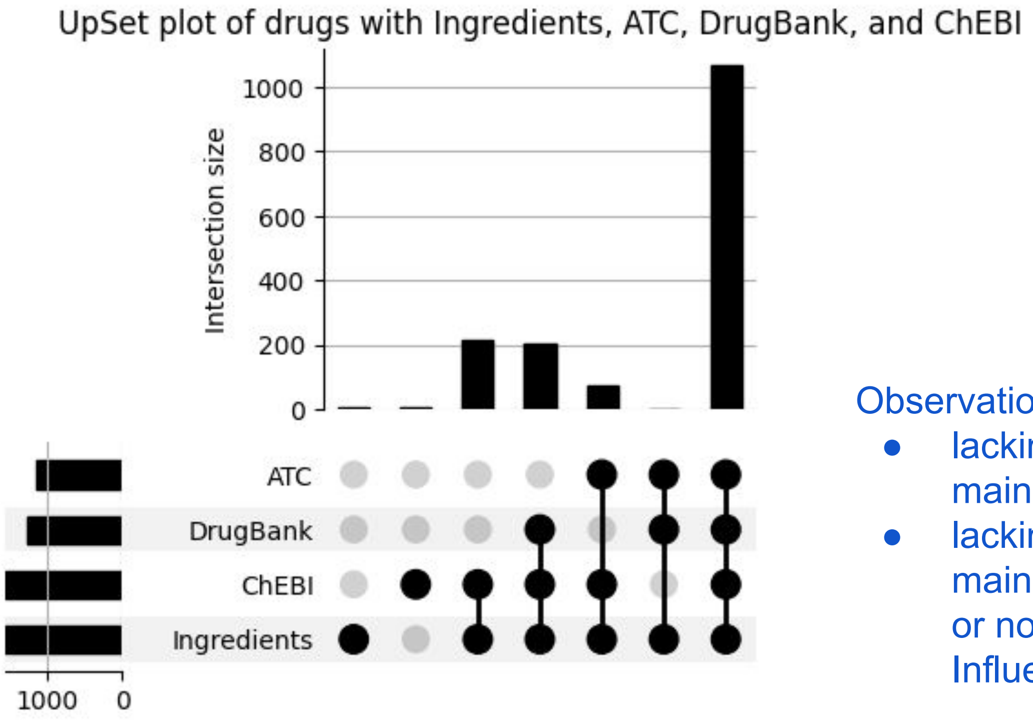
ATC 3rd& 4th levels

Chemical, Pharmacological or Therapeutic subgroup

e.g.

- N **NERVOUS SYSTEM**
- N06 **PSYCHOANALEPTICS**
- N06A **ANTIDEPRESSANTS**
- N06AA **Non-selective monoamine reuptake inhibitors**
- N06AA01 **desipramine**

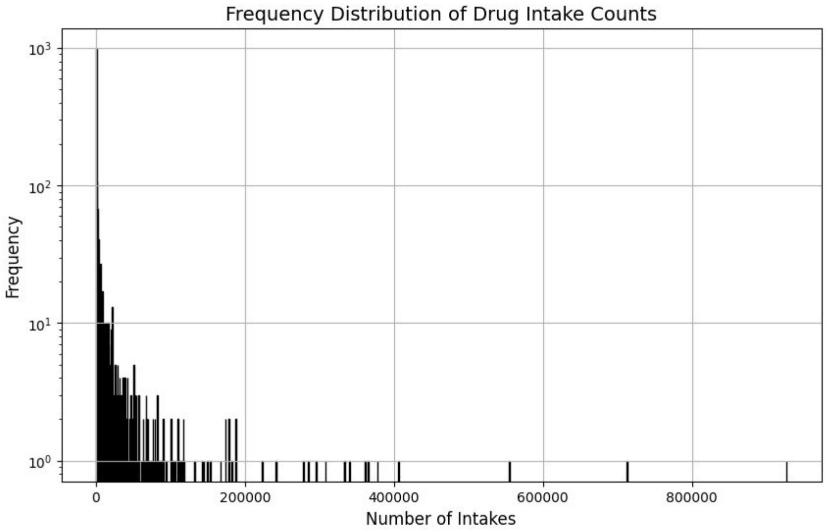
Drug Information

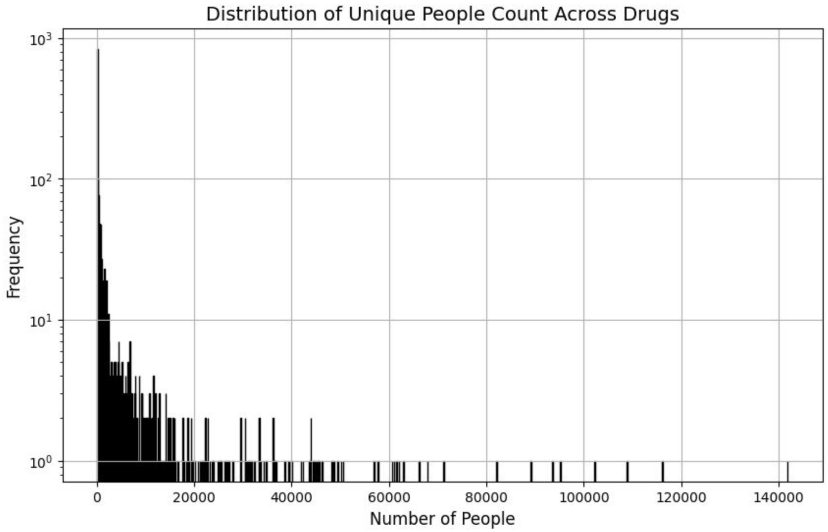


Observation:

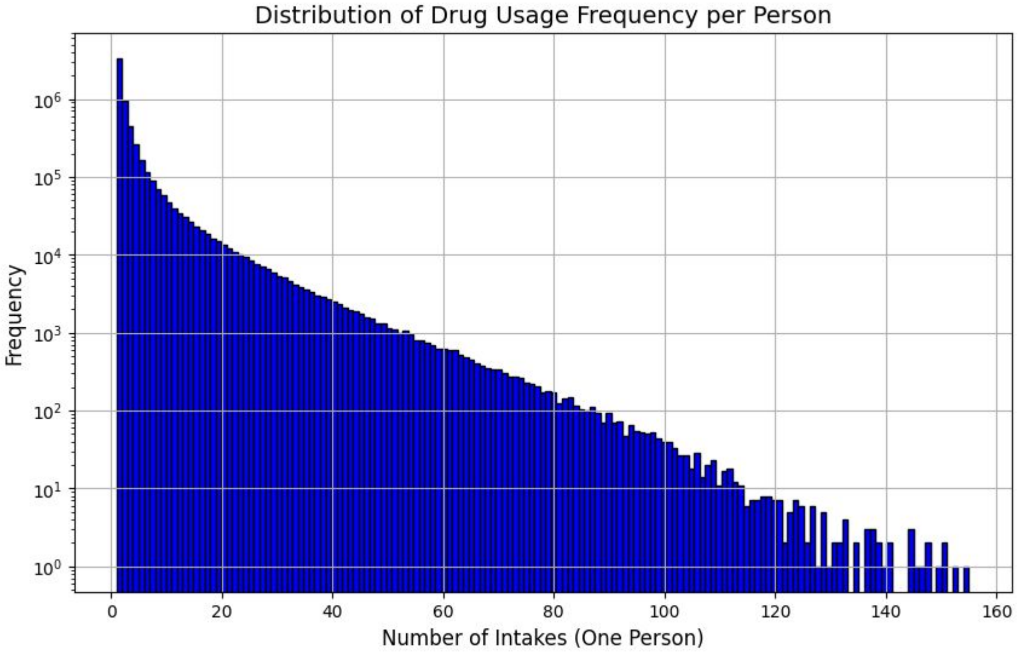
- lacking CHEBI/Ingredient: very rare cases, mainly due to dataset error
- lacking ATC code/DrugBank: sometimes, mainly due to non-therapeutic ingredients or non-therapeutic ingredients e.g. Influenza A virus

Individual-Level Drug Era


Frequency Distribution of Drug Intake Counts

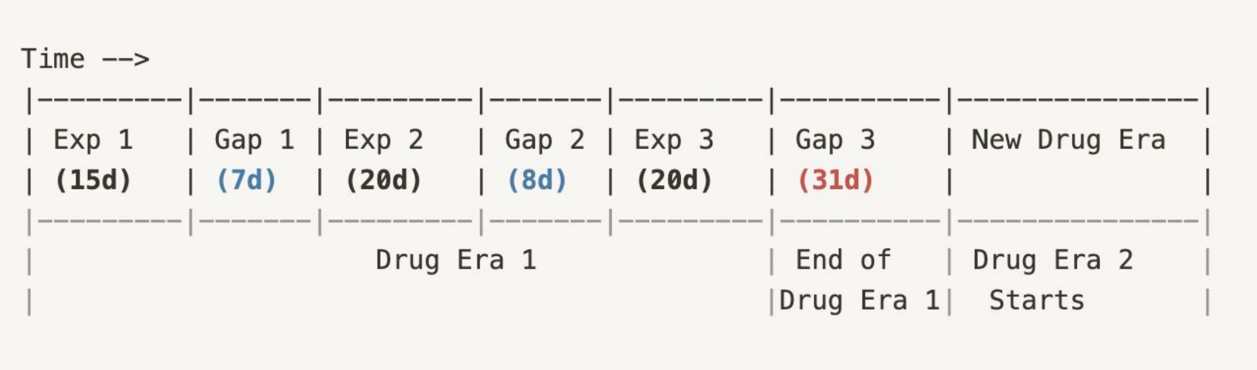

Distribution of Unique People Count Across Drugs

Individual Drug Era



Drug Era Data Structure

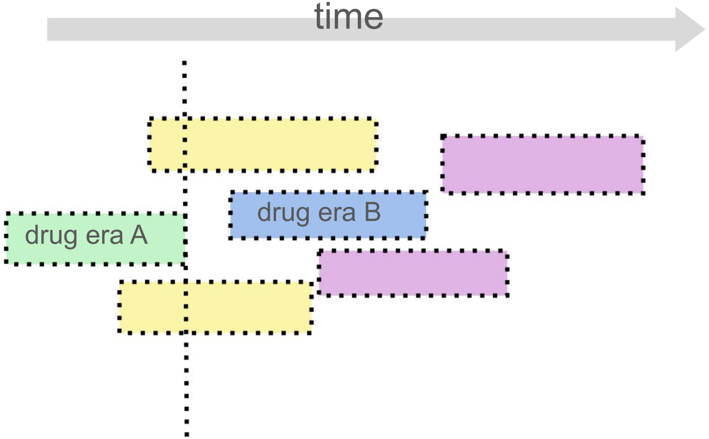
- **drug exposure**: a strictly continous period of using one drug (same active substance)
- **drug era**: a period of drug use for a person (allowing short gaps ≤ 30 days)
- **component of a drug era**: which person, what drug, from when to when, how many times (exposures), total gap days within in the drug era



```
Time -->
|---------|-------|---------|-------|---------|----------|---------------|
| Exp 1   | Gap 1 | Exp 2   | Gap 2 | Exp 3   | Gap 3    | New Drug Era  |
| (15d)   | (7d)  | (20d)   | (8d)  | (20d)   | (31d)    |               |
|---------|-------|---------|-------|---------|----------|---------------|
|                     Drug Era 1                | End of   | Drug Era 2    |
|                                               |Drug Era 1|  Starts       |
```

Drug Switch Pattern

- defining **drug switch (drug era A → drug era B)**:
  - after stopping drug era A, what drug era did the person first begin?
  - for each drug era A, we find its **closest subsequent drug era**: the drug era that first started after A ends



- switch pattern: drug A → drug B

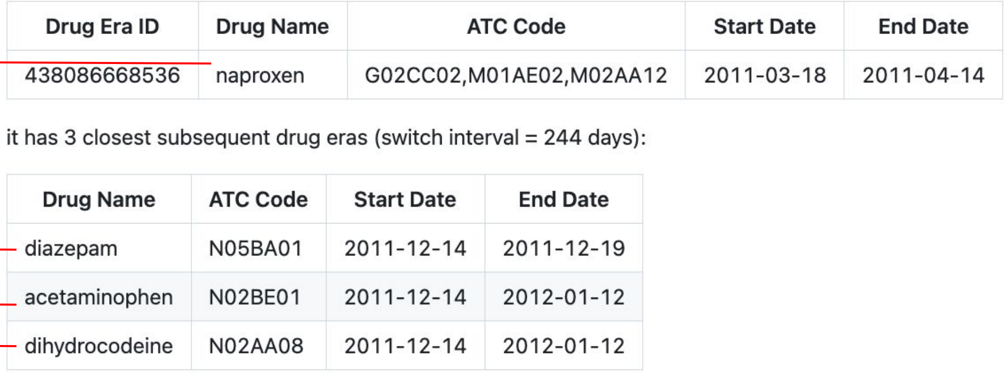
Drug Switch Pattern

- number of **drug eras**: approx. 20 million
  (44% of them have multiple switches)
- total number of **drug switches** founded approx. 37 million

Drug Switch Pattern

- number of **drug eras**: approx. 20 million (44% of them have multiple switches)
- total number of **drug switches** approx. 37 million

an example on A → multiple B



Drug Era ID	Drug Name	ATC Code	Start Date	End Date
438086668536	naproxen	G02CC02,M01AE02,M02AA12	2011-03-18	2011-04-14

it has 3 closest subsequent drug eras (switch interval = 244 days):

Drug Name	ATC Code	Start Date	End Date
diazepam	N05BA01	2011-12-14	2011-12-19
acetaminophen	N02BE01	2011-12-14	2012-01-12
dihydrocodeine	N02AA08	2011-12-14	2012-01-12

treat pain & inflammatory diseases

treat anxiety disorders

treat mild to moderate pain

treat moderate to severe pain

PMI on switch source & destination

- how significant/meaningful is a drug switch pattern drug A → drug B ?

- unique drug switch pattern (drug A, drug B): approx. 400k
  - we can compare to find what switch happened most frequently;
  - biased

- use PMI score to determine whether a switch happens by chance

PMI on switch source & destination

- use PMI score to determine whether a switch happens by chance
- PMI: Pointwise Mutual Information

$$\mathrm{PMI}(x, y) = \log \frac{p(x, y)}{p(x)p(y)}$$

p(x,y) = p(x)p(y), PMI = 0, indenpendent, co-occur by chance
p(x,y) < p(x)p(y), PMI < 0, co-occur less than by chance
p(x,y) > p(x)p(y), PMI > 0, co-occur more than by chance

PMI on switch source & destination

- use PMI score to determine whether a switch happens by chance

Pr (A→B)

PMI for a drug switch A → B

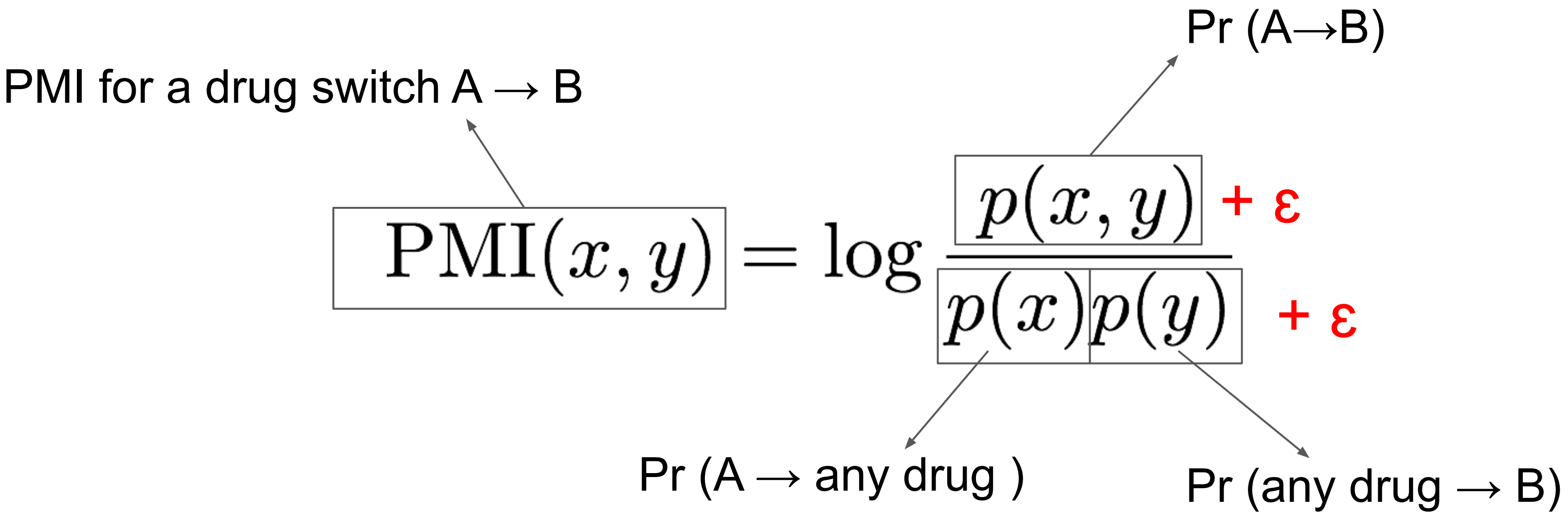
$$\boxed{\mathrm{PMI}(x,y)} = \log\frac{\boxed{p(x,y)}}{\boxed{p(x)}\boxed{p(y)}}$$

Pr (A → any drug)
A is the source of the switch

Pr (any drug → B)
B is the destination of the switch

PMI on switch source & destination

- use PMI score to determine whether a switch happens by chance



PMI for a drug switch A → B

Pr (A→B)

$$\boxed{\mathrm{PMI}(x, y)} = \log \frac{\boxed{p(x, y)} + \varepsilon}{\boxed{p(x)}\boxed{p(y)} + \varepsilon}$$

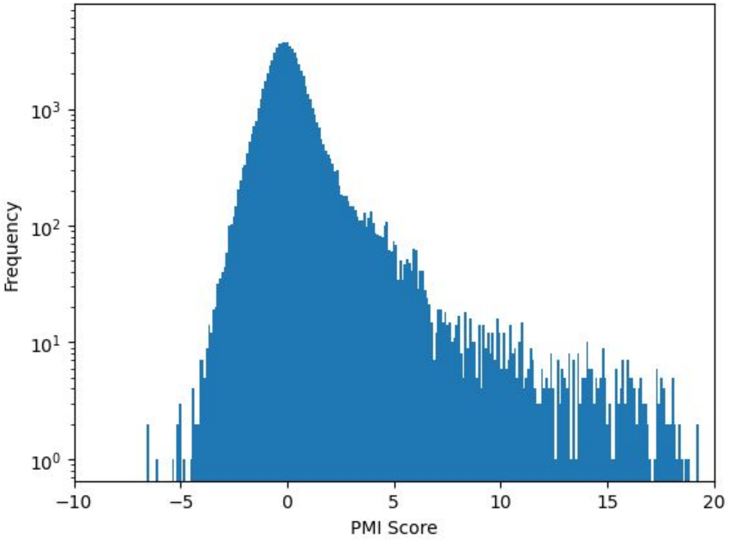
Pr (A → any drug)

Pr (any drug → B)

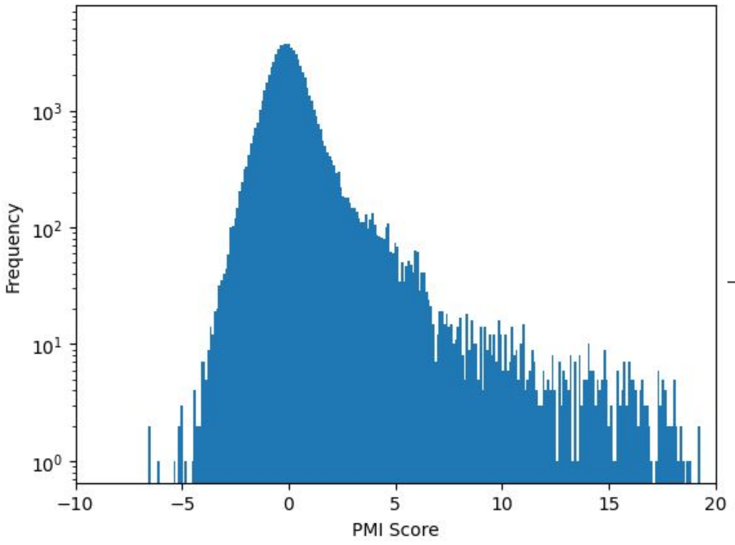
- too uncommon switches can have misleading high PMI values → filtered out

PMI on switch source & destination

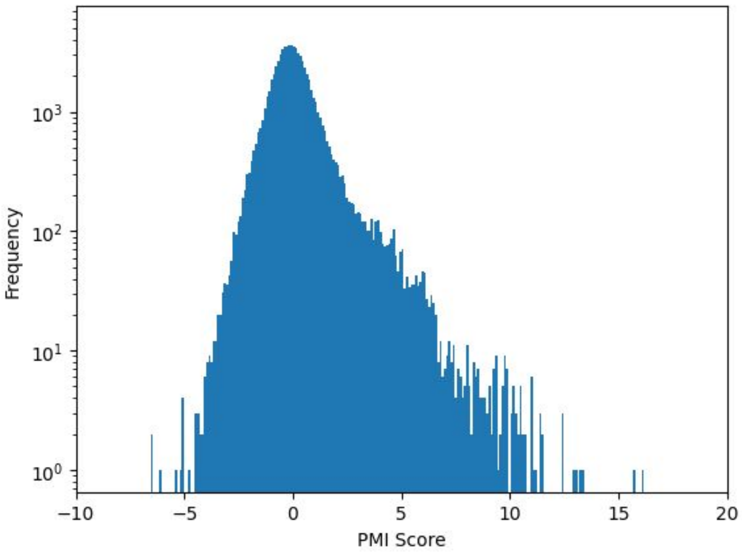
- 76,152 distinct drug switch patterns after filtering



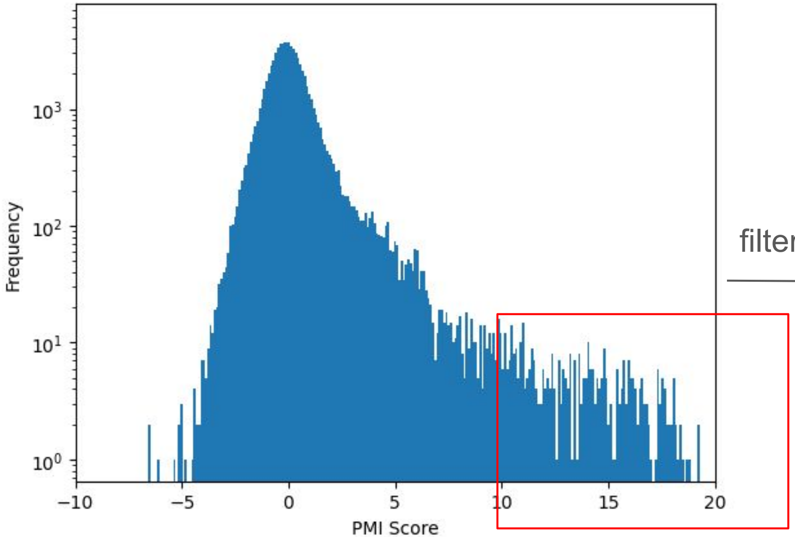
PMI on switch source & destination



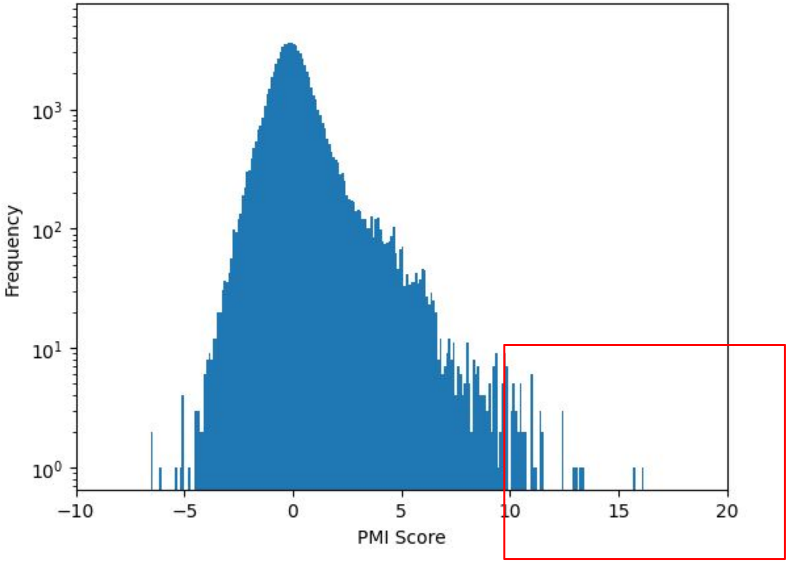
filter out self-switches



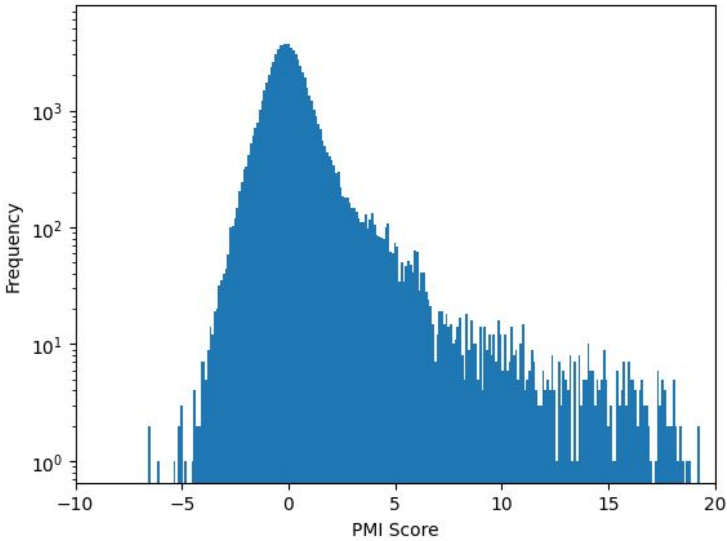
PMI on switch source & destination



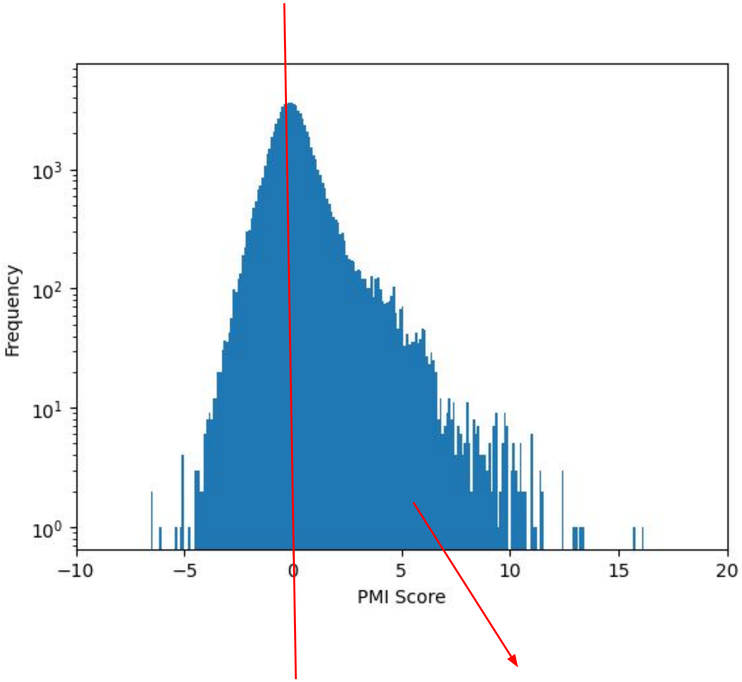
filter out self-switches



PMI on switch source & destination



filter out self-switches



independent

meaningful
drug switches

PMI on switch source & destination

an example:

oxycodone → naloxone
PMI score: 11.522 (really high)

oxycodone being source: 4910 times
naloxone being destination: 251 times
oxycodone → naloxone: 105 times

naloxone → oxycodone
PMI score: 11.255 (really high)

naloxone being source: 212 times
oxycodone being destination: 7061 times
naloxone → oxycodone: 106 times

Oxycodone: opioid, treating pain
Naloxone: treating opioid overdose and respiratory or mental depression due to opioids

PMI on switch source & destination

an example:

oxycodone → naloxone
PMI score: 11.522 (really high)

oxycodone being source: 4910 times
naloxone being destination: 251 times
oxycodone → naloxone: 105 times

naloxone → oxycodone
PMI score: 11.255 (really high)

naloxone being source: 212 times
oxycodone being destination: 7061 times
naloxone → oxycodone: 106 times

↓

indicates strong addiction

PMI on switch pattern 1 & switch pattern 2

Is drug switch pattern A → B related to pattern C→D ?

- only drug switches between drugs under ATC code category “N” (Nervous System)
- 1,875 kinds of switches after filtering
- Combination (1875, 2) = 1,756,875 switch pairs

PMI on switch 1 & switch 2

- use PMI score to determine whether **a person have two switches** by chance

Pr (a person has had both A → B & C → D)

PMI for a A → B & C → D

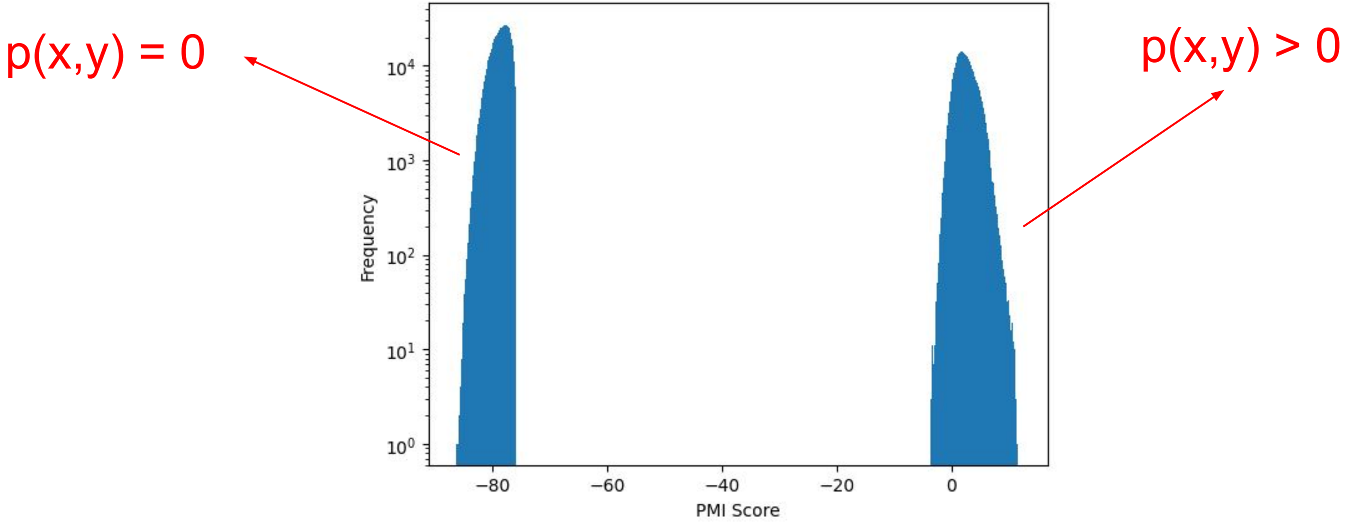
$$\boxed{\mathrm{PMI}(x,y)} = \log \frac{\boxed{p(x,y)} + \varepsilon}{\boxed{p(x)}\boxed{p(y)} + \varepsilon}$$

Pr (a person has had A → B)

Pr (a person has had C → D)

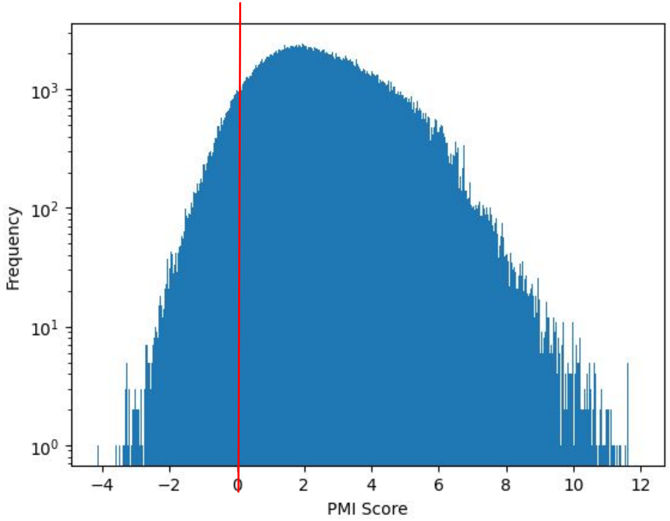
PMI on switch 1 & switch 2

Is drug switch A → B related to C→D ?
Combination (1875, 2) = 1,756,875 switch pairs



very small epsilon is used;
p(x,y) = 0 vs. 1e-6 can make huge difference in PMI score

PMI on switch 1 & switch 2



- drug switch A → B and C→D tend to co-occur
- possible underlying factors

PMI on switch 1 & switch 2

example:

Switch 1: ergotamine → amitriptyline
Switch 2: amitriptyline → caffeine
PMI score: 11.316

- **Ergotamine**: treat migraines
- **Amitriptyline**: treat chronic pain; preventive treatment for migraines
- **Caffeine**: sometimes used to enhance the effects of pain relievers or manage certain headaches

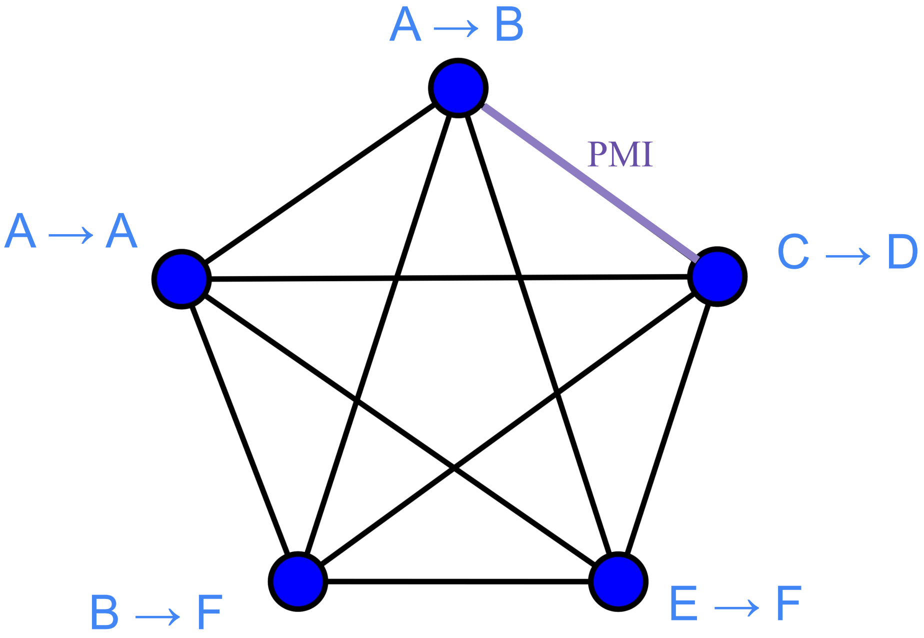
⬇

might indicates a fine-tuning of migraine treatment or managing side effects of amitriptyline (e.g., fatigue or sedation).

PMI on switch 1 & switch 2

Graph

Nodes: 1875 switches;
Undirected Edges;
Weight: PMI scores



PMI on switch 1 & switch 2

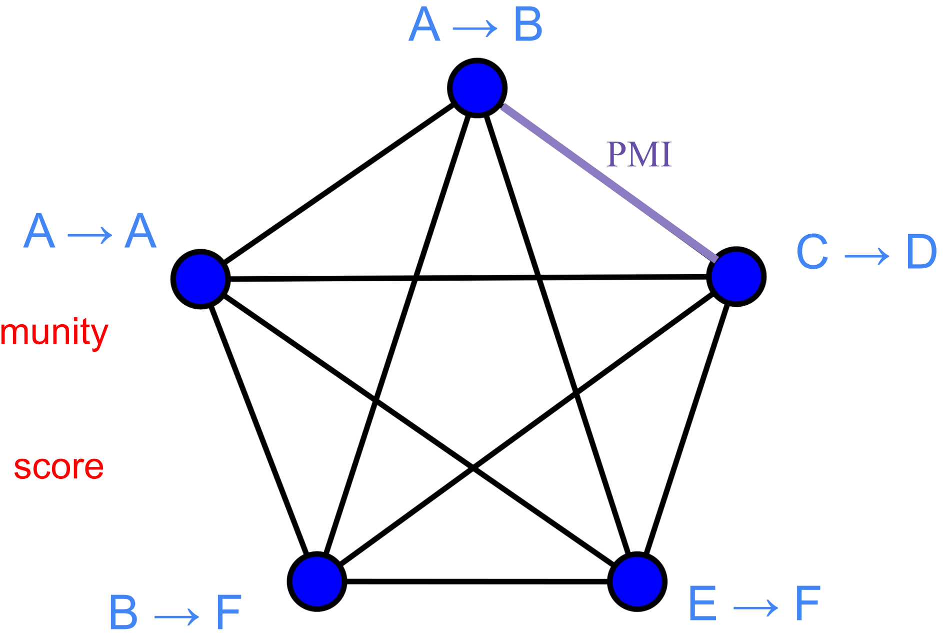
Graph

Nodes: 1875 switches;
Undirected Edges;
Weight: PMI scores

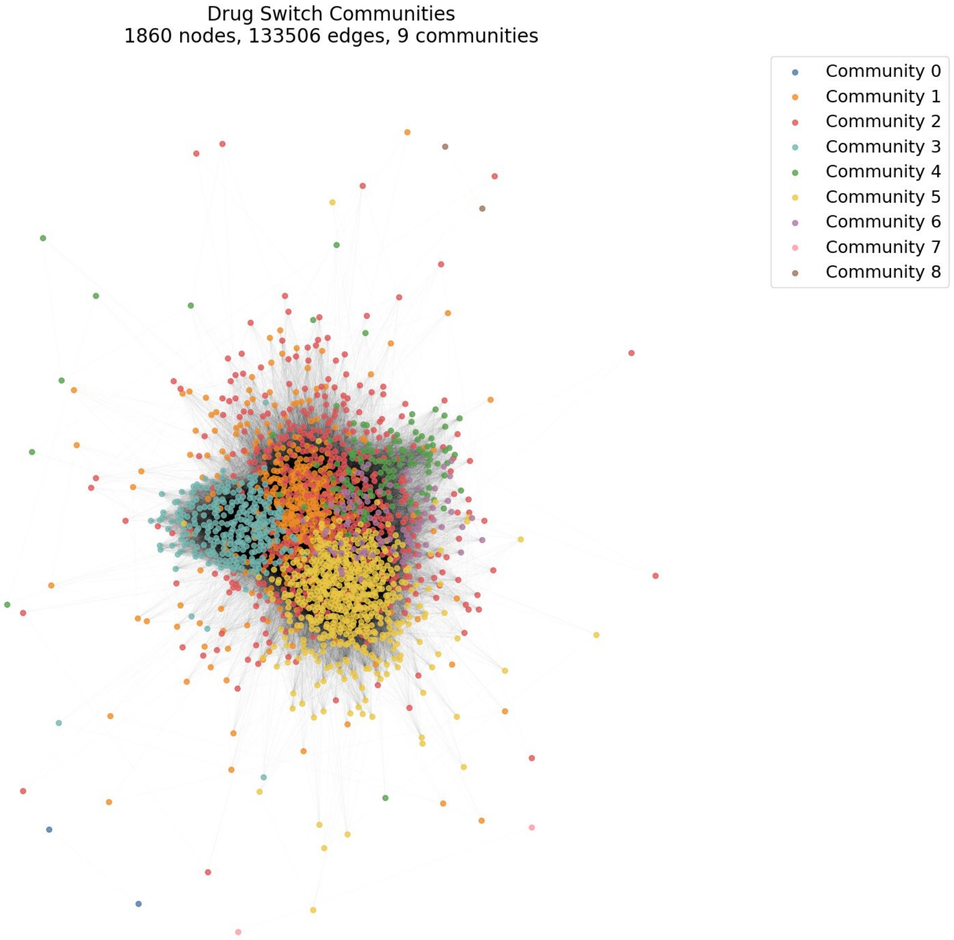
clustering: Louvain community detection algorithm

maximizing the avg PMI score within each cluster

9 clusters of drug switches



PMI on switch 1 & switch 2



- Nodes closed together: strongly connected nodes (high PMI)
- Nodes farther apart: weak connection

Patterns of Gap Days

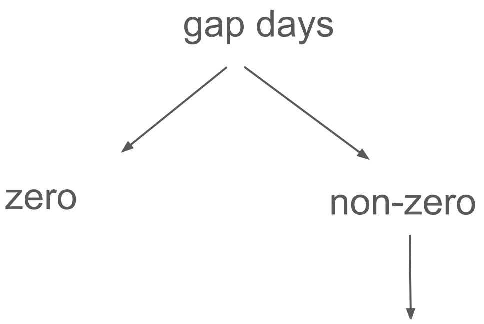
gap: the patient’s adherence to the treatment

- what drug causes people to have more gaps?
- can we predict the number of gap days given what drug it is and the drug era’s duration?

f(drug, duration) = gap days

Patterns of Gap Days

clean dataset (16 drugs, excluding outliers); zero-inflated model; two-stage approach



gap days

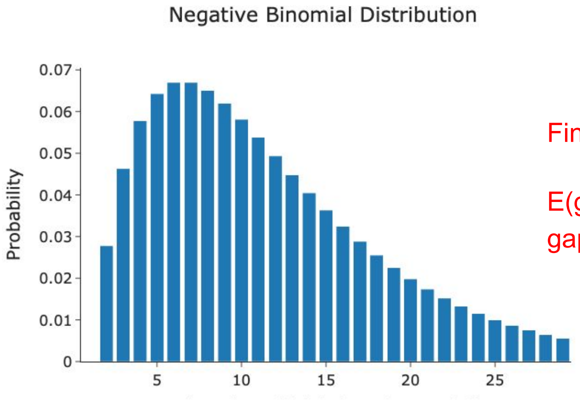
zero

non-zero

$\text{logit}(P(\text{gap days} > 0)) = \gamma 0 + \gamma 1X1 + \gamma 2X2 + \cdot\ \cdot\ \cdot$

$\log(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \log(\text{duration})$

μ : E(gap days)



Final Probability Model:

E(gap days) = P(gap days > 0) × E(gap days | gap days > 0)

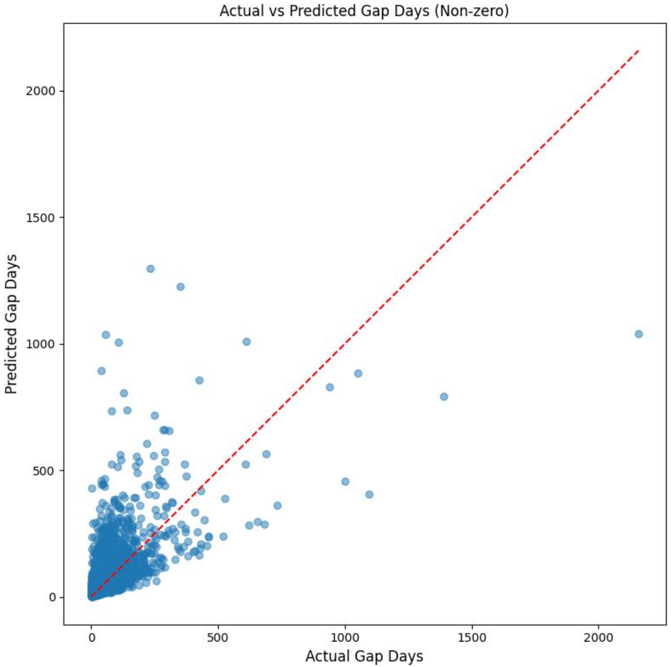
Patterns of Gap Days

one hot encoding for different drugs : [0, 1, 0, 0, …]

$$\log(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \log(\text{duration})$$

$$\mu : E(\text{gap days})$$

- small LLR (Log-Likelihood Ratio) p-value
- pseudo R-square (11.09%): limited predictive power



Patterns of Gap Days

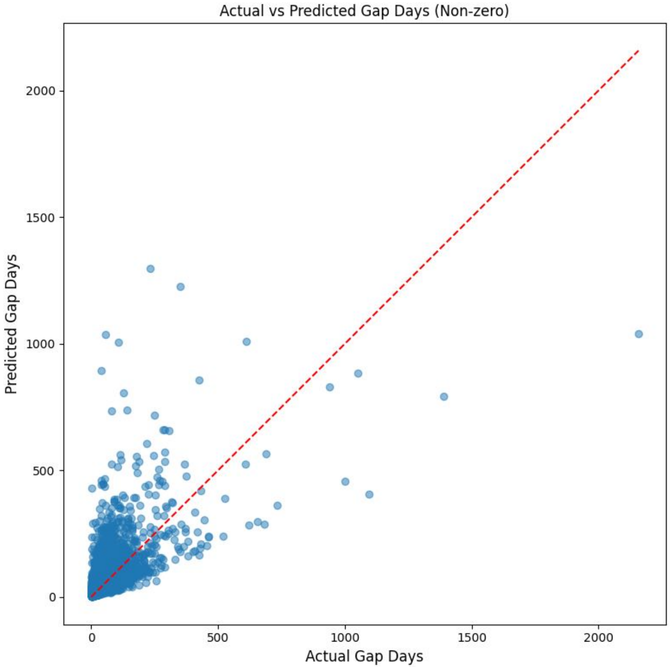
one hot encoding for different drugs : [0, 1, 0, 0, …]

$$\log(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \log(\text{duration})$$

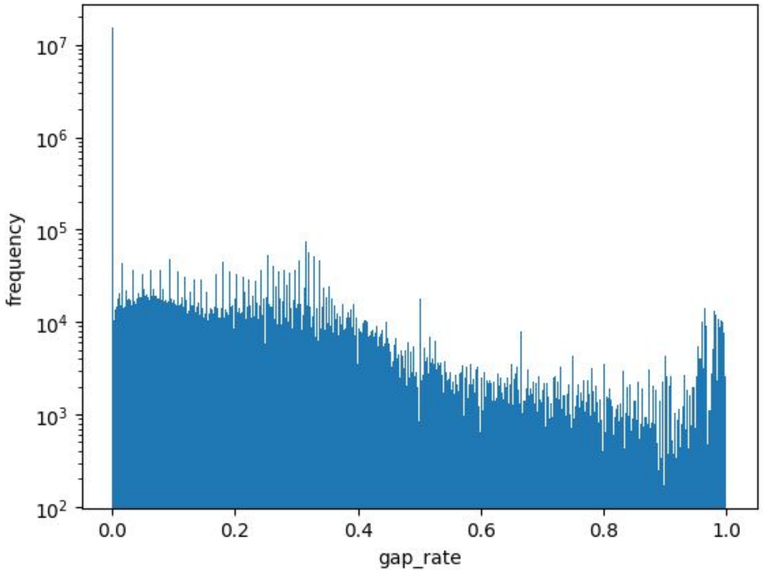
$$\mu : E(\text{gap days})$$

- small LLR (Log-Likelihood Ratio) p-value
- pseudo R-square (11.09%): limited predictive power

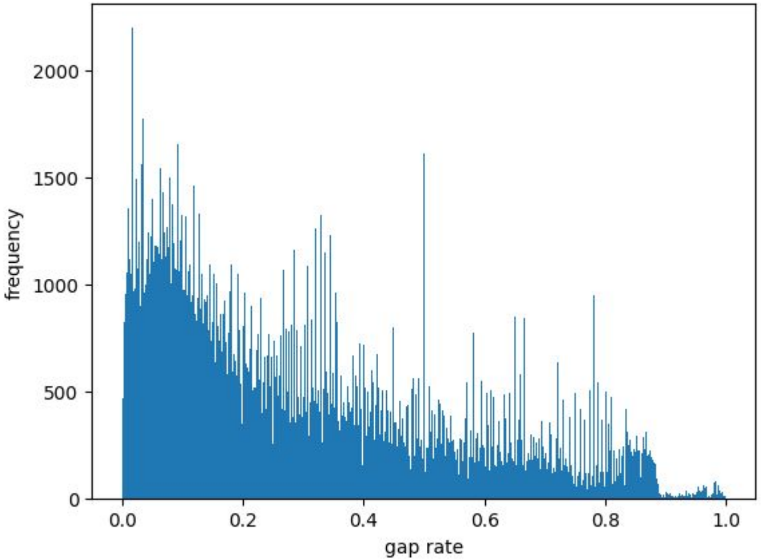
- gap is affected by drug (and duration); but not fully - population-level factors
- different drugs may affect gap days in different ways



drug’s impact on gap
gap rate = gap days / duration



gap rate for **all** drug eras

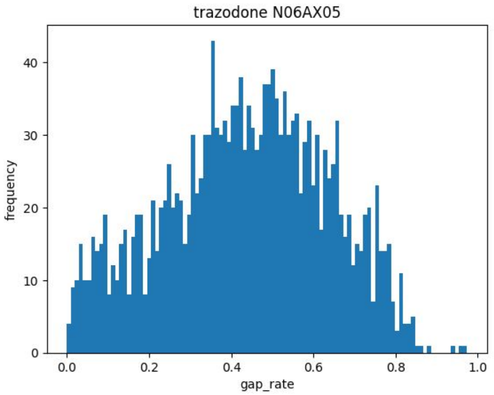


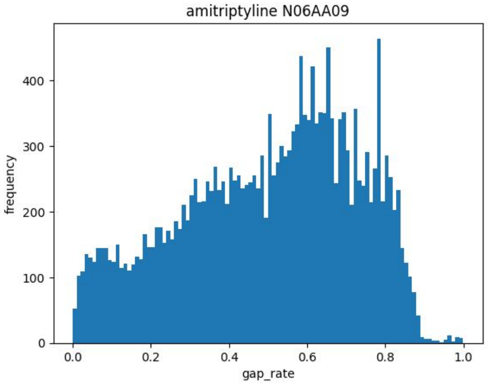
gap rate for drug eras with
N06 (psychoanaleptics)

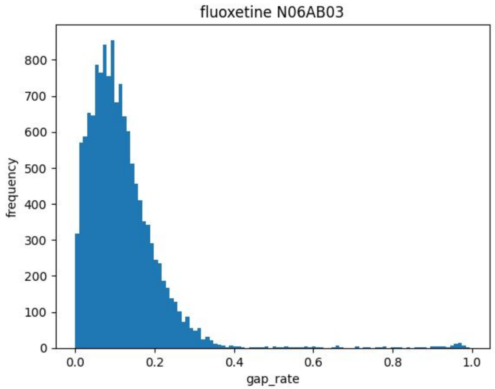
drug’s impact on gap

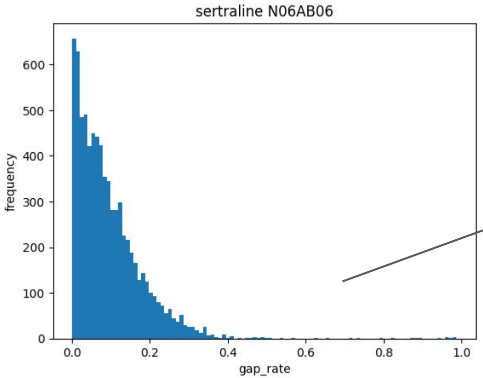
gap rate = gap days / duration

4 kinds of antidepressant (N06):









less gap,
more adherence;
probably less side effect

Thank you for listening