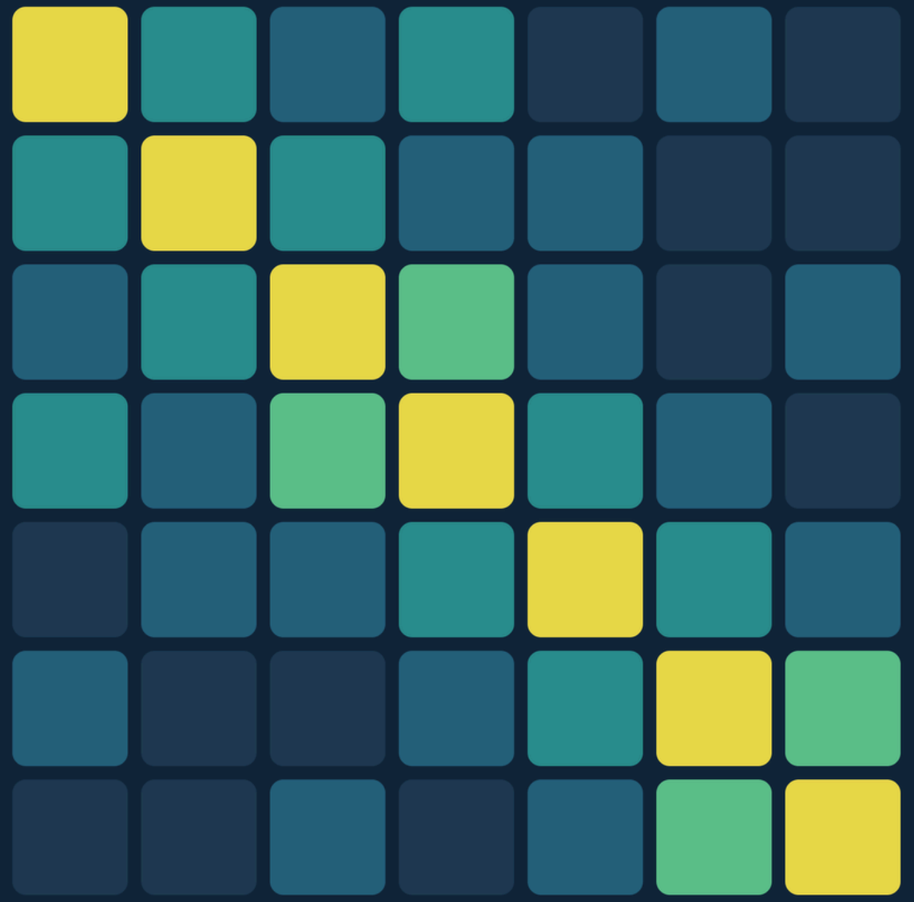
MIMIC-IV Demo · Comorbidity analysis

Which diseases travel together?

A market basket-style co-occurrence analysis of comorbidities using cosine similarity



Safiya Arif & Raviha Okasha

Problem statement

Diseases rarely travel alone

1 · At the clinic



Just my blood pressure.

Hypertension

2 · Soon to come along



High cholesterol

Diabetes

Other conditions often come along.

3 · In the records



1,310 conditions, 100 patients.
Which pairs really go together?

Illustration only. Our data shows which diseases go together, not which comes first.

About the Data Set

MIMIC-IV Clinical Database Demo

100
patients

1,310
distinct conditions

2
coding systems:
ICD-9 and ICD-10

≥ 5
patients needed to
keep a condition

Public and de-identified, released through PhysioNet.

Literature overview

Closest prior work

	Wang et al., 2024	Our study
Data	MIMIC-IV ICU, 55,486 patients; eICU and COHD to validate	MIMIC-IV demo, 100 patients
Focus	Comorbidities of one disease: acute kidney injury	All frequent conditions, no single target
Method	Apriori association rules on disease pairs	Binary matrix with cosine similarity

Wang et al., BMC Med Inform Decis Mak 24:35 (2024)

Exploring the Dataset:

The data was cleaned , patients were placed on rows, the unique diagonosis was put on the columns, if a patient was diagonosed 1 was placed else 0 was placed in the matrix.

subject_id	Abnormal coagulation profile	Accidents occurring in residential institution	Acidosis	Acute kidney failure	Acute respiratory failure	Acute respiratory failure with hypoxia	Alcohol use disorder	Alkalosis	Anemia	Anxiety
10000032	0	0	0	0	0	0	1	0	0	0
10001217	0	0	0	0	0	0	0	0	0	0
10001725	0	0	0	0	0	0	0	0	0	1
10002428	0	0	1	1	1	0	0	0	1	0
10002495	0	0	0	1	0	0	0	0	0	0

Methodology

The math behind:

The co-occurrence matrix counts the patients who have both conditions in a pair:

$$C = X^\top X, \qquad C_{ij} = \sum_{p=1}^{n} X_{pi}\, X_{pj}. \tag{1}$$

Raw counts are dominated by common conditions, so we normalized them with **cosine similarity**:

$$S_{ij} = \frac{C_{ij}}{\sqrt{C_{ii}}\,\sqrt{C_{jj}}}. \tag{2}$$

Here C_{ii} is the number of patients with condition i. S_{ij} ranges from 0 (the conditions never occur together) to 1 (they always occur together). Normalization **reduces**, but does not fully remove, the effect of prevalence.

The co-occurence matrix made using python:

	Abnormal coagulation profile	Accidents occurring in residential institution	Acidosis	Acute kidney failure	Acute respiratory failure	Acute respiratory failure with hypoxia	Alcohol use disorder	Alkalosis	Anemia	Anxiety
Abnormal coagulation profile	7	2	3	5	1	1	0	0	6	2
Accidents occurring in residential institution	2	13	4	6	3	3	3	1	10	2
Acidosis	3	4	19	14	6	4	9	2	14	3
Acute kidney failure	5	6	14	35	7	7	11	4	26	8
Acute respiratory failure	1	3	6	7	12	1	4	1	8	1

Results

Three groups of conditions

Group 1

Cardiometabolic

Hypertension · High cholesterol · Diabetes · Tobacco use · Anemia

≈ 0.70

Hypertension + high cholesterol

Group 2

Renal–cardiac

Acute kidney failure · CKD · Heart failure · Anemia

≈ 0.65

Kidney failure + CKD, and + anemia

Group 3

Acute illness

Sepsis · Acidosis · Thrombocytopenia · Pneumonia · Acute kidney failure

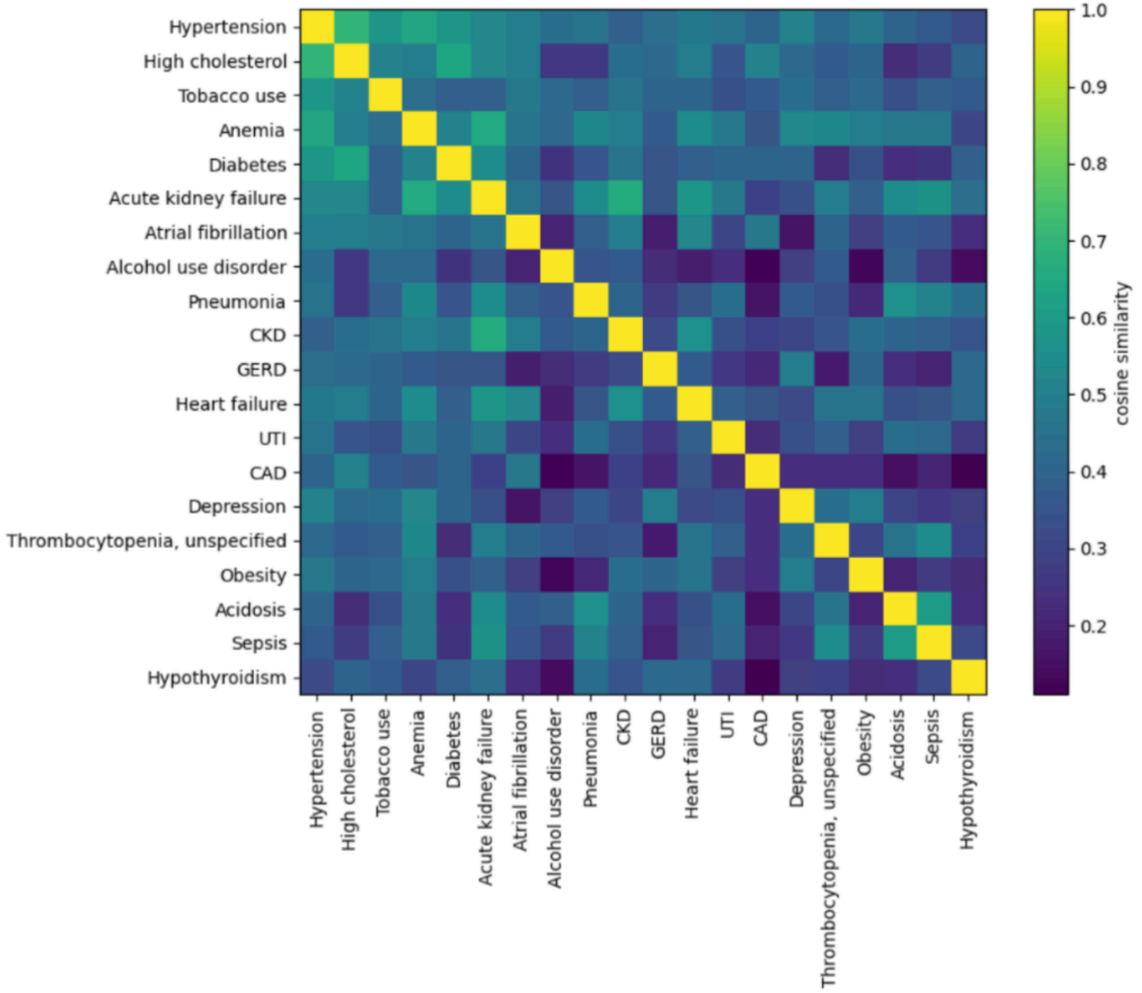
≈ 0.60

Sepsis + acidosis

Results

Cosine similarity map

- Often together
- Moderately together
- Rarely together



Results

Most common conditions



Share of the 100 patients. Most of the 1,310 conditions appear in only one patient.

Discussion

What it means

Matches clinical knowledge

- ✓ Shared risk factors: blood pressure, lipids, diabetes
- ✓ Kidney and heart disease interact
- ✓ Sepsis goes with acidosis

⚠ Handle with care

- ⚠ Only 100 patients, so exploratory
- ⚠ Association, not causation
- ⚠ Common conditions inflate scores
- ⚠ Names grouped by keyword rules

Plausible patterns from diagnosis codes alone.

Summary

Key findings

Three comorbidity groups

Cardiometabolic, renal–cardiac and acute illness, as clinicians expect.

Mapped market basket analysis onto patient records

Diagnoses become items in a patient × condition matrix.

Next steps

Validate on full MIMIC-IV and add lift and association rules.