



Project ATHENA

Augmenting Therapeutic Effectiveness through Novel Aalytics

Valerie Vandeweerd

Scientist Physician – Janssen Clinical Innovation

Charlotte Herzeel

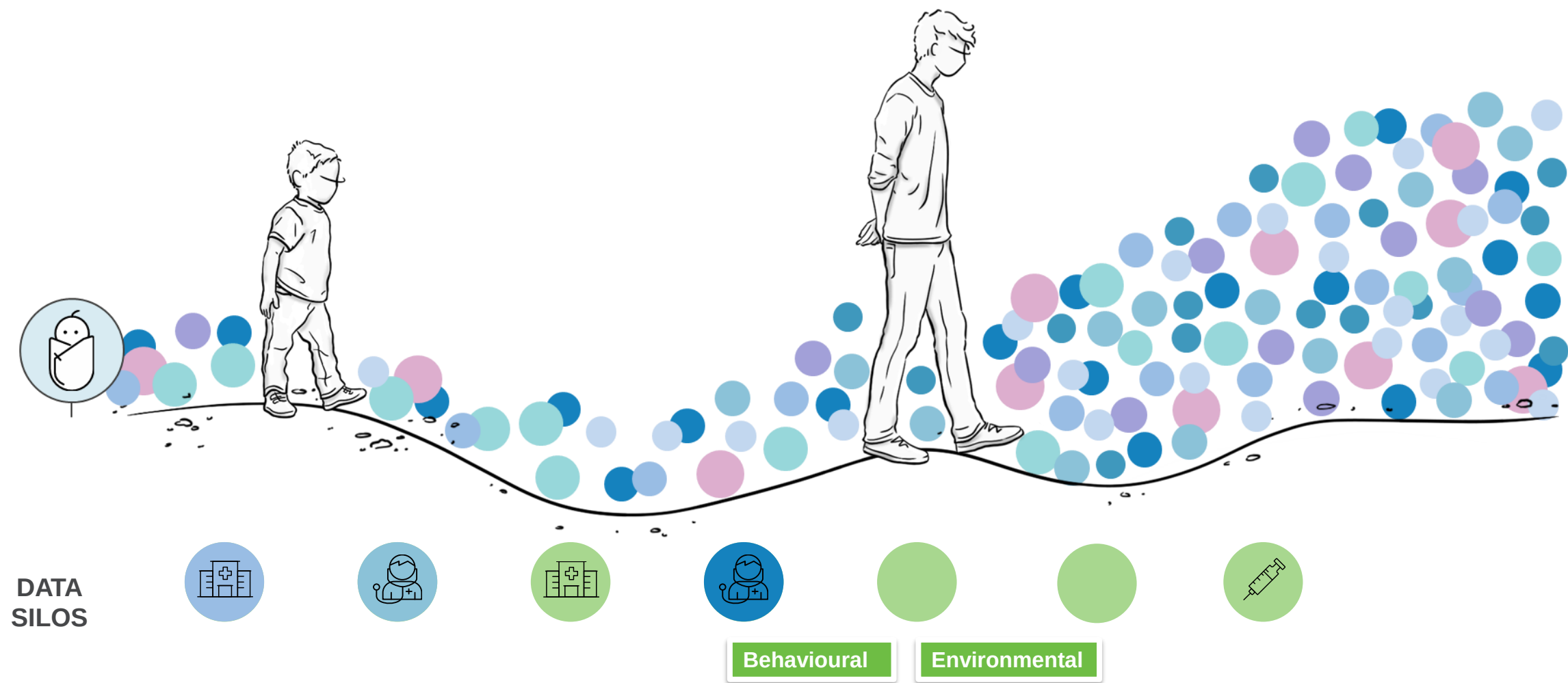
Senior Researcher - IMEC



Project number ATHENA: HBC.2019.2528

More and more data is being gathered

Ushering in the age of personalized healthcare

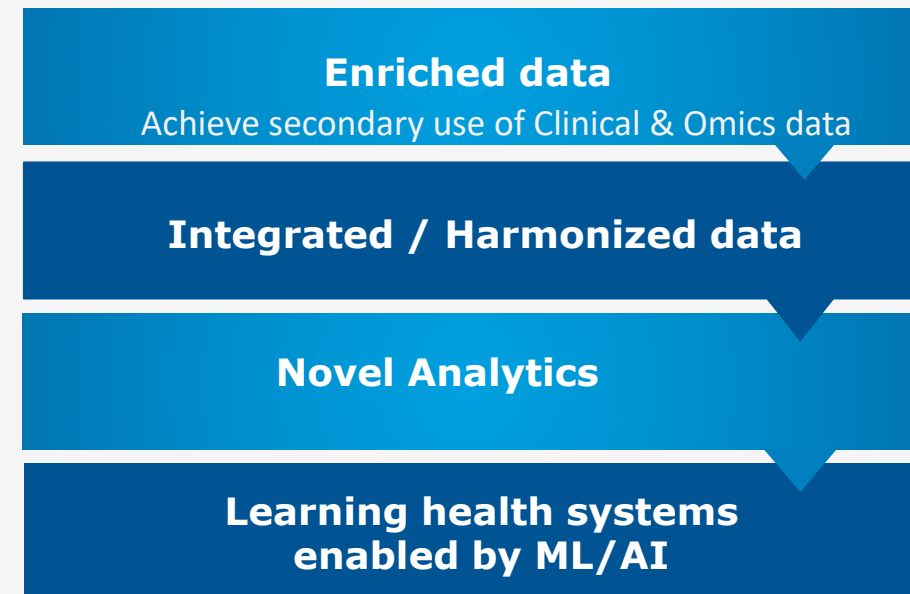


Harnessing real-world evidence for the development of novel cancer therapies*

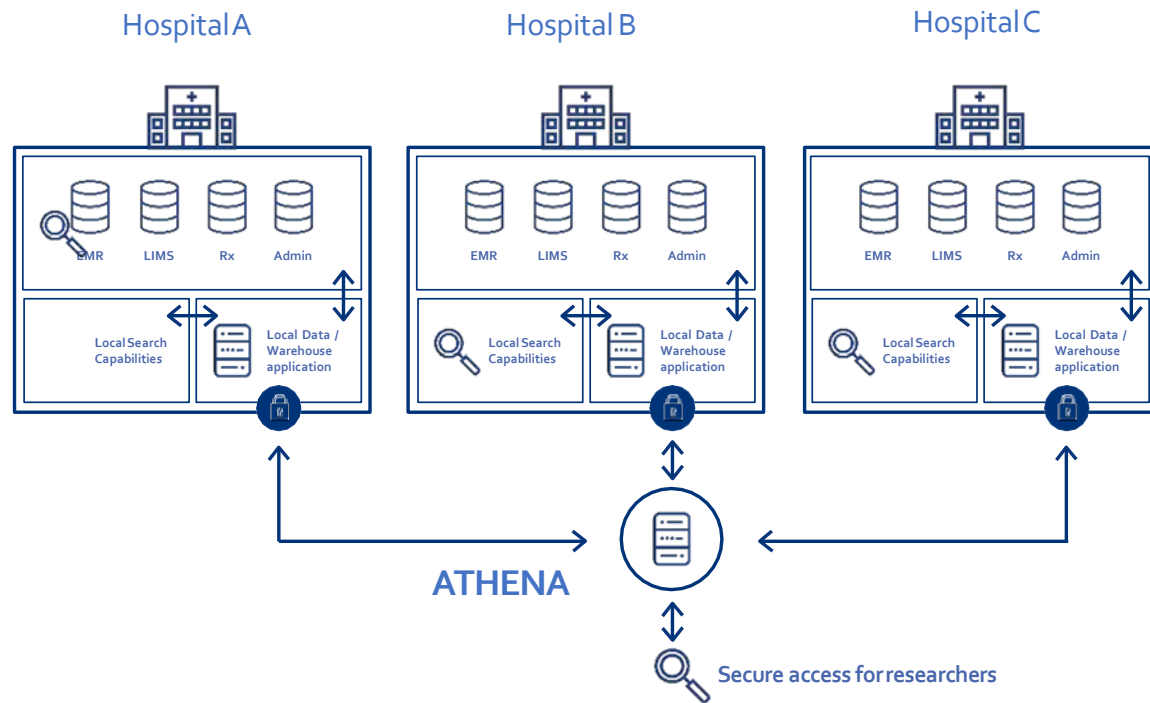


- **Novel analytic techniques** can bring together diverse data types to expand novel insights gained from RWE.
- **Enrichments of high-quality data coupled with machine learning** can help find areas of therapeutic need, enhance clinical trial interpretation, and even identify novel targets.

*Greshock et al., 2020



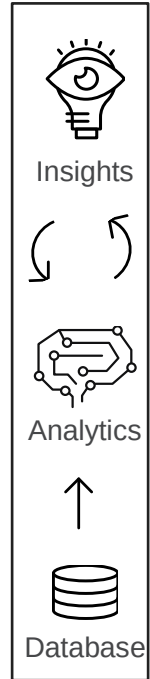
ATHENA's privacy-preserving federated ML platform



- Data remains under the control of the data owner
- No patient level data leaves the owner's site, only aggregated counts, thereby ensuring patient privacy
- **GDPR – 'Privacy by Design'**
- Analyses are brought to the data, rather than creating central data repository
- Use of common data model allows for efficient search/analyses across multiple data sets
- Local legal and ethical approvals
- Requires close collaboration with data owners which builds trust

Zooming into a single hospital

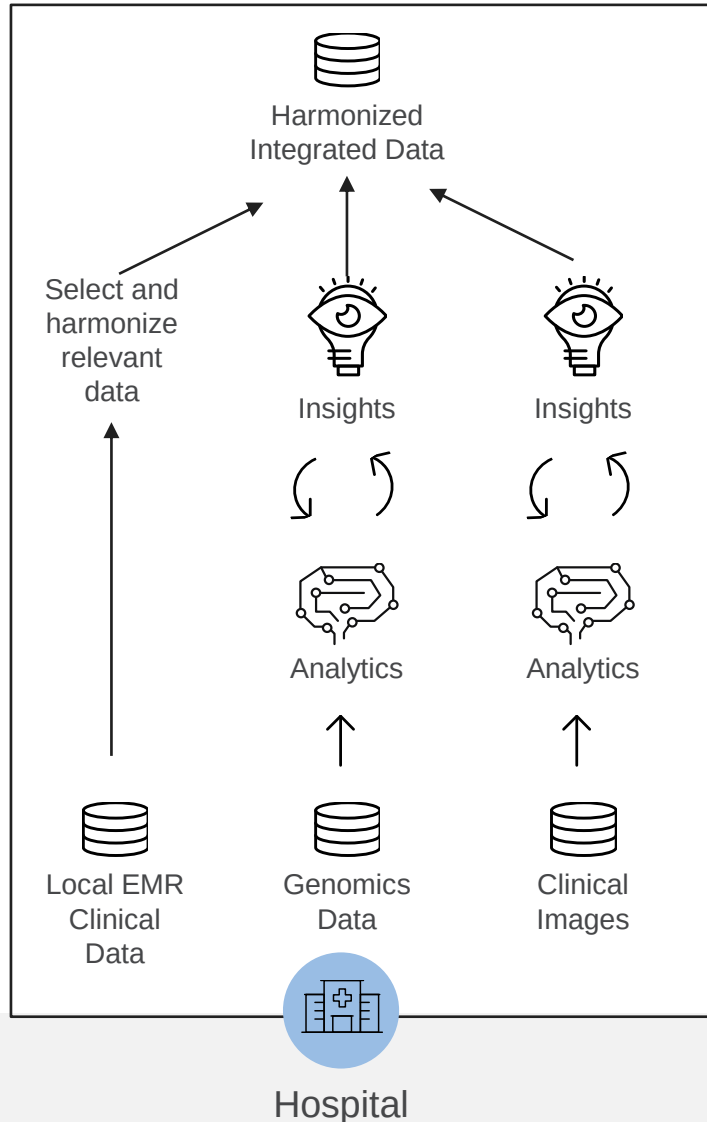
Even within a single hospital there are already multiple flows



**DATA
SILOS**

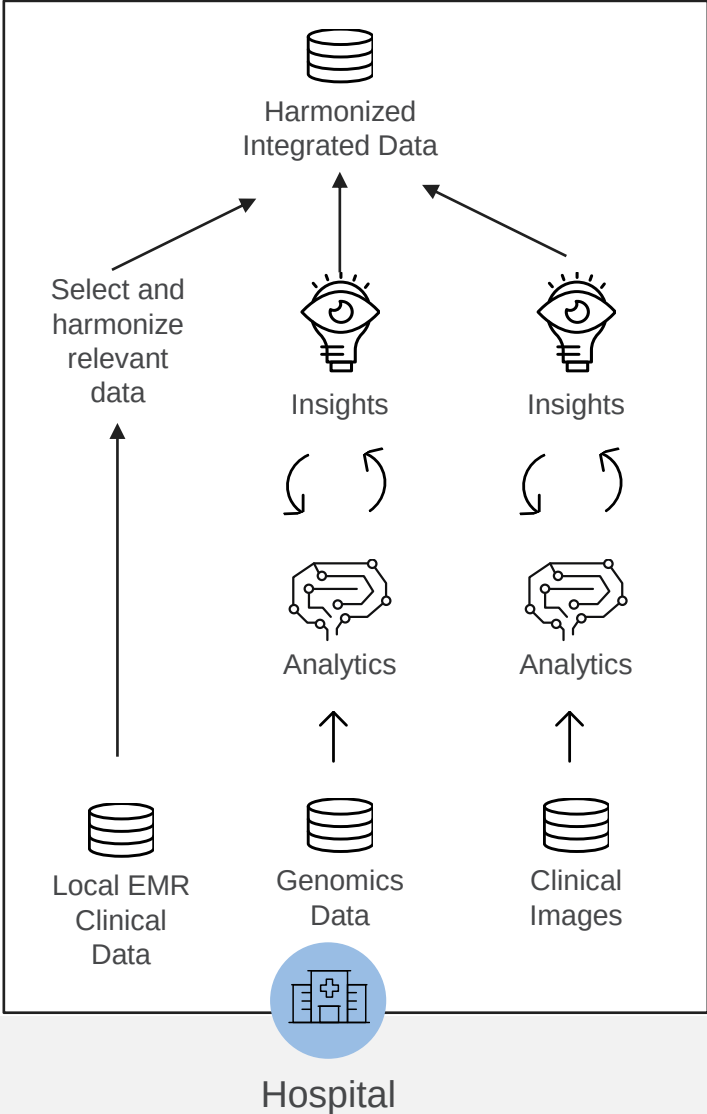
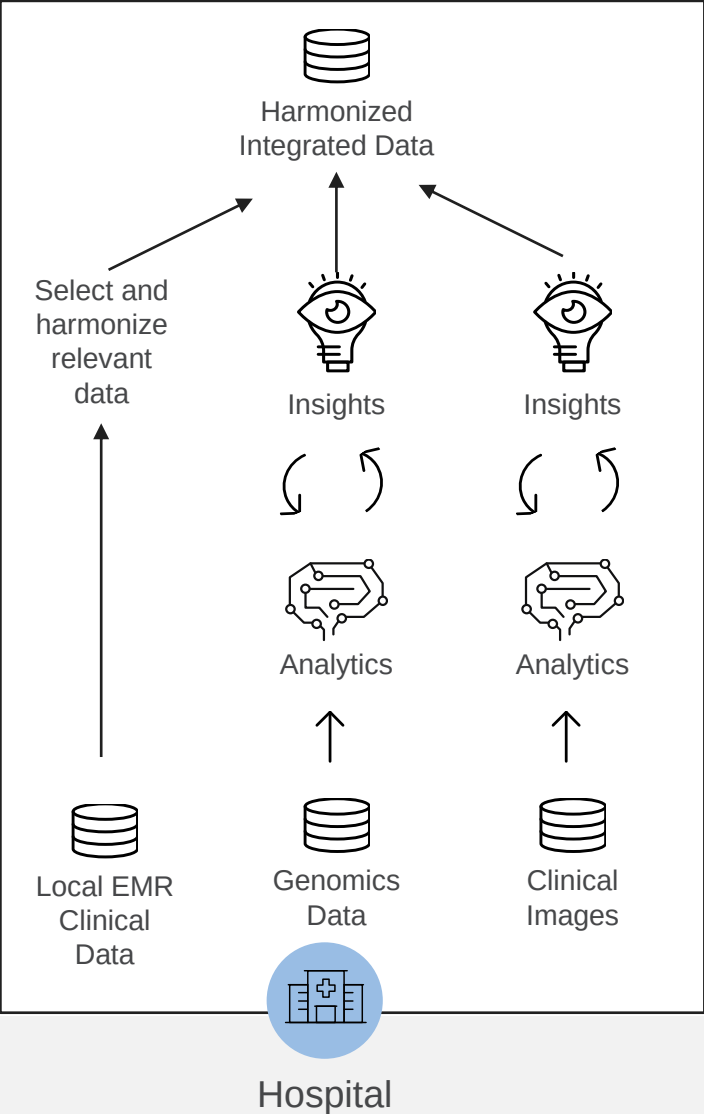
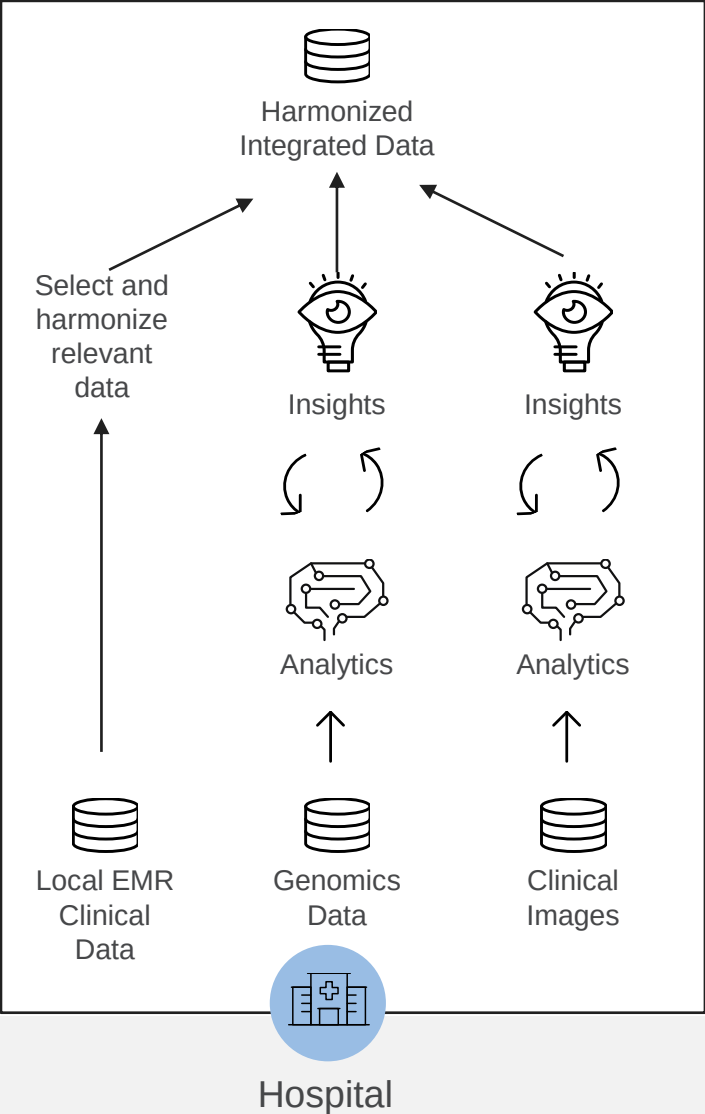


Zooming into a single hospital

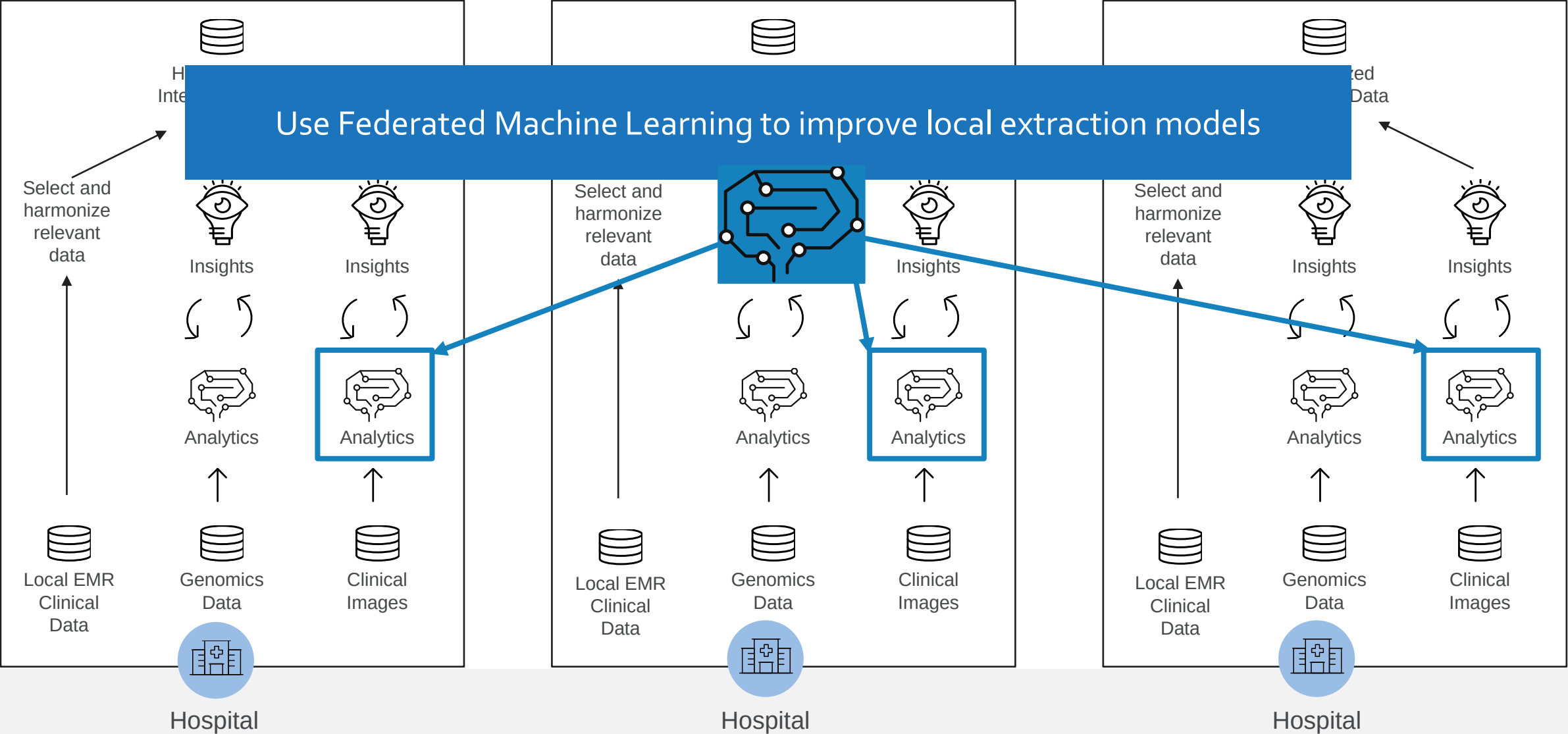


- Main idea:
 - Data from various local databases is extracted and stored in the harmonized integrated data store
 - EHR data is OMOPd
 - Diagnoses, treatments, lab results, ...
 - Clinical images are processed by an AI model, resulting in insights
 - Tumor size, stage, ...
 - Genomics data is processed by an AI model, resulting in insights
 - Disease affecting variants, ...

Can we let multiple hospitals collaborate ?



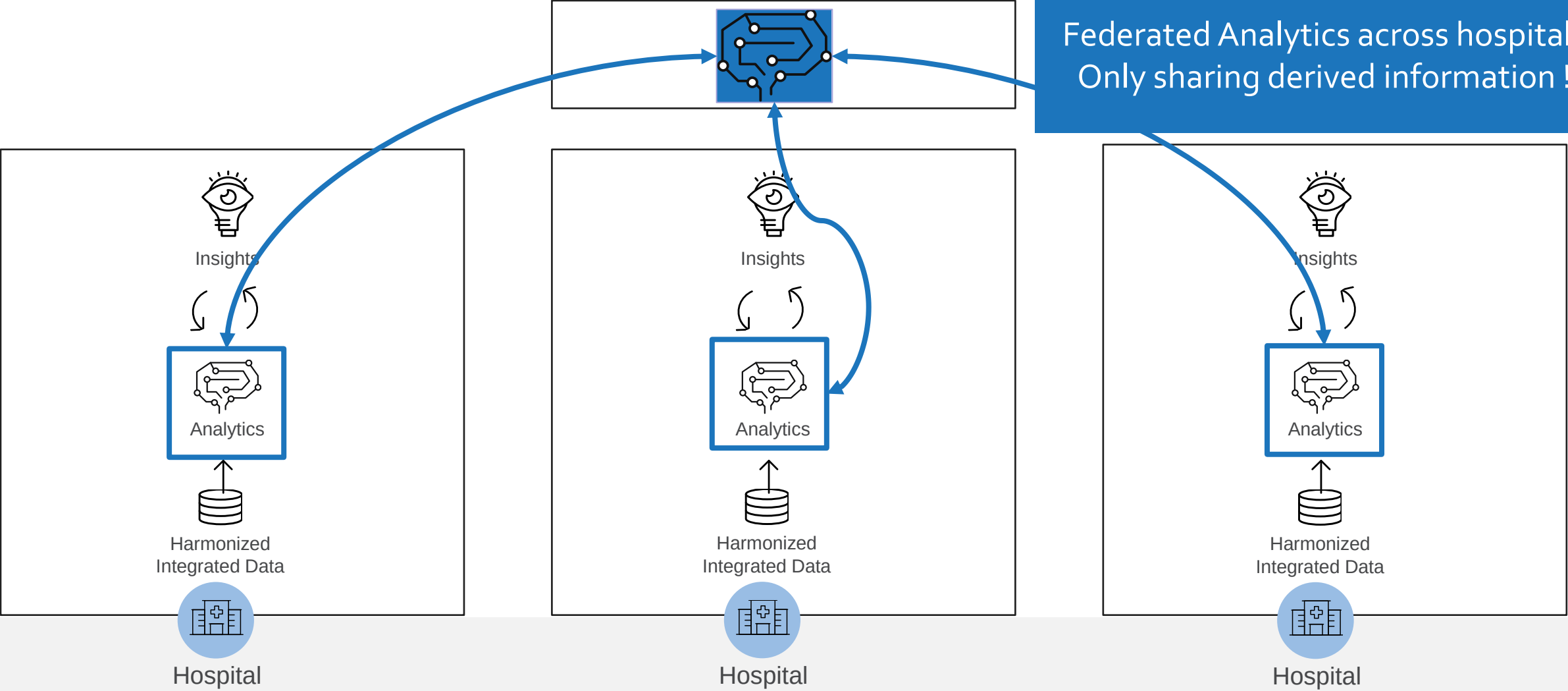
Opportunities for federation across hospitals



Can we let multiple hospitals collaborate ?

Secure Federated analytics server

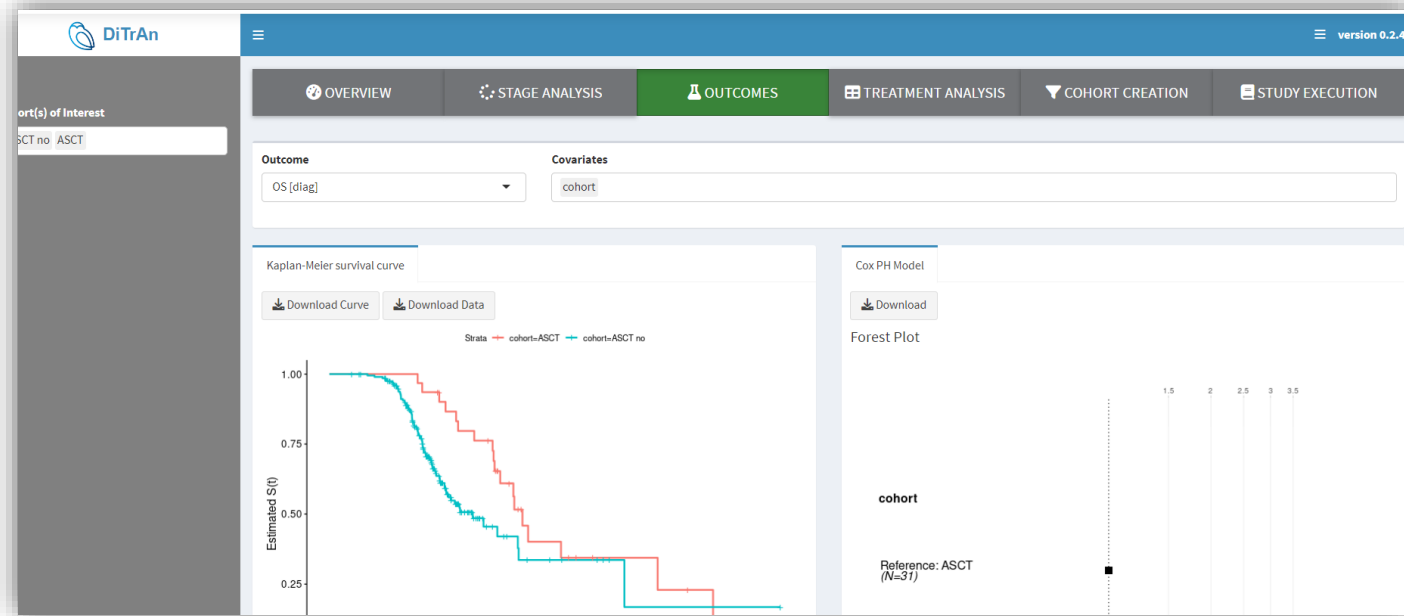
Federated Analytics across hospitals
Only sharing derived information !



Use case 1: longitudinal patient data exploration by physicians

Visualization of their
own hospital data

Enables rapid cohort
creation and comparison

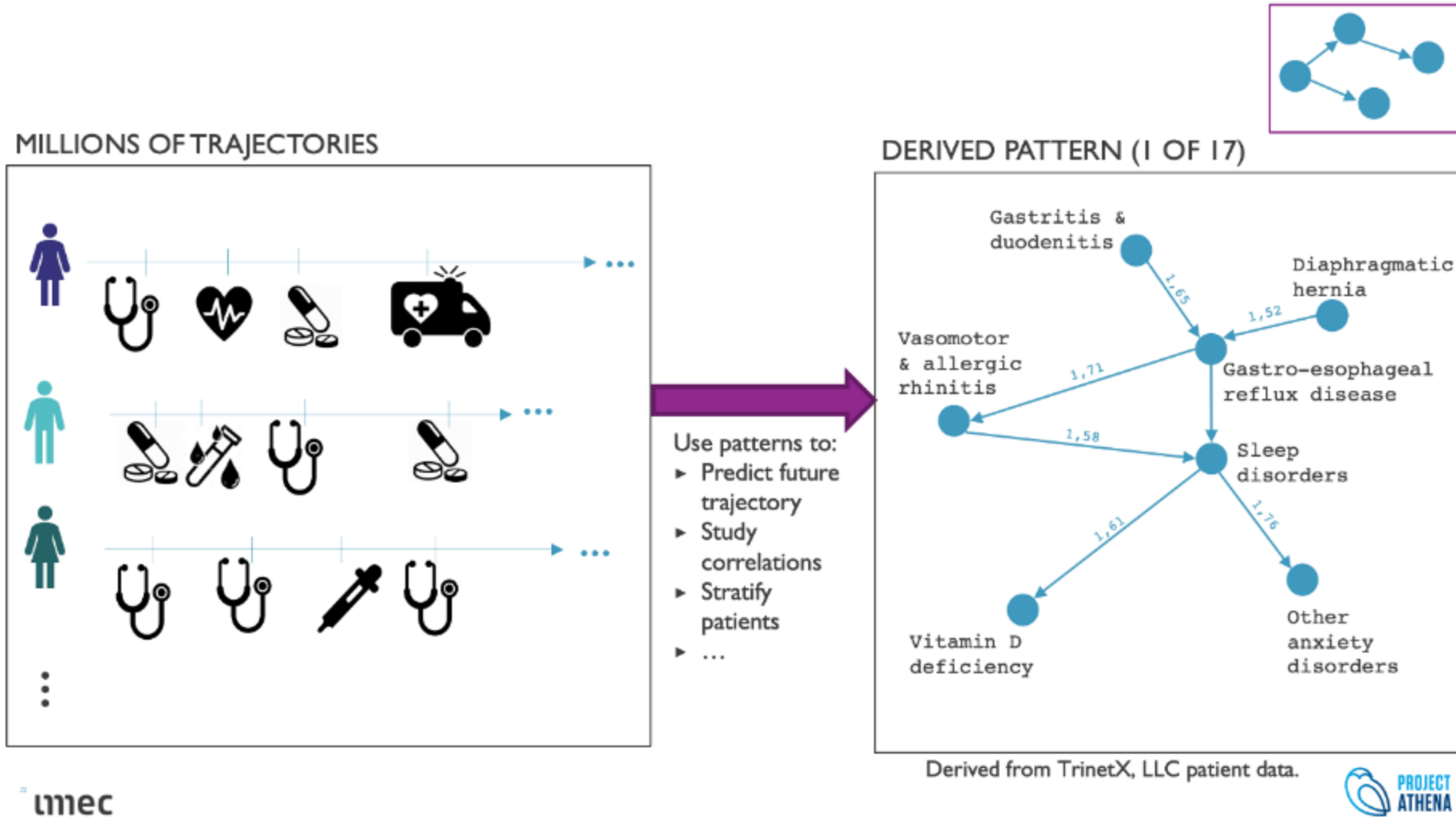


Interactive design

Enables instant analyses
on site data

“It is safe to say this exceeds my expectations!”
Prof. Dr. Michel Delforge (UZ Leuven)

Use case 2: A software package for efficient patient trajectory analysis



Context

Definition of patient trajectory

Context

- Extracting temporal disease trajectories from patient disease histories
- Patient histories are made from diagnostic codes (ICD10, MedDRA, etc)
- E.g:
 - H1: Cough, Broken leg, Dyspnea, Depression, Viral infection, Cough, COPD, ...
 - H2: Cough, BMI>30, Viral infection, Cough, High Blood pressure, Dyspnea, Diabetes, COPD ...
 - H3: BMI>30, Tooth ache, Viral infection, Cough, High blood pressure, Broken leg, Diabetes, ...
 - ... HN

Definition of patient trajectory

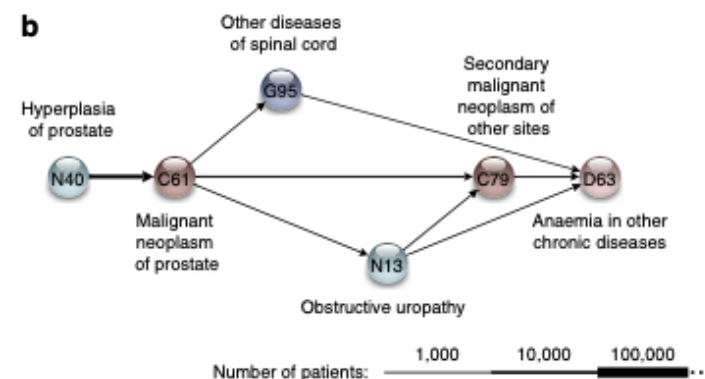
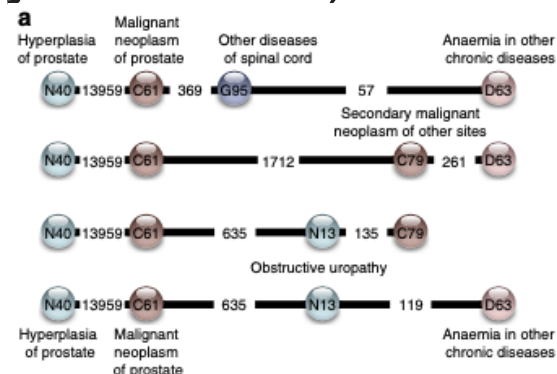
Context

- Extracting temporal disease trajectories from patient disease histories
- Patient histories are made from diagnostic codes (ICD10, MedDRA, etc)
- E.g:
 - H1: Cough, Broken leg, Dyspnea, Depression, Viral infection, Cough, COPD, ...
 - H2: Cough, BMI>30, Viral infection, Cough, High Blood pressure, Dyspnea, Diabetes, COPD ...
 - H3: BMI>30, Tooth ache, Viral infection, Cough, High blood pressure, Broken leg, Diabetes, ...
 - ... HN
 - Possible statistically relevant trajectories:
 - Cough → Dyspnea → COPD
 - BMI>30 → High Blood pressure → Diabetes

Jensen&Brunak algorithm a.k.a “The Danish method”

Context

- Jensen et al, Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients, Nature Communications, 2014
- A “3-step” algorithm:
 - Select diagnosis pairs $A \rightarrow B$ based on RR score
 - Build trajectories $A \rightarrow B \rightarrow \dots \rightarrow Z$ by chaining pairs
 - Cluster trajectories using MCL
- Outcome (from Jensen, 2014):



Project Athena, TriNetX Bladder Cancer data set

Context

- All patients from TriNetX with bladder cancer diagnosis

- Multiple CSV files:

patient_id, sex, race, ethnicity, year_of_birth, age_at_death, patient_regional_location, postal_code, marital_status, reason_yob_missing, month_year_death, source_id

patient.csv

patient_id, encounter_id, code_system, code, principal_diagnosis_indicator, admitting_diagnosis, reason_for_visit, date, derived_by_trinetx, source_id

diagnosis.csv

- Very dense & detailed data set → Challenging for performance & interpretation

Name	#Patients	#Diagnoses	Encoding
TriNetX	126k	47M	ICD09/ICD10
Thin	273k	5M	MedDRA
J&B	6.2M	101M	ICD10

Status

Patient Trajectory Analysis (PTRA) implementation

Status

- An extension of Dtract (°2017)
- Wrote code to parse TriNetX data to input into PTRA:
 - JSON parsing for ICD09→ICD10 mapping
 - XML parsing for WHO ICD10→medical name mapping
 - CSV parsing for exported TriNetX database files
 - CCSR mapping of ICD
- Fixed bugs
- Changed clustering to work on trajectories using Jaccard similarity scores
- Added patient/trajectory filter mechanisms to steer analysis
- Added example Julia code for cluster interpretation
- Performance:
 - Parallelised RR calculation and trajectory building
 - Profiling shows Random Number generation is bottleneck (also after switch to specific library)
- ±4500 LOC in Golang & Julia, on our local git, will be open sourced

Patient Trajectory Analysis (PTRA) paper (in review)

Status

A software package for efficient patient trajectory analysis

Charlotte Herzeel¹, Valerie Vandeweerd², Wouter Botermans², José Felipe Golib Dzib², Ellie D'Hondt¹, Murat Akand³, Frank Van der Aa³, Roel Wuyts¹, Wilfried Verachtert¹,

1 ExaScience Lab, imec, Leuven, Belgium

2 Janssen Research & Development, a division of Janssen Pharmaceutica NV, Beerse, Belgium

3 Department of Urology, University Hospitals Leuven, Leuven, Belgium

* charlotte.herzeel@imec.be

Abstract

We present PTRA, a software package for explorative analysis of disease development. PTRA provides the tools for extracting statistically relevant trajectories from the medical event histories of a patient population. These trajectories can additionally be clustered for visual inspection and identifying key events in disease development. The algorithms of PTRA are based on a statistical method developed previously by Jensen et al, but we contribute several modifications and extensions to implement a practical tool. This includes a new clustering strategy, filter mechanisms for controlling analysis to specific cohorts and for controlling trajectory output, a parallel implementation that executes on a single server rather than an HPC cluster, etc. PTRA is furthermore open source and the code is organized as a framework so researchers can reuse it to analyze new data sets. We illustrate our tool by discussing trajectories extracted from the

Algorithms

Relative Risk scores

PTRA implementation – Step 1: Selection of diagnosis pairs

- RR example:
 - #Patients = 100, #Smoking = 50, #COPD = 35
 - #(Smoking→COPD) = 30, #($\neg$ Smoking→COPD) = 5
- What is the Relative Risk (RR) for COPD when exposed to Smoking?
- Exposure Table: Exposure = Smoking, Outcome = COPD:

	Outcome: Yes	Outcome: No
Exposure: Yes	30 (a)	20 (b)
Exposure: No	5 (c)	45 (d)

- $RR = (a / a + b) / (c / c + d) = 3/5 / 1/50 = 6$
- 6x more chance for COPD when Smoking
- Count outcome only 1x per Patient

RR with Sampling experiments

PTRA implementation – Step 1: Selection of diagnosis pairs

- $\forall$ disease pairs $A \rightarrow B$ calculate RR. What is the risk for developing B when exposed to A?
- $\forall$ pair $A \rightarrow B$:
 - P1 $\leftarrow$ get #patients exposed to A
 - P2 $\leftarrow$ **sample** #P1 similar patients $\neg$ exposed to A
 - Count $\#(A \rightarrow B)$, Count $\#(\neg A \rightarrow B)$
 - Calculate RR

RR with Sampling experiments

PTRA implementation – Step 1: Selection of diagnosis pairs

- $\forall$ disease pairs $A \rightarrow B$ calculate RR. What is the risk for developing B when exposed to A?
- $\forall$ pair $A \rightarrow B$:
 - (P1 $\leftarrow$ get #patients exposed to A
 - P2 $\leftarrow$ **sample** #P1 similar patients $\neg$ exposed to A
 - Count $\#(A \rightarrow B)$, Count $\#(\neg A \rightarrow B)$
 - Calculate RR) $\times 10.000$
 - Calculate p-value

RR with Sampling experiments

PTRA implementation – Step 1: Selection of diagnosis pairs

- $\forall$ disease pairs $A \rightarrow B$ calculate RR. What is the risk for developing B when exposed to A?
- $\forall$ pair $A \rightarrow B$:
 - (P1 $\leftarrow$ get #patients exposed to A
 - P2 $\leftarrow$ **sample** #P1 similar patients $\neg$ exposed to A
 - Count $\#(A \rightarrow B)$, Count $\#(\neg A \rightarrow B)$
 - Calculate RR) $\times 10.000$
 - Calculate p-value
- Perform 10.000 samplings per pair to get within 0.01 of true p-value (400 for 0.05)
 - Time consuming
 - Random sampling necessary to avoid biased comparisons
 - Sampling patients similar in terms of: Age, Sex, (Diagnosis Season, admission type)

RR with Sampling experiments

PTRA implementation – Step 1: Selection of diagnosis pairs

- Do not perform RR calculation for all $A \rightarrow B$ pairs:
 - Filter based on total #A and #B counts as in paper
 - Filter out all diagnosis codes in following ICD10 categories:
 - Pregnancy XV and XVI
 - General symptoms and signs not lined to disease (XVIII)
 - External cause (XIX and XX)
 - Administration (XXI)
- Counting $\#(A \rightarrow B)$ must take into account time constraint between A and B
- ICD10 is a hierarchy with 7 levels; by default we collapse to level 3:
 - I Certain infectious and parasitic diseases
 - A00-A09 Intestinal infectious diseases
 - A01 Typhoid and paratyphoid fevers
 - » A01.0 Typhoid fever, A01.1 Paratyphoid fever A, A01.2 Paratyphoid fever B, A01.3 Paratyphoid fever C, A01.4 Paratyphoid fever, unspec
 - Reduces pairs to consider, otherwise combinatorial explosion

Selecting diagnosis pairs based on RR

PTRA implementation – Step 1: Selection of diagnosis pairs

- Choose all pairs $A \rightarrow B$ |
 - $RR > 1.0$
 - $RR > 1.0$ & $RR \text{ reversed} > 1.0 \rightarrow$ directionality test using Binomial experiment to choose
 - $\#(A \rightarrow B) \geq \text{min number of Patients}$

Building trajectories

PTRA implementation – Step 2

- Chaining selected RR pairs $A \rightarrow B$:
 - E.g. $A \rightarrow B$ and $B \rightarrow C$ form a new T: $A \rightarrow B \rightarrow C$
 - For each extension $A \rightarrow B \rightarrow C$ check that $\#P$ following the new T $\geq$ min required $\#P$
- Parallelized $\forall A \rightarrow B$

Trajectory clustering

PTRA implementation – Step 3

- A typical run produces trajectories with a lot of overlap:

Other anemias → Cardiomyopathy → Chronic ischemic heart disease

Other anemias → Cardiomyopathy → Heart failure

Other anemias → Heart failure → Anemia in chronic diseases classified elsewhere Chr

Other anemias → Heart failure → Chronic ischemic heart disease → Atrial fibrillatio

Other anemias → Heart failure → Chronic ischemic heart disease → Hypertensive chron

Other anemias → Acute kidney failure → Anemia in chronic diseases classified elsewh

Other anemias → Acute kidney failure → Disorders of mineral metabolism Chronic kidn

Other anemias → Acute kidney failure → Heart failure → Chronic ischemic heart disea

Other anemias → Acute kidney failure → Iron deficiency anemia → Chronic kidney dise

Other anemias → Acute kidney failure → Hypertensive chronic kidney disease → Anemia

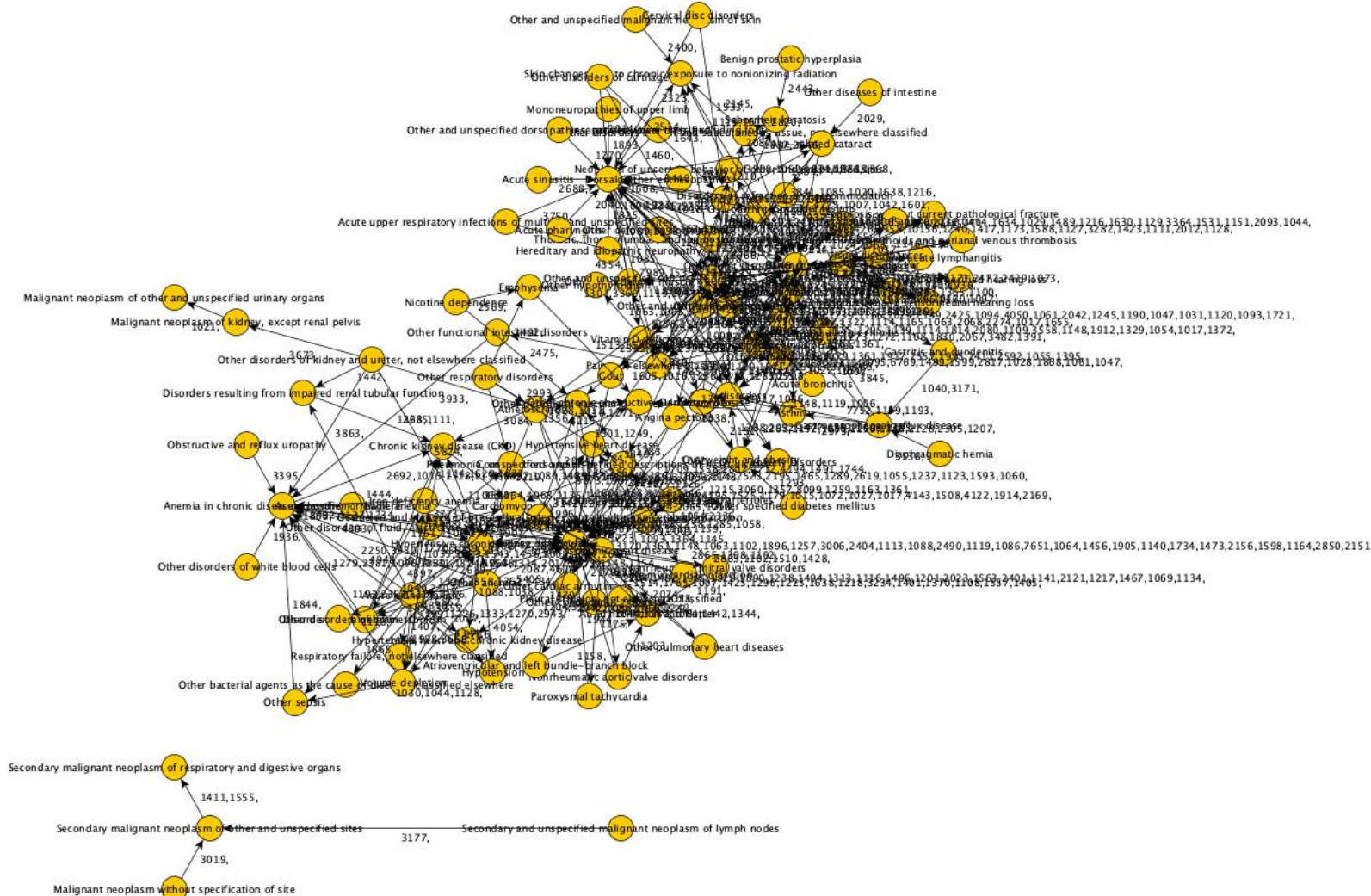
Other anemias → Acute kidney failure → Hypertensive chronic kidney disease → Anemia

Other anemias → Acute kidney failure → Hypertensive chronic kidney disease → Heart

- Clustering can help

Trajectory clustering

PTRA implementation – Step 3 – clustering can help



Trajectory clustering

PTRA implementation – Step 3 – Jensen & Brunak approach

- Cluster the diagnostic codes
 - Make a graph: diagnostic codes = nodes, RR pairs = edges
 - Place a similarity score on each edge using the Jaccard Index:
 - $\text{Jaccard}(A \rightarrow B) = (\#T \mid A \rightarrow B \in T) / (\#T \mid A \in T + \#T \mid B \in T - \#T \mid A \rightarrow B \in T)$
 - Normalized value $\in [0, 1] \forall$ pairs $A \rightarrow B$
 - Cluster the graph with Markov Chain Clustering (MCL), resulting in diagnostic code clusters
- Assign T to the cluster C if $\forall A \in T \mid A \in C$
- This leads to many T not assigned to any cluster:
 - Allow to assign T with all $A \in T$ except one
 - Manually merge clusters that seems good to merge
- For the TriNetX data set, more than 2/3 of T unassigned and manual merge is impossible

Trajectory clustering

PTRA implementation – Step 3 – Our direct clustering approach

- Cluster the trajectories directly
 - Make a graph: trajectory IDs = nodes, fully connected
 - Place a similarity score on each edge using the Jaccard Index to measure the similarity between T:
 - View T as sets of Diagnostic codes
 - $J(T_1, T_2) = |T_1 \cap T_2| \div |T_1| + |T_2| - |T_1 \cap T_2|$
 - Cluster the graph with Markov Chain Clustering (MCL), resulting in trajectory clusters
- Assign T to the cluster C if $TID \in C$
- Allows assigning all T to a cluster without manual intervention
- Jaccard Index ignores order of events, might be better to use another similarity score, e.g. Dynamic Time Warping?

Results: General disease development patterns in bladder cancer patients in TriNetX Research

PTRA outputs

Example

- **Parameters:**

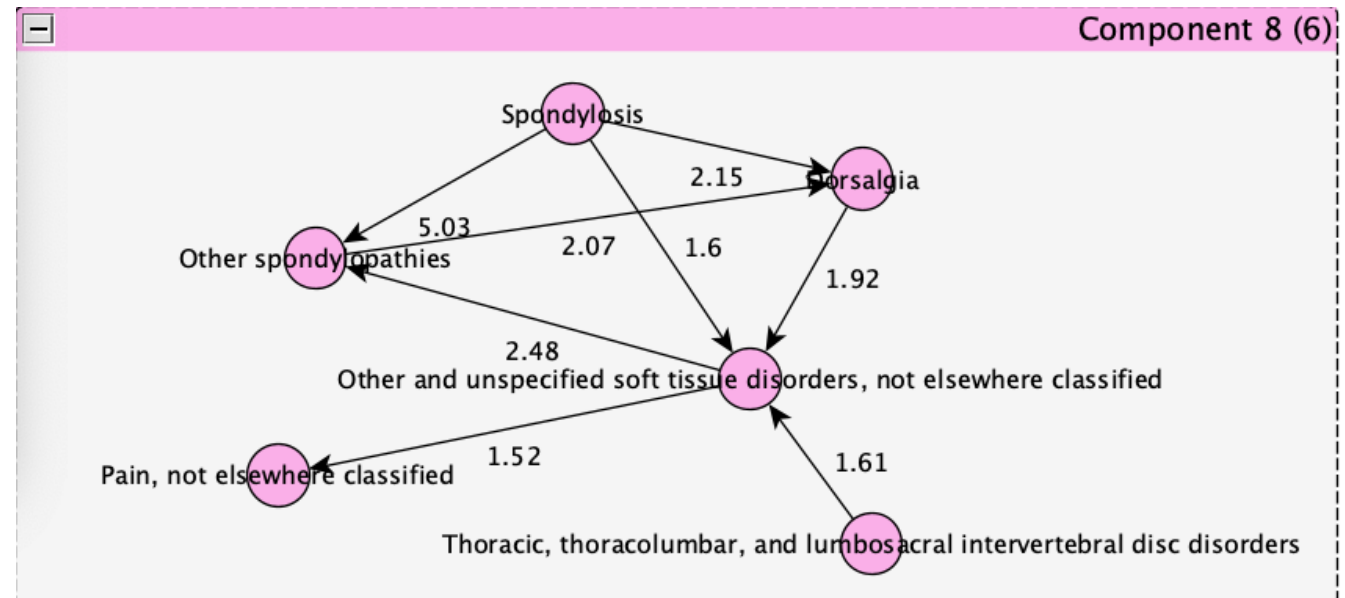
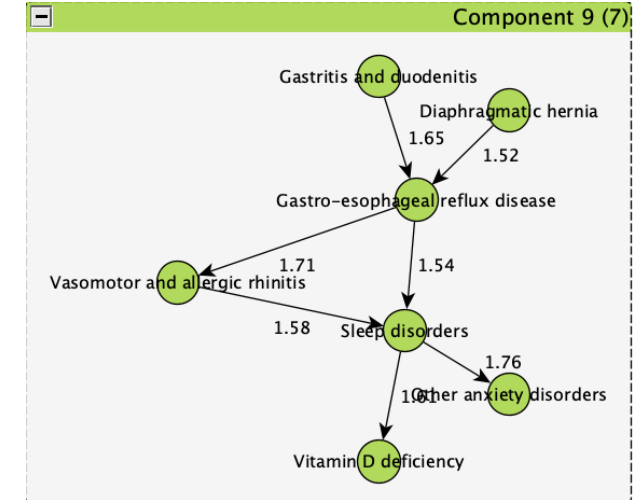
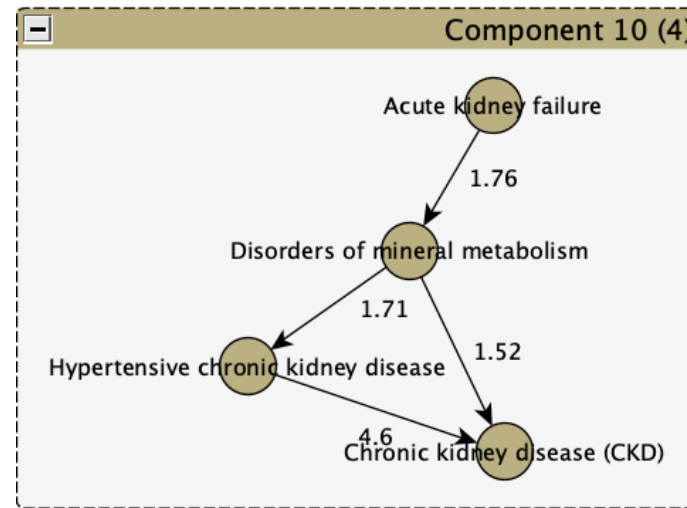
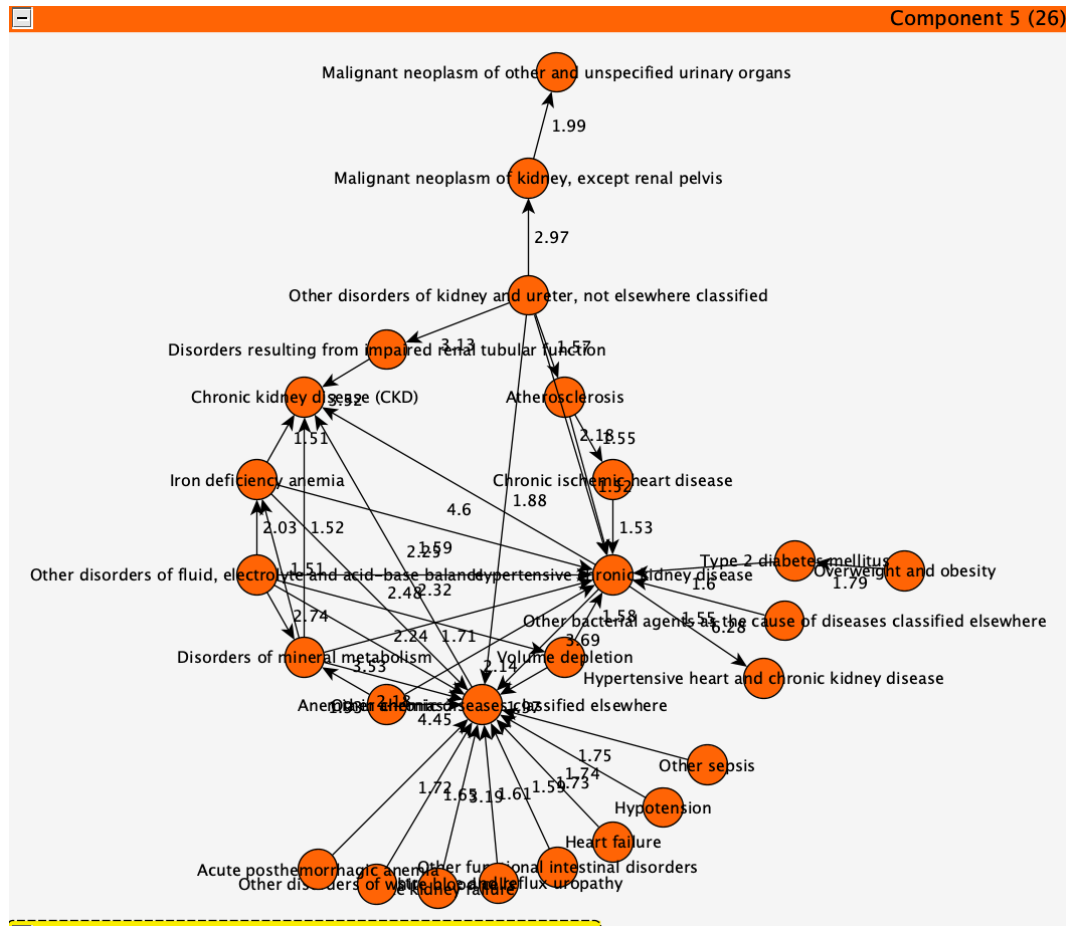
```
ptr patient.csv icd10cm_tabular_2022.xml diagnosis.csv ./result/ --  
RR 1.50 --nofAgeGroups 10 --minTrajectoryLength 3  
--maxTrajectoryLength 5 --lvl 2 --minPatients 1000 --iter 400
```

- **Runtime:** 19h14m on 1 server with 2x18-core Intel Xeon E5-2699v3 CPUs, 256GB RAM
- Incremented RR, minPatients parameters, lowered iter to 400 to handle data density.
- **Outputs:**
 - 1036 pairs: 1 tab file
 - 775 trajectories: 1 tab file
 - 17 clusters (direct clustering): 1 GML file (RR), 1 GML file (#P), 1 CSV file assigning P, T to cluster, 1 CSV file with patient info [Our method]
 - 24 clusters/332 trajectories not clustered (disease clustering): 1 GML file [Jensen&Brunak method]

Trajectories unclustered

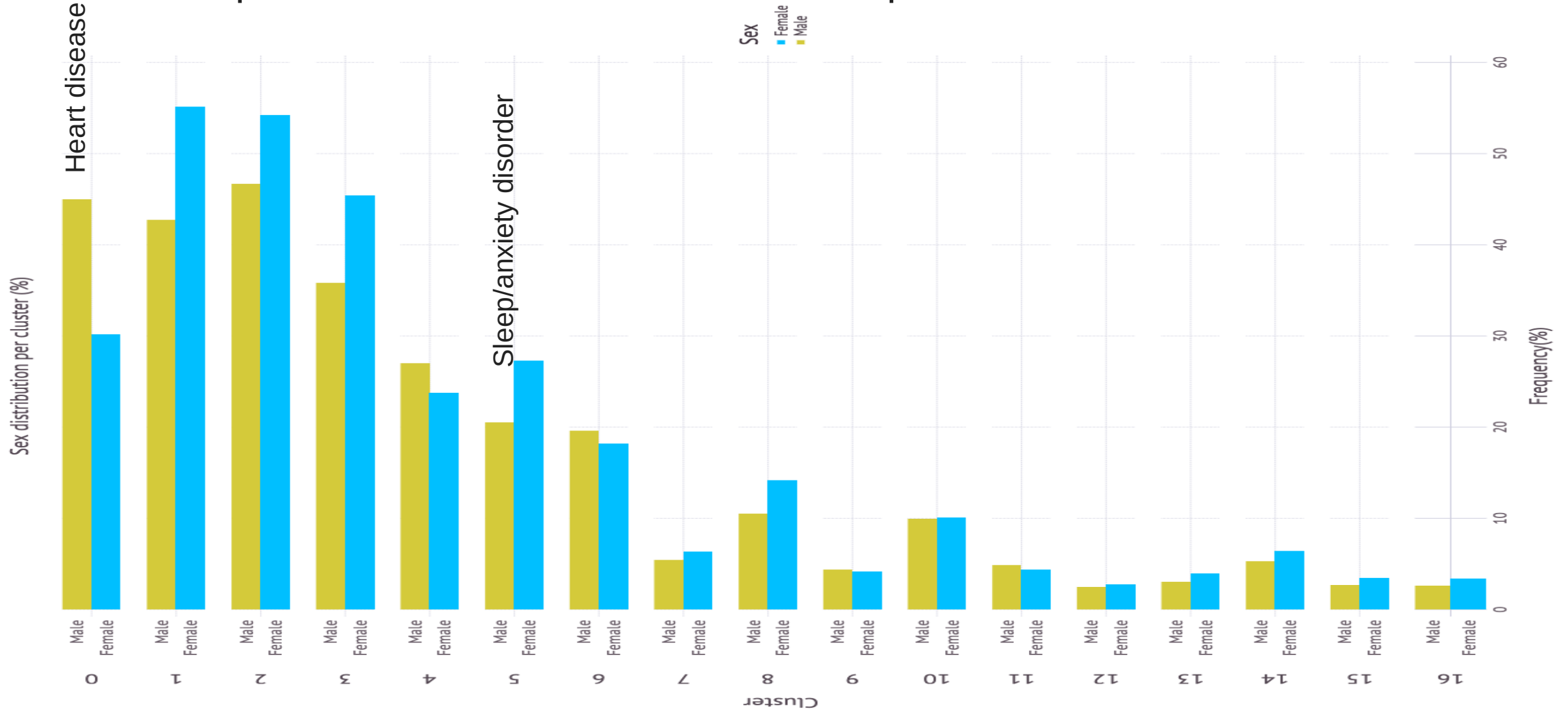


General PTRA outputs



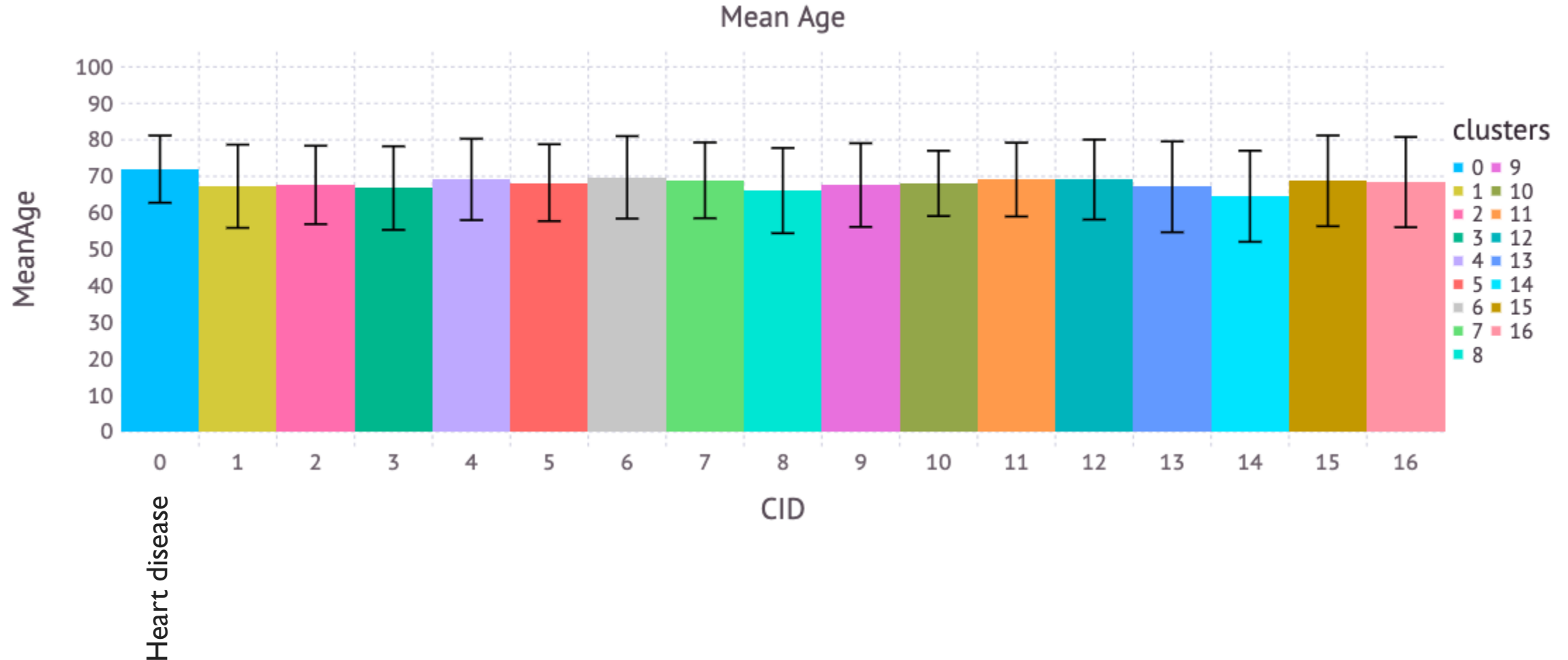
PTRA outputs

Cluster interpretation – Distribution of male/female per cluster



PTRA outputs

Cluster interpretation – Mean age per cluster<



PTRA outputs

Full bladder cancer cohort

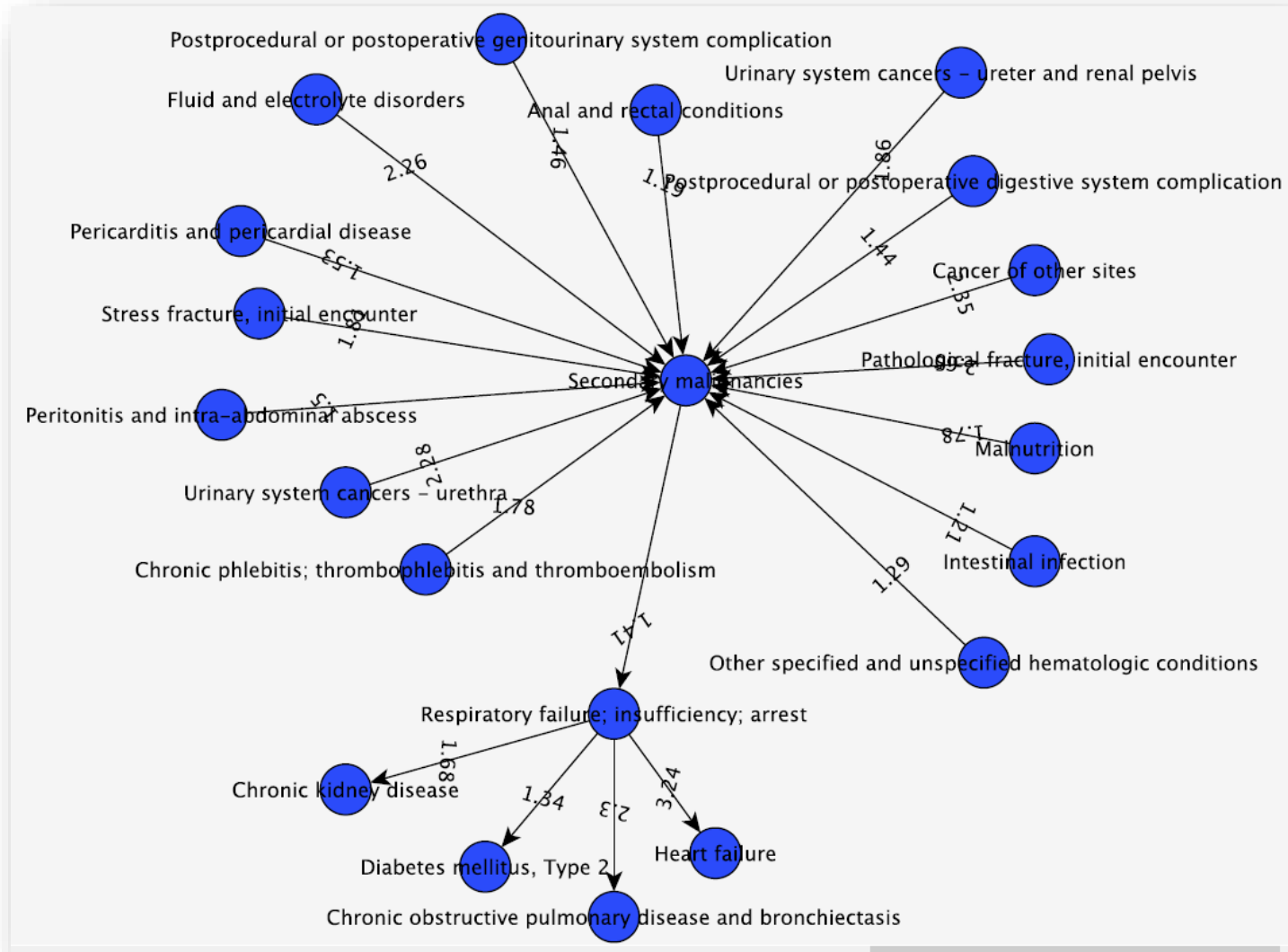
- On the full bc cohort, PTRA outputs clusters of highly prevalent, well-known chronic diseases in an aging population such as bladder cancer patients
 - Cardiovascular disease, diabetes, dorsalgia, lung diseases, etc
 - Medically valid, but not necessarily bladder cancer specific
- PTRA features to focus on bc specific trajectories:
 - CCSR categorization of ICD10 codes
 - Addition of bladder cancer specific treatments to the diagnosis histories
 - Radical cystectomy, MVAC chemo, intravesical therapy
 - Use trajectory filters to include at least 1 bc diagnosis (ICD code) or treatment
 - C67, C77, C78, C79 or radical cystectomy, MVAC chemo, intravesical therapy

Results: Bladder cancer development patterns in TriNetX Research

PTRA outputs

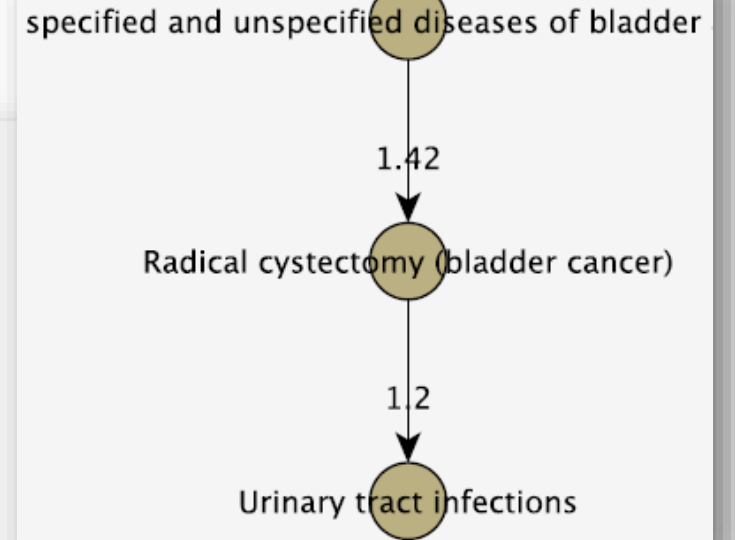
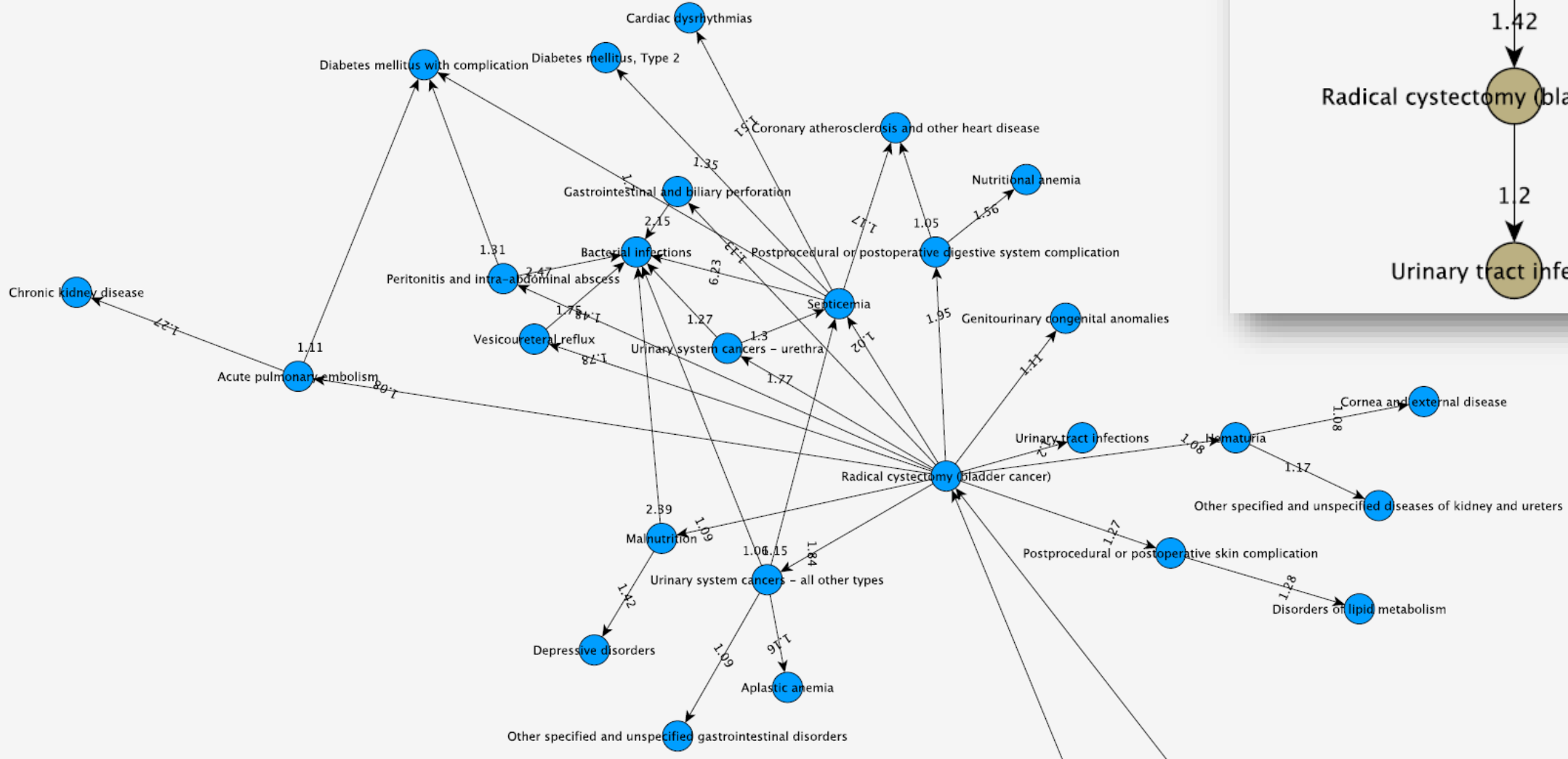
Specific treatment trajectory analysis

Secondary malignancy



PTRA outputs

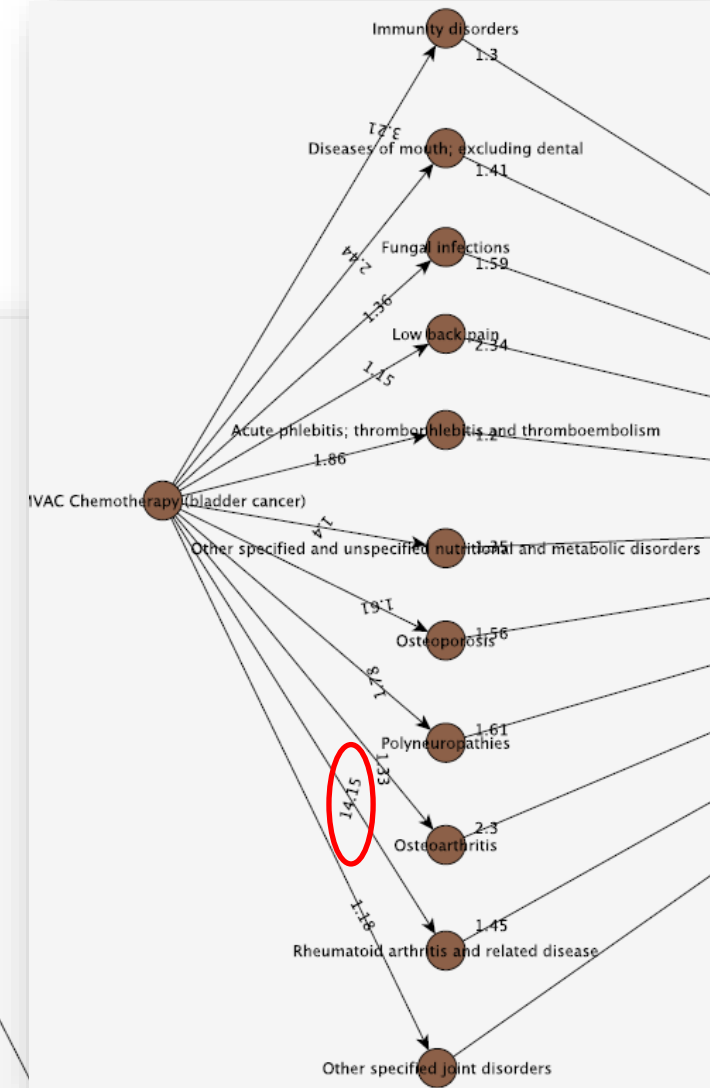
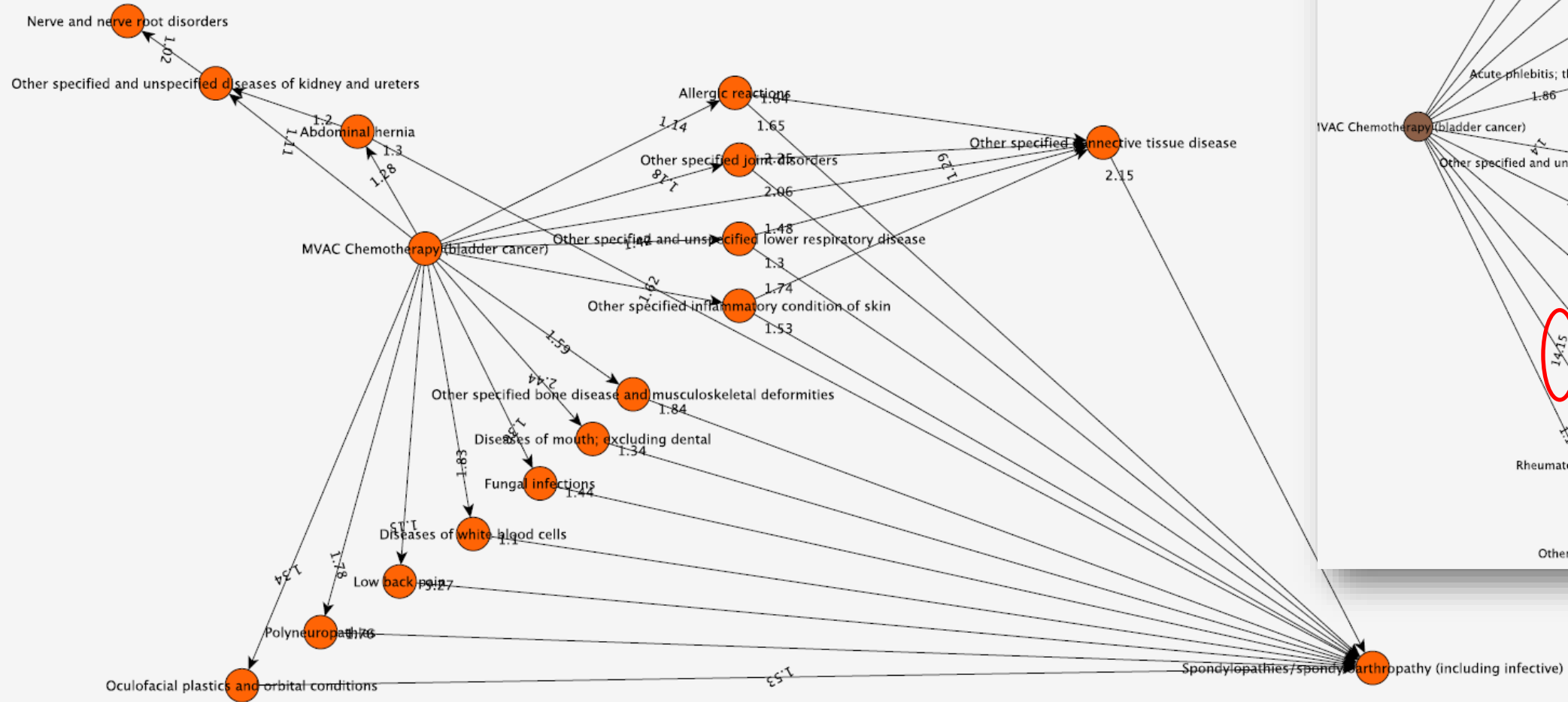
Specific treatment trajectory analysis



Radical
Cystectomy

PTRA outputs

Specific treatment trajectory analysis

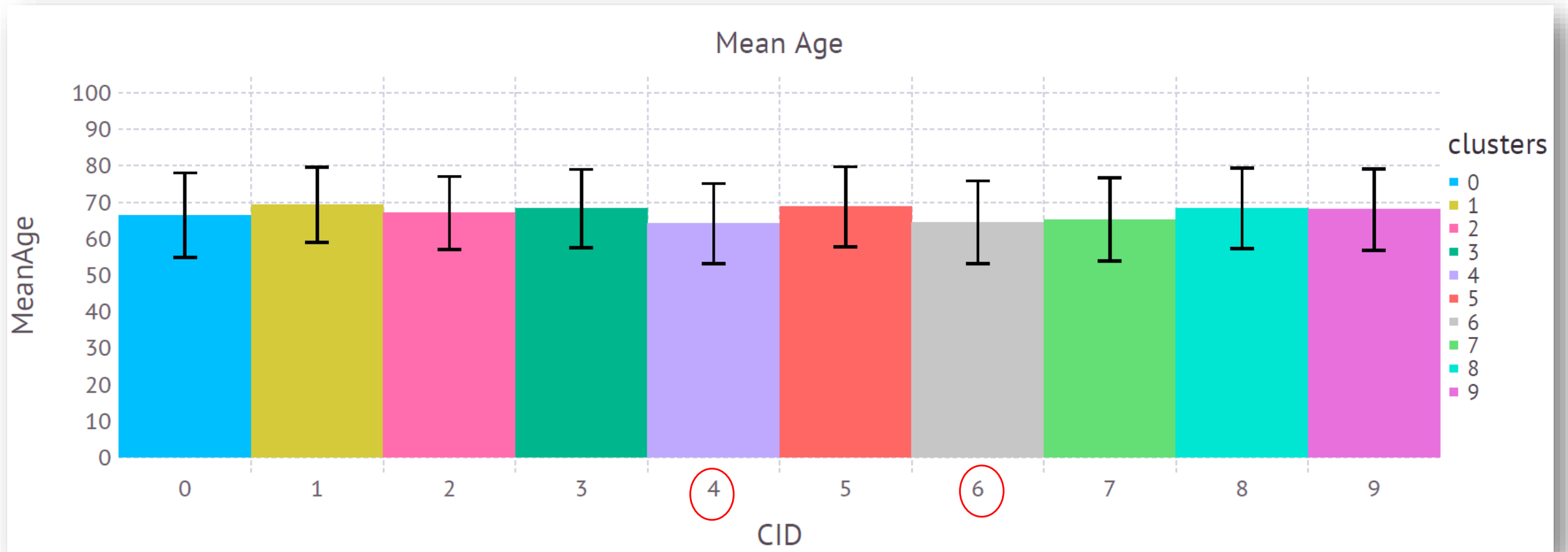


MVAC

PHARMACEUTICAL COMPANIES OF

PTRA outputs

Specific treatment trajectory analysis



Conclusions

Summary & future work

- Patient Trajectory Analysis tool that works on TriNetX:
 - Parsing of ICD10 hierarchy, ICD9->ICD10 mappings, TriNetX exported CSVs
 - Changed clustering to work on trajectories
 - Added support for cluster interpretation
 - Generalized medical events can be used as inputs to the tool:
 - Added Bladder cancer-specific treatments to diagnosis histories
 - Added filter mechanisms to focus analysis on bc specific trajectories
 - Performance:
 - Parallel RR calculation and trajectory building
 - Sampling iterations are the expensive part
 - Random number generation is costly
 - $\pm$ 4500 LOC in Golang & Julia, plan is to open source
 - Code is organized as a framework
 - Practical command-line tool
 - Evaluation of our tool on TriNetX Research bladder cancer cohort
 - Draft paper for PLOS Digital Health
- Future work
 - Improve computational performance
 - Improve clustering
 - Improve interpretation
 - Interactive GUI