A Deep Learning Approach to Identifying Missing Hierarchical Relations in SNOMED CT

Rashmie Abeysinghe

Mentor: Dr. Olivier Bodenreider

U.S. National Library of Medicine
Lister Hill National Center for Biomedical Communications

Outline

- SNOMED CT
- Terminology Quality Assurance
- Non-Lattice Subgraphs and structural-lexical auditing methods
- Graph Neural Networks
- Sample Generation
- Architecture of the neural network
- Results

SNOMED CT

- Developed by SNOMED International*
- Largest clinical healthcare terminology in the world
- More than 350,000 active concepts
- Has 19 top-level sub-hierarchies

*http://www.snomed.org/

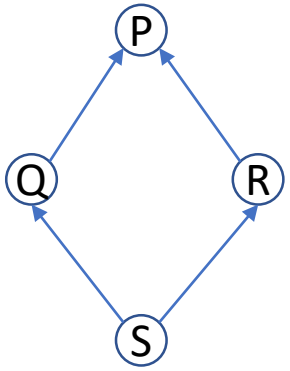
Terminology Quality Assurance (TQA)

- Errors exist in terminologies*
- TQA: Essential part of terminology management lifecycle
- Manual auditing: labor intensive & time consuming
- Automating TQA: active area of research

*Rector AL, Brandt S, Schneider T. Getting the foot out of the pelvis: modeling problems affecting use of SNOMED CT hierarchies in practical applications. Journal of the American Medical Informatics Association. 2011 Apr 21;18(4):432-40.

Non-Lattice Subgraphs (NLSs)

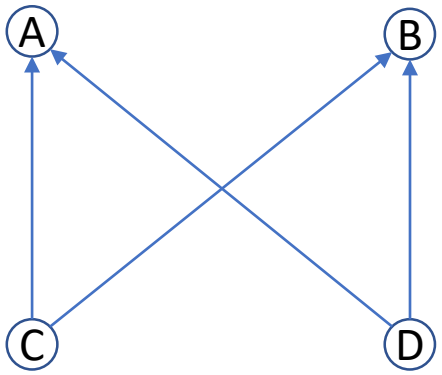
- Lattice – a desirable property for a well-formed terminology*
- Lattice – a DAG such that any two nodes have a unique maximal common descendant as well as a unique minimal common ancestor



Lattice subgraph

Upper bound

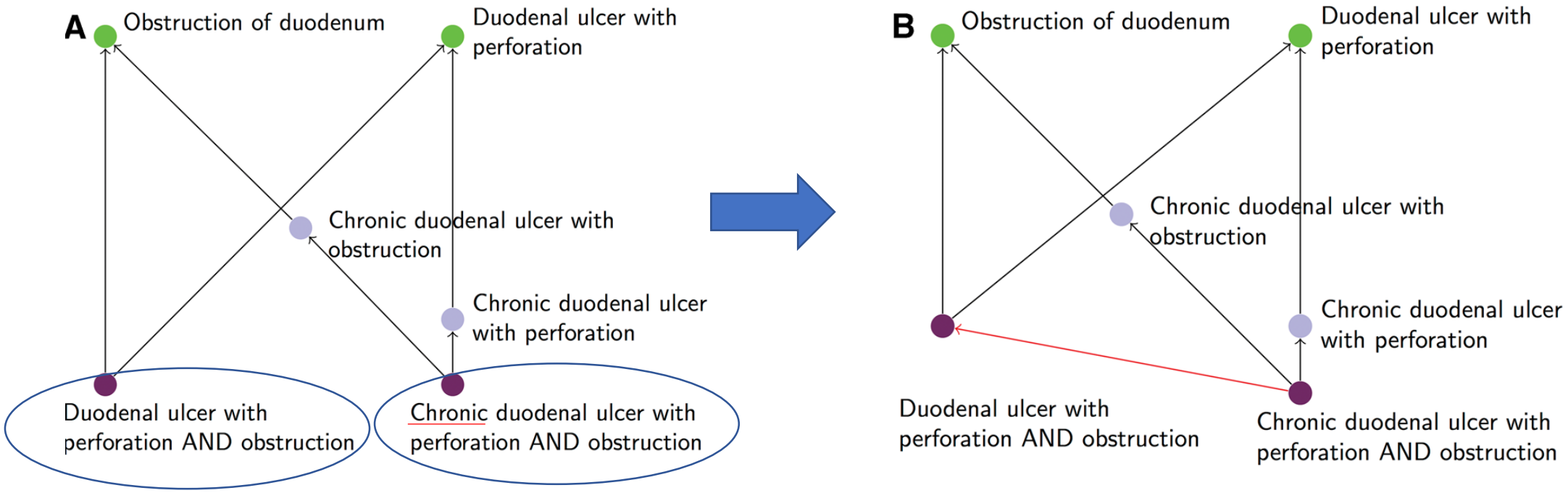
Lower bound



Non-lattice subgraph

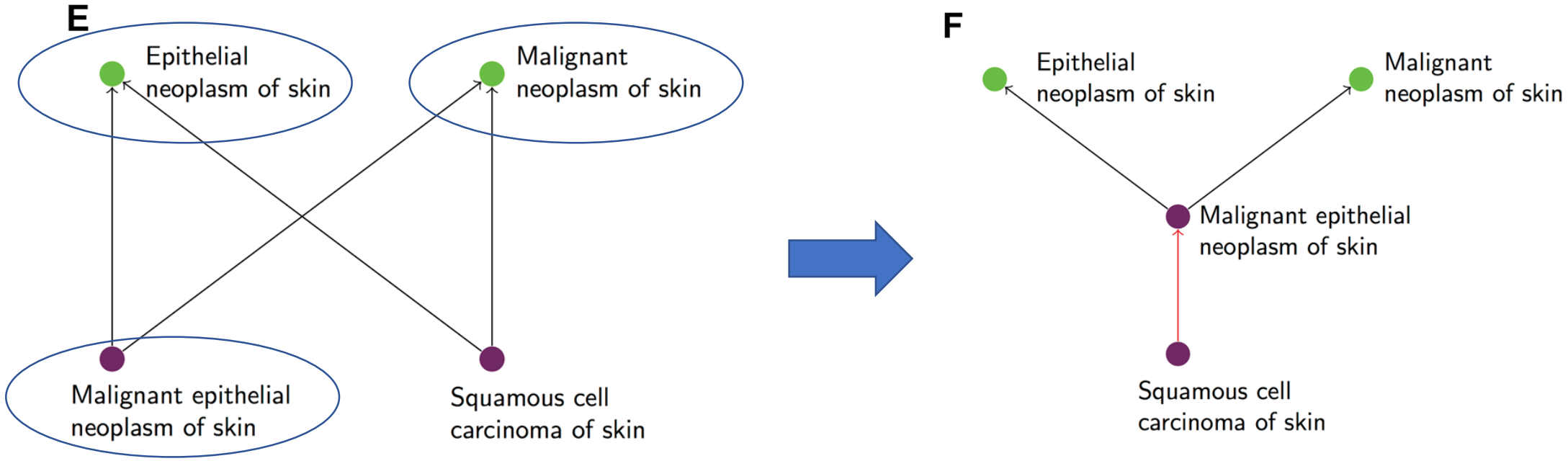
*Zhang GQ, Bodenreider O. Large-scale, exhaustive lattice-based structural auditing of SNOMED CT. AMIA Annual Symposium Proc. 2010;922-26.

Containment lexical pattern



*Cui L, Zhu W, Tao S, Case JT, Bodenreider O, Zhang GQ. Mining non-lattice subgraphs for detecting missing hierarchical relations and concepts in SNOMED CT. JAMIA. 2017 Jul 1;24(4):788-798

Union lexical pattern



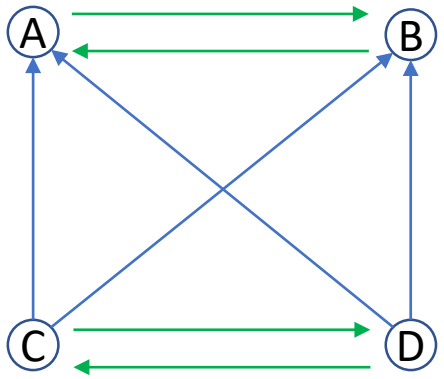
*Cui L, Zhu W, Tao S, Case JT, Bodenreider O, Zhang GQ. Mining non-lattice subgraphs for detecting missing hierarchical relations and concepts in SNOMED CT. JAMIA. 2017 Jul 1;24(4):788-798

Lexical Patterns

- Manually curated
- Majority of NLSs does not exhibit any of the lexical patterns currently identified
- Not sustainable

Objective

- Develop a deep learning approach that can identify the existence of and the direction of a missing is-a among two concepts in the upper/lower bounds of a non-lattice subgraph in SNOMED CT



Upper bound

Lower bound

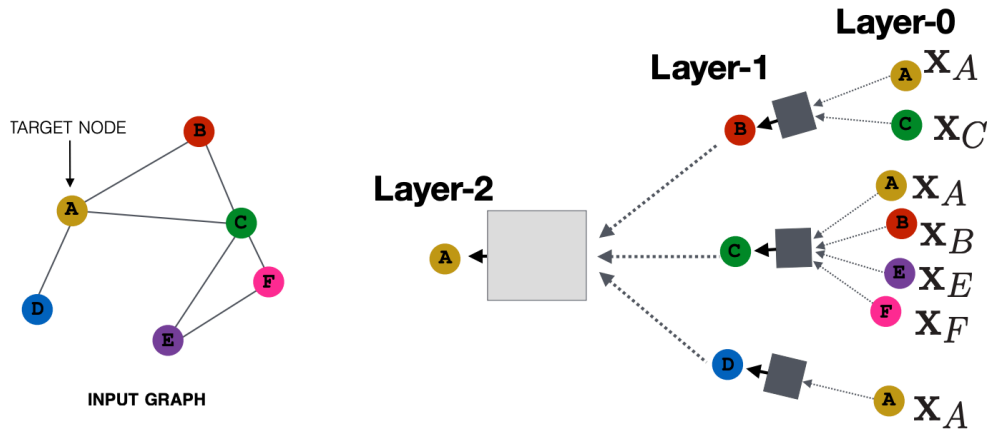
Handling Graph Data with Neural Networks

- Traditional neural network architectures does not handle graph data that well

- Graph data are irregular:
  - Each graph has a variable size of nodes
  - Each node has a different number of neighbors
  - Neighbors of a node cannot be ordered in a graph

Graph Neural Networks (GNNs)

- Key idea: Generate node embeddings based on local neighborhoods.
- Intuition: Nodes aggregate information from their neighbors using neural networks

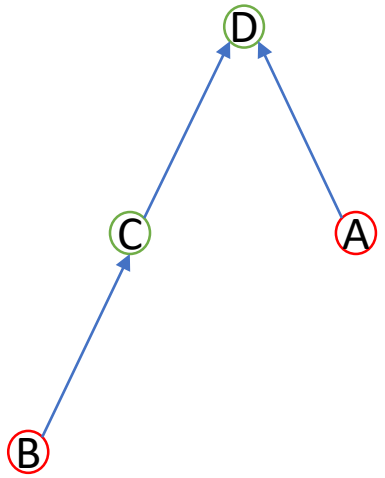


*Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907. 2016 Sep 9.

Sample Generation

- 117,625 concepts in Clinical Findings
- 210,349 relations
- 14 billion non-relations!

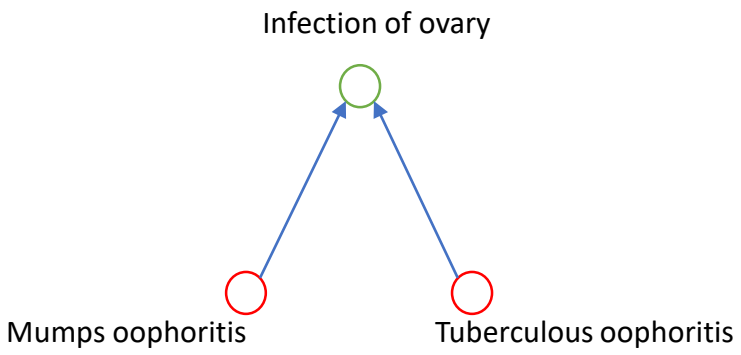
- Most non-relations may be uninteresting
- Only non-relations types that occur in NLSs needs to be used
- To identify the non-relation type, we use edge-separation:
  - Edge-separation: sum of edges from each concept to a common ancestor



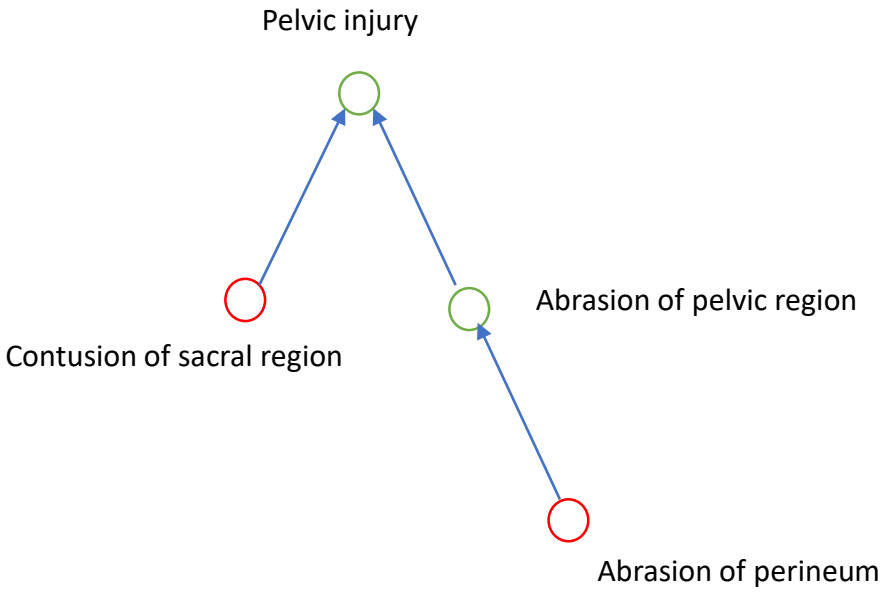
Edge-separation of A & B = 3

Sample Generation

- Edge-separation 2



Edge-separation 3



Edge-separation of non-relations in NLSs

Edge-separation	Number of non-edges	Percentage
2	24514	63.04%
3	5342	13.74%
4	3500	9.00%
5	2148	5.52%
6	1368	3.52%
7	926	2.38%
8	568	1.46%
9	290	0.75%
10	148	0.38%
11	64	0.16%
12	12	0.03%
13	2	0.01%
14	2	0.01%

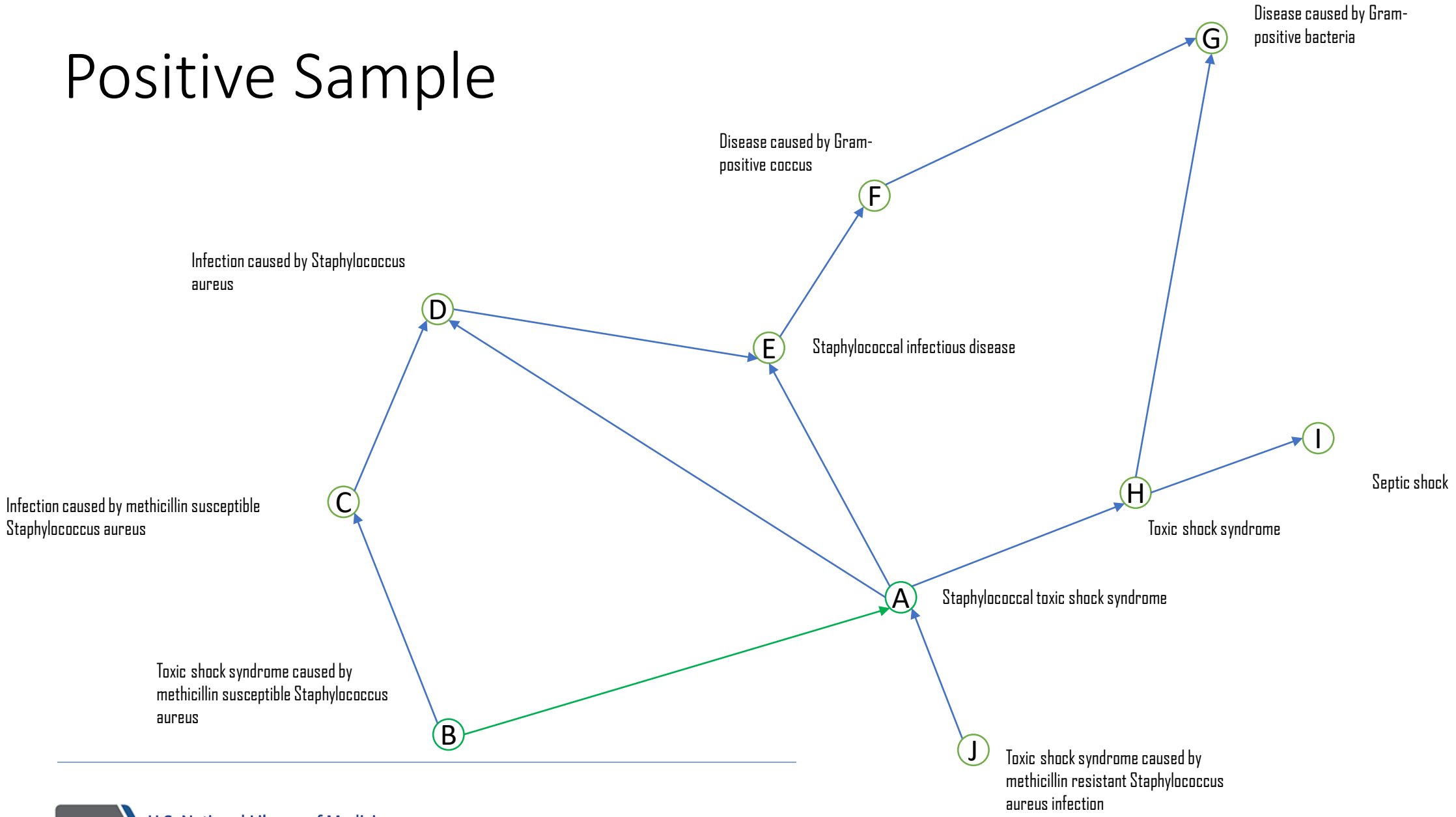
Sample Generation

- Positive samples:
  - 210,349 relations
- Negative samples:
  - 210,349 non-relations with same edge-separation distribution as NLS non-relations
  - 210,349 non-relations obtained by reversing existing is-a relations
- 90% for training and validation:
  - 567,942 samples
- 10% for testing:
  - 63,105 samples

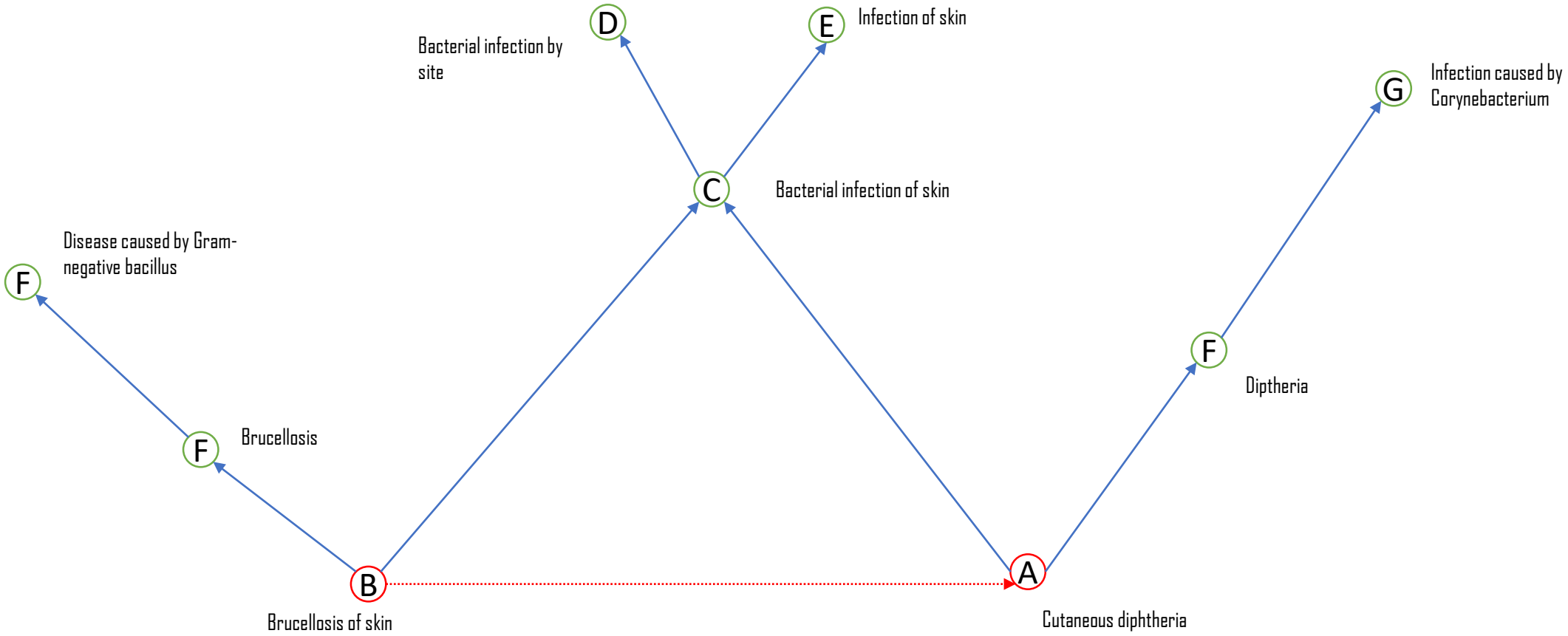
Sample Subgraph Generation

- Context of a concept: consists of ancestors and descendants up to two levels.
- Concept-pair of the sample and the concepts in their contexts will define a subgraph

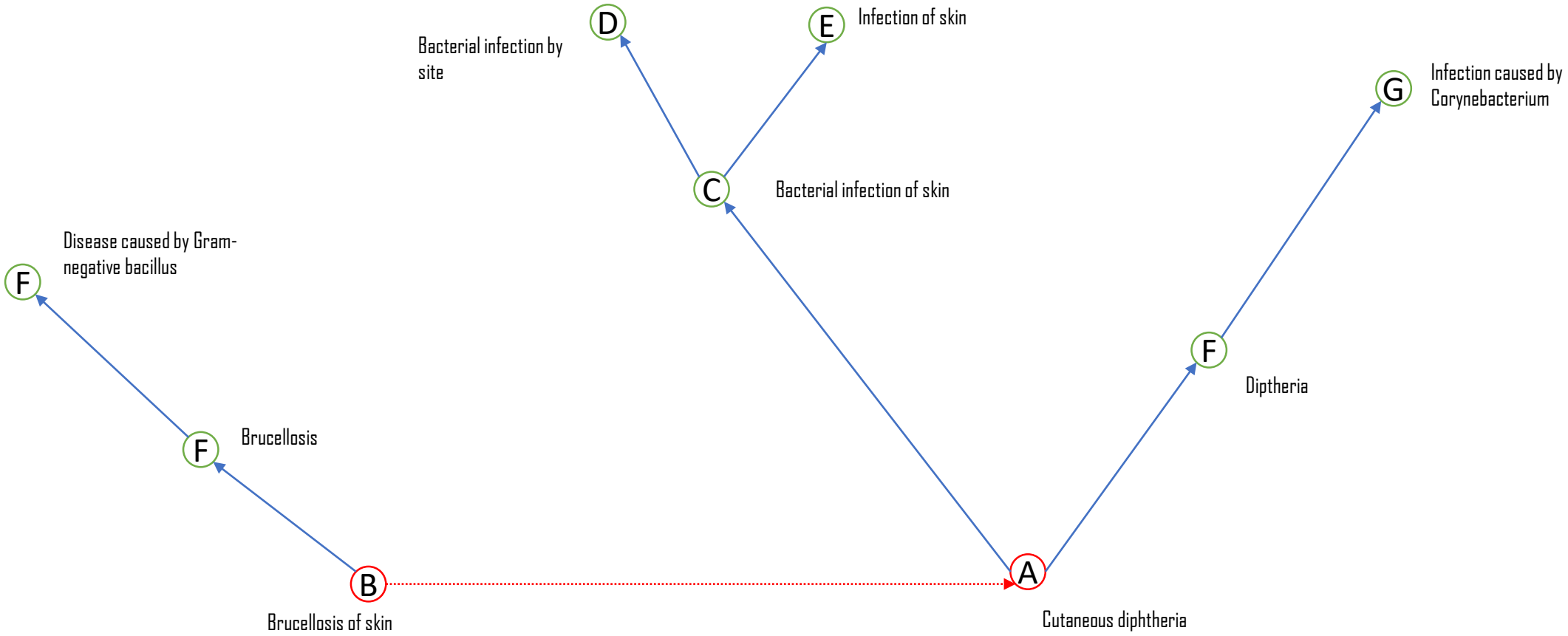
Positive Sample



Negative Sample



Negative Sample

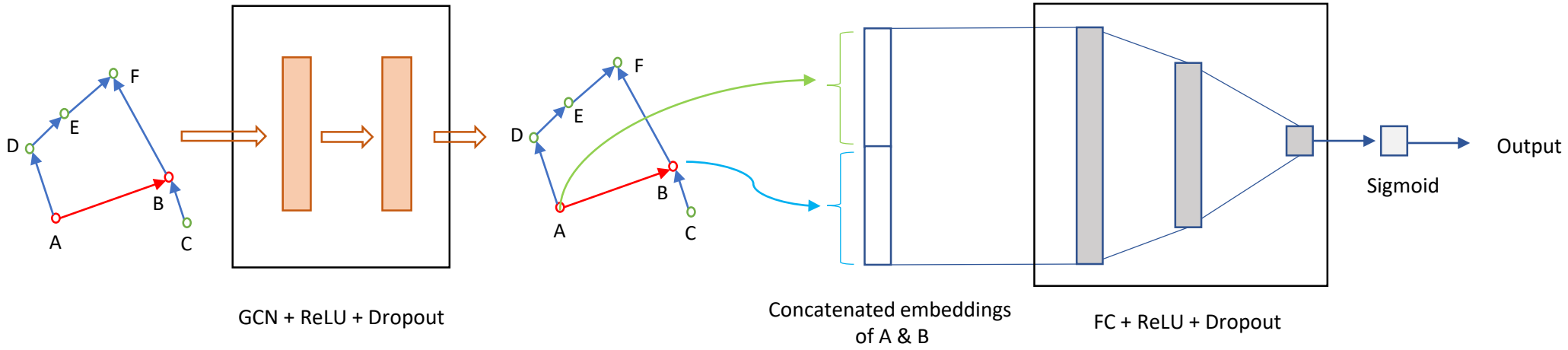


Node Features (at Layer-0 of GNN)

- Trained a Doc2Vec* model to obtain embeddings for preferred terms of each concept
  - Doc2Vec: Unsupervised algorithm to generate vectors for sentences/paragraphs/documents etc.
  - Vector Size: 150

*Le Q, Mikolov T. Distributed representations of sentences and documents. In International conference on machine learning 2014 Jan 27 (pp. 1188-1196).

Architecture of the Network



Training

- GNN library: Deep Graph Library (DGL): A python package dedicated to deep learning on graphs. Built atop Pytorch.
- Ran on NVIDIA Tesla K20X GPUs (Biowulf)
- Loss function: Binary Cross Entropy
- Optimizer: Adam
- Learning rate: 0.001
- Batch size: 128
- 100 epochs
- Performed 6-fold cross validation

Performance

	Cross validation	Test set
Precision	0.8629	0.8439
Recall	0.8475	0.8291
F1	0.8552	0.8364

Relations identified correctly (true positives)

Child	Parent
Duodenal papilla not found	Digestive system finding
Cyproterone adverse reaction	Antineoplastic adverse reaction
Mosaic trisomy 5 syndrome	Anomaly of chromosome pair 5
Acute hepatitis	Inflammatory disease of liver

Non-relations identified correctly (true negatives)

Child	Parent
Chronic gingivitis	Chronic fibrous gingivitis
Finding of urine drug level	Acetaminophen in urine
Open fracture of proximal phalanx of left thumb	Closed fracture thumb proximal phalanx, head
Dextrotransposition of aorta	Transposition of aorta

Relations not identified (false negatives)

Child	Parent
Traumatic arthropathy of metacarpophalangeal joint	Traumatic arthropathy of the hand
Primary basal cell carcinoma of right lower limb	Basal cell carcinoma of lower extremity
Pulmonary aspiration of gastric contents	Pulmonary aspiration of fluid
Lesion of radial nerve	Radial neuropathy

Non-relations not identified (false positives)

Child	Parent
Implanted defibrillator generator infection	Infected pacemaker
Open fracture of right angle of mandible	Open fracture of zygoma
Finding of cochlear function	Cochlear microphonic
Aluminum hydroxide overdose	Kaolin overdose

Applying the trained model on NLSs

- 7,141 small (size 4,5,6) non-lattice subgraphs in Clinical Finding subhierarchy
- 37,404 non-relations in lower and upper bounds
- 11,943 missing is-a predicted

Comparison with lexical patterns

Pattern	# missing is-a obtained by pattern	# common missing is-a obtained by classifier
Containment	406	150
Union	110	43
Intersection	542	370

*Cui L, Zhu W, Tao S, Case JT, Bodenreider O, Zhang GQ. Mining non-lattice subgraphs for detecting missing hierarchical relations and concepts in SNOMED CT. JAMIA. 2017 Jul 1;24(4):788-798

Conclusions and Future Work

- We employ a deep learning approach to identify missing relations in SNOMED CT
- Samples are generated based on existing relations and non-relations in SNOMED CT
- After training the model, we apply it to lower and upper bounds of non-lattice subgraphs to identify missing relations:
- Future work: train the model on relations and non-relations similar to those that appear in non-lattice subgraphs

Acknowledgement

- Dr. Olivier Bodenreider
- Dr. Licong Cui
- Dr. Paul Fontelo
- Dr. Vinh Nguyen
- Joey Yip
- Karan Luthria

Thank you!