Emerging techniques for aggregating SNOMED CT encoded patient

Ming Zhang | Donna Truran | Michael Lawley

INTRODUCTION

SNOMED International
SNOMED CT EXPO 2020
Virtual Conference October 8-9

Abstract

We have developed and evaluated a method named Iterated Least Common Ancestor (ILCA). It is a dynamic and data-driven way to perform aggregation of SNOMED encoded data, without the need to pre-define static categories or 'reporting' groups that to manually specify. In general, the approach operates on a set of unique SNOMED CT concepts and produces a candidate set of aggregating concepts. This method allows the '**data to speak for itself**'. It is then possible to run iteratively until the clinical terminologist gets a useful aggregation result. Several heuristic rules are predefined to stop the aggregation producing overly general categories.

Background

The SNOMED user community, especially clinicians, are increasingly looking to derive new information and knowledge about their clinical practice from the EMR data they have collected for their patients, encoded in SNOMED. The use cases they bring forward focus on cohort identification, prognosis and outcomes analyses, clinic performance monitoring, and grouping data to obtain a holistic overview of their patient caseload. The most prominent requirement is for SNOMED CT encoded Patient data to be aggregated into "clinically meaningful" categories, while still allowing access to the original and specific SNOMED concepts assigned to each patient case.

Topic

- What is Iterated Least Common Ancestor (ILCA)
- Prepare Input concept for ILCA calculation form exiting data collection.
- ILCA calculation for two input concepts.
- ILCA for a set of SNOMED CT concepts.
- Run the method iteratively.
- Working with Patient Data.
- Data analysis result.

Emerging techniques for aggregating SNOMED CT encoded patient

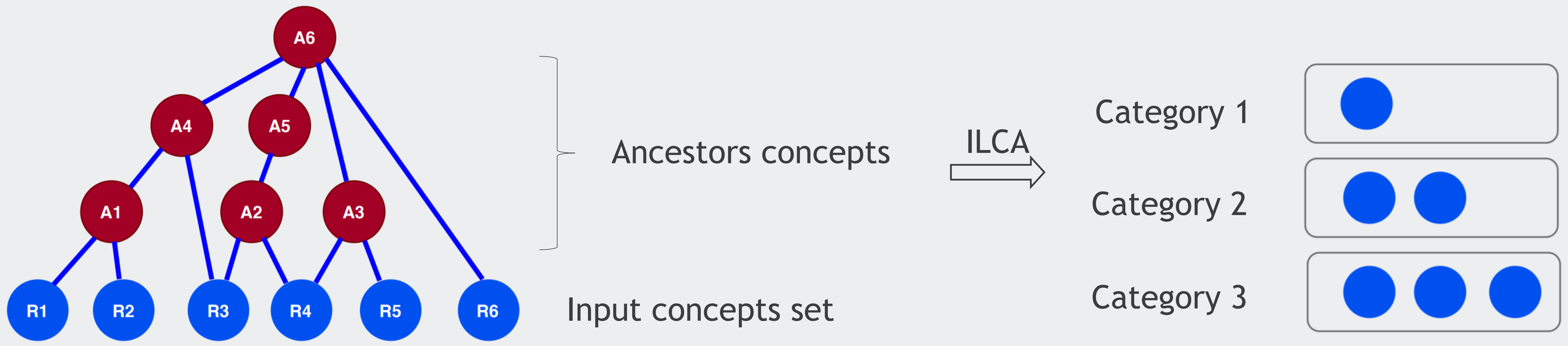
Ming Zhang | Donna Truran | Michael Lawley

SNOMED International
SNOMED CT EXPO 2020
Virtual Conference October 8-9

Pari-wised LCA – Pari-wised Least Common Ancestor

The ILCA method is to find the higher-level SNOMED CT concepts which can be used for categorization purposes, from an input set of SNOMED CT concepts.

The patient data is a set of SNOMED CT concept ids and the initial categorization needs to be performed through the SNOMED CT hierarchy, and this method is to be used after the initial categorization has been calculated. A disjointed set of SNOMED CT concepts are the input of this method.



The process starts to calculate the least common ancestor (LCA) for each pair of concepts in the input sets, briefly, calculated Pari-wised LCA, the below example shows that finding Pair-wise LCA for R1 and R3.

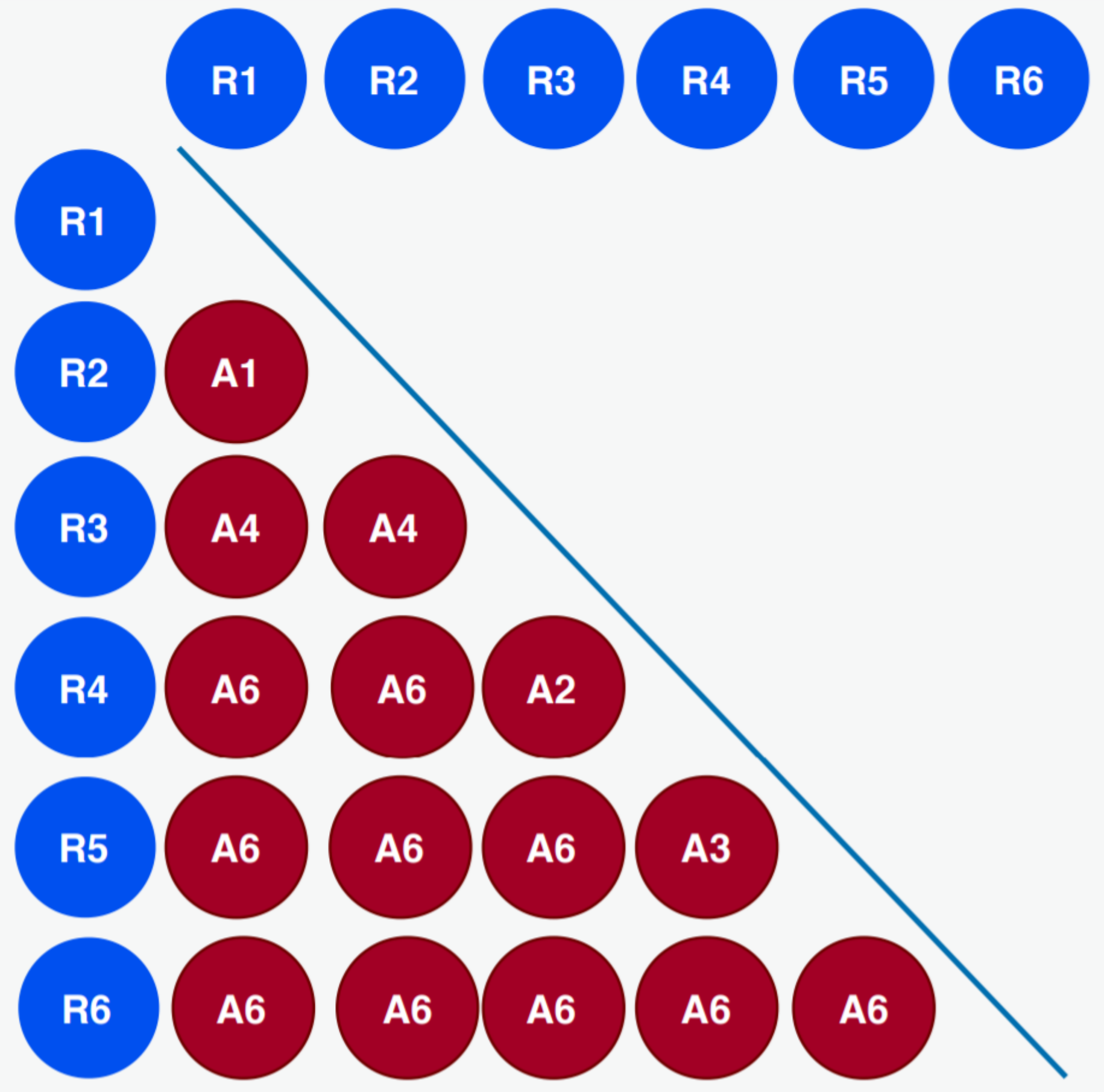
Definition:
Least Common Ancestor: The leaf concept from all common ancestors of two concepts

Step 1. Find all ancestors of R1 and R3, the result set consists of A1, A2, A4, A5, A6,
Step 2. Find all common ancestors, the result set is a4 and A6,
Step 3. Find specific ancestor (leaf) from the result of step 2, then the LCA concept is A4

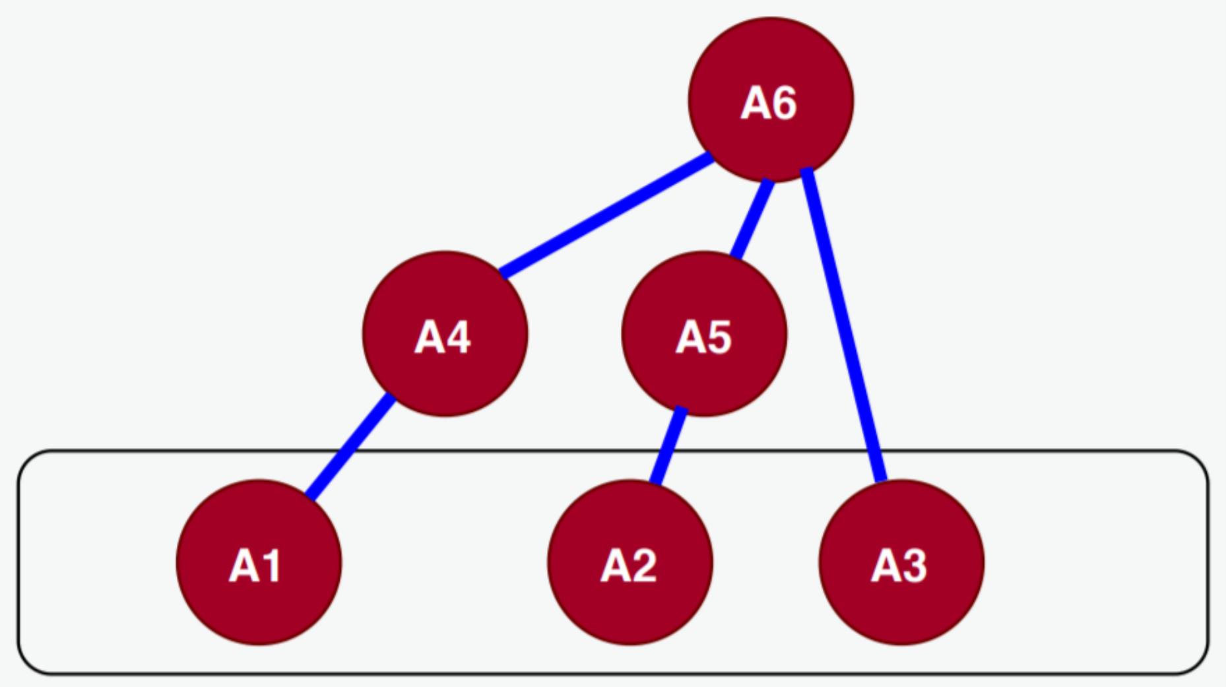
Pair-wise LCA(R1 R3) = A4

LCA – from a set of input

Next step, calculate all Pair-wised LCA for all concept pairs in the input set, and the result is shown on the left side below, then based on the existing hierarchy relationship of calculated Pair-wise LCA result, select the leaves concepts and the LCA result. In the example, the selected result is A1, A2, A3.



	R1	R2	R3	R4	R5	R6
R1						
R2	A1					
R3	A4	A4				
R4	A6	A6	A2			
R5	A6	A6	A6	A3		
R6	A6	A6	A6	A6	A6	



LCA Result

Last step. For any input concept which is not the descendants of the current LCA result, this concept itself will be added into LCA result as its category, in the example, R6 is added in.

LCA (R1 R2 R3 R4 R5 R6) = A1 A2 A3 R6

Emerging techniques for aggregating SNOMED CT encoded patient

Ming Zhang | Donna Truran | Michael Lawley

RESULTS

SNOMED International
SNOMED CT EXPO 2020
Virtual Conference October 8-9

ILCA – Iterated Least Common Ancestor

The LCA method will aggregate a set of input SNOMED CT concepts into clustered categories. The previous example shows one round of LCA calculation from a given input set. To achieve more general clustered categories, the result from this current round will become the input concept in the next round of calculation. The iteration of LCA calculation will stop when there is no new LCA result calculated. Full aggregation history will be achieved by Iterated Least Common Ancestor(ILCA) method.

The table below shows an example of running ILCA, the number is the count of the LCA concepts in each round.

Input set	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
48	34	27	20	15	11	10

SNOMED CT Encoded Data

SNOMED encoded data from a major tertiary hospital, for both acute (emergency department) and admitted patient episodes was used.

1. In the total 320K acute episodes, there are 13200+ unique diagnosis codes
2. In the total admitted patient 500K episodes , there are 19000+ unique diagnosis codes

The ILCA method is to discover the clusters from those unique diagnosis codes, and it will be used in future analysis.

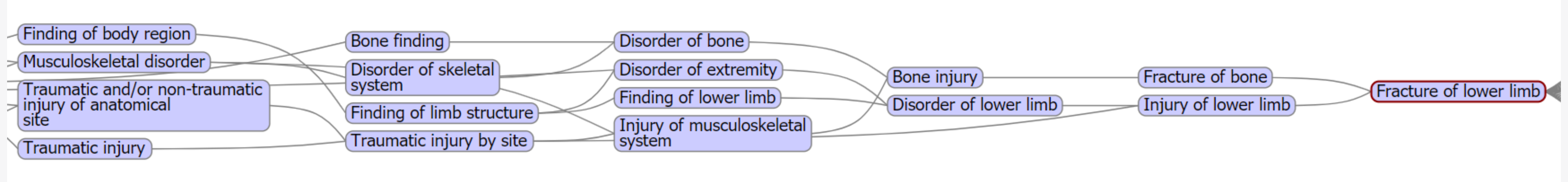
Data in SNOMED CT Hierarchy

The original diagnosis data was not constrained to concepts drawn from the Clinical Finding hierarchy. But for this experiment we are only interested in aggregating diagnoses. Clinical Finding concepts comprised 92.59% of the dataset and were used for ILCA calculation.

SCT Hierarchy	Percentage
Clinical finding	**92.59%**
Procedure	1.70%
Situation with explicit context	2.02%
Other Hierarchy	3.69%

Disorder and Finding

In SNOMED CT clinical finding hierarchy, the semantic tag for the concepts are either disorder and finding, and many disorder concepts also have multiple finding ancestor. We have separated disorder and finding concepts in ILCA analysis. That is we deploy the algorithm in two partitions, once for Findings and once for Disorders. This prevents all disorders from aggregating to their finding ancestor concepts.



Emerging techniques for aggregating SNOMED CT encoded patient

Ming Zhang | Donna Truran | Michael Lawley

SNOMED International
SNOMED CT EXPO 2020
Virtual Conference October 8-9

The clustered categories

Acute (emergency department) Data				Inpatient Data			
Disorder		Finding		Disorder		Finding	
Round	**Count**	**Round**	**Count**	**Round**	**Count**	**Round**	**Count**
Total	9859	Total	2801	Total	12979	Total	4719
Input	127	Input	203	Input	135	Input	237
Round 1	92	Round 1	140	Round 1	97	Round 1	182
Round 2	79	Round 2	93	Round 2	84	Round 2	138
Round 3	71	Round 3	63	Round 3	79	Round 3	81
Round 4	69	Round 4	46	Round 4	77	Round 4	60
Round 5	68	Round 5	35	Round 5	76	Round 5	42
		Round 6	32			Round 6	41
						Round 7	40
						Round 8	37

The clustered categories for concepts appearing in the data collection are

- Acute data 9859 disorder concepts cluster into 68 categories in 5 ILCA round.
- Acute data 2801 finding concepts cluster into 32 categories in 6 ILCA round.
- Inpatient data 12979 disorder concepts cluster into 76 categories in 5 ILCA round.
- Inpatient data 4719 finding concepts cluster into 37 categories in 8 ILCA round.

Conclusions

The ILCA Method

- is a dynamic and data-driven way to perform aggregation from SNOMED CT encoded patient data
- needs the input data to be prepared before processing start
- produce full history of aggregation and the original highly specific data is still available (untransformed) for use in decision making
- relies on SNOMED CT native hierarchy only
- Is efficient and reproducible
- Requires some initial specification, but low ongoing maintenance

Future Directions

There are challenges raised during the aggregation processing. Data collection that is highly heterogeneous will almost always contain SNOMED concepts that lie outside the scope of ancestor calculation. Also, some questions would need to be considered while using this method for clustering. Such as, how high up the SNOMED CT hierarchy is too high to provide ‘clinically sensible’ groupings? How many reporting groups are too many or too few?