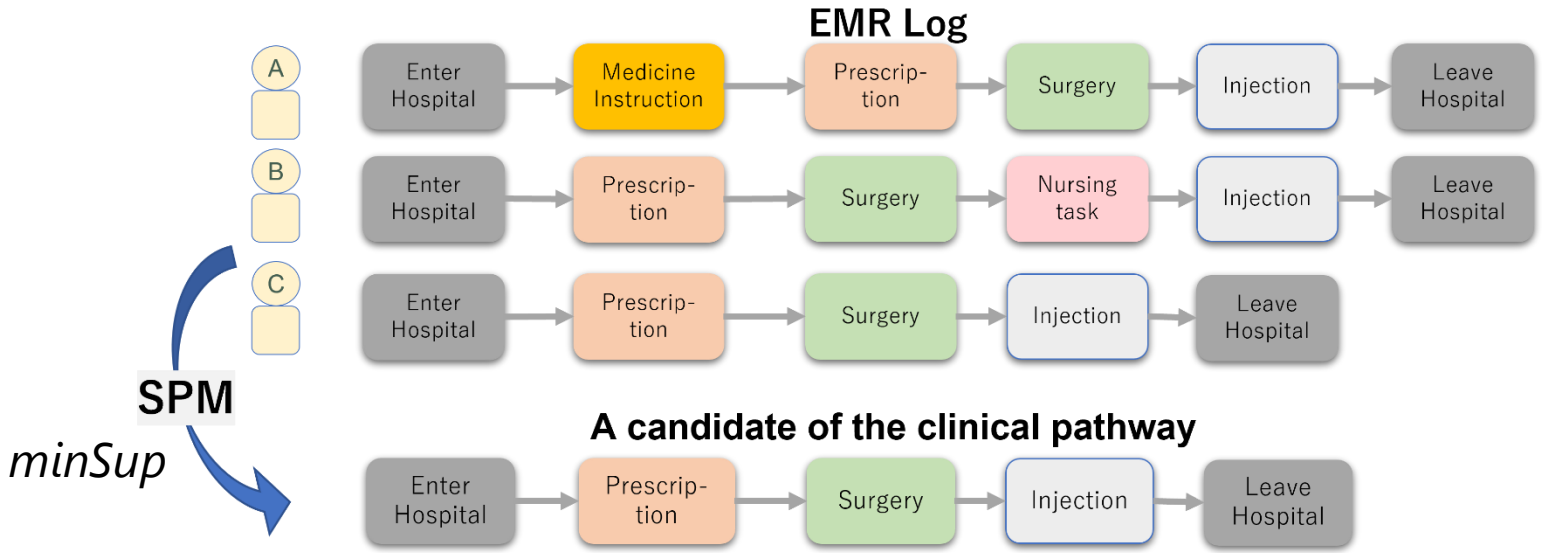
A Clustering-based Sequence Variants Analysis Method for Electronic Medical Records of Multimedical Institutions

Hieu Hanh Le[1], Yuki Yasumitsu[2], Ryosuke Matsuo[1], Tomoyoshi Yamazaki[1], Haruo Yokota[1,3]

[1] *Ochanomizu University,* [2] *Tokyo Institute of Technology,* [3] *Josai University*

2 EMR Data Analysis

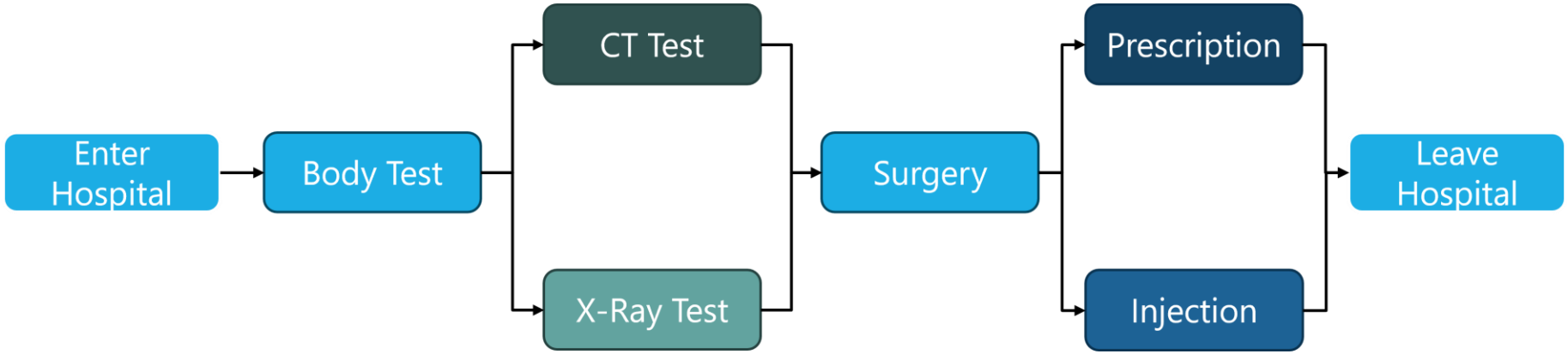
- EMR has been widely used in many hospitals, and the data has been analyzed to improve medical tasks
  - Generate clinical pathways by using SPM techniques



EMR: Electronic Medical Records System, SPM: Sequential Pattern Mining

3 One Institute Data Analysis

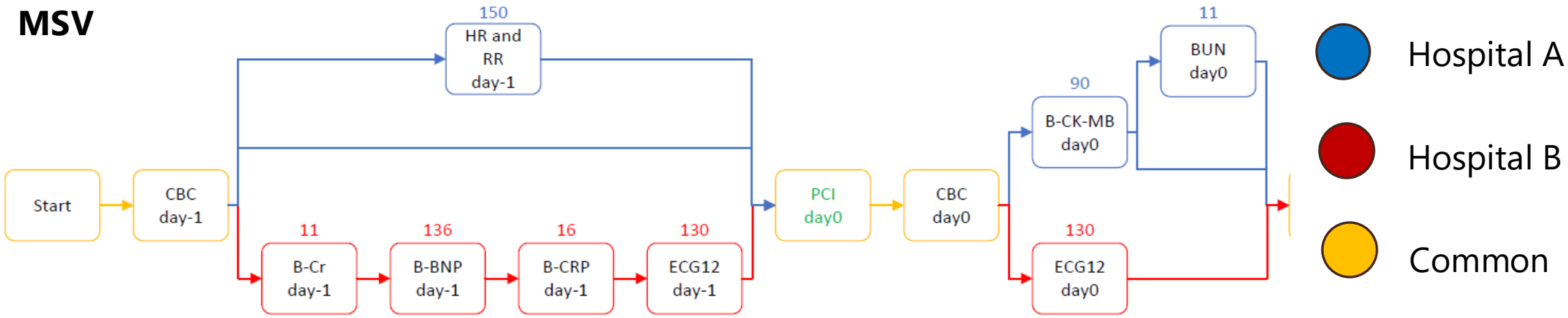
- Visualize the generated clinical pathways as Sequence Variant (SV) [Honda+, DMAH2019]
  - SV is the extension of a sequence with branches
- Understand the background reasons that led to the branches [Le+, ACM Healthcare 2023]
  - Test results, medical background, gender, age, etc.



An SV example

4 Multiple Institutes Data Analysis

- Analyzing data from multiple medical institutes is desired
  - Compare the treatment patterns with those in other hospitals [Li+, DEXA2022]
    - A hospital can confirm its characteristics and improve medical practices by referring to the treatment patterns of other



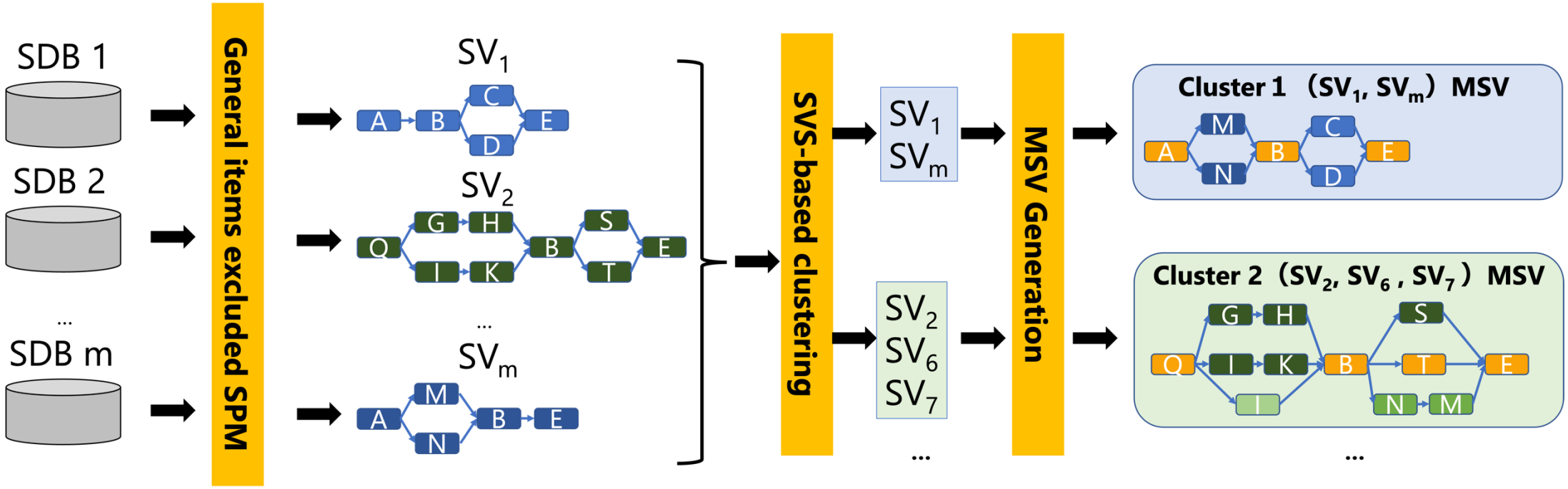
Merged Sequence Variants (MSV) of the two hospitals' frequent treatments

5 Challenges

- **Comparing the differences in the treatment patterns for more than three medical institutes**
  - As the number of medical institutions increases, the commonalities decrease, making it difficult to grasp the characteristics accurately
- **General medical items that appear in almost all sequences should be considered optional**
  - Body or blood tests that are commonly performed but do not provide helpful medical information, such as those related to the prescription of injections

6 Proposal

- A method to efficiently compare SVs from more than three medical institutions



SDB: Sequence DataBase, SVS: Sequence Variant Similarity

General Items Excluded SPM

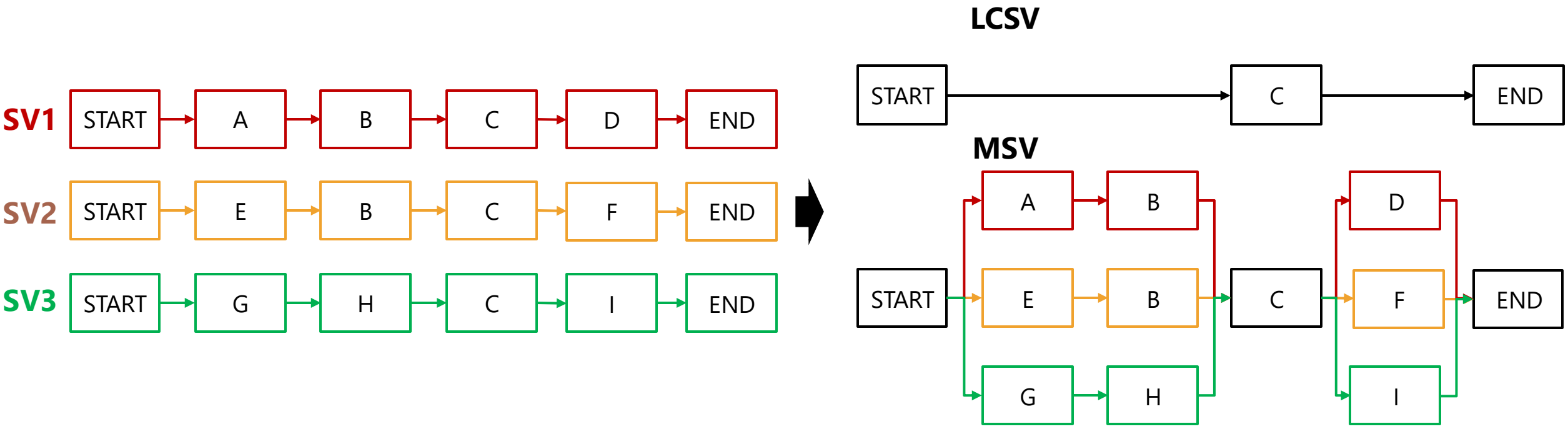
- Introduce an indicator Term Occurrence per Sequence (TO/S) to evaluate the frequency of occurrences
  - $TO/S(item, institution) = \frac{\#occurences(item, institution)}{\#sequences(insitution)}$
- Test items with high TO/S scores across institutions will be excluded
  - The threshold is set when excluding the test items doesn't affect on the SPM results

SVS-based Clustering
SVS (Sequence Variant Similarity)

- Clusters are created based on the dendrogram produced by hierarchical clustering
  - For clustering, the similarity of SVs needs to be defined
    - $SVS(SV_1, SV_2) = 1 - \frac{2 \times |LCSV(SV_1, SV_2)|}{|SV_1| + |SV_2|}$
      - $|SV|$: number of nodes in SV
  - Apply five hierarchical clustering methods
    - Single linkage, complete linkage, centroid linkage, average linkage, and Ward
- Elements to choose the appropriate clustering method
  - The number of clusters is minimum
  - The existence of the LCSV in clusters

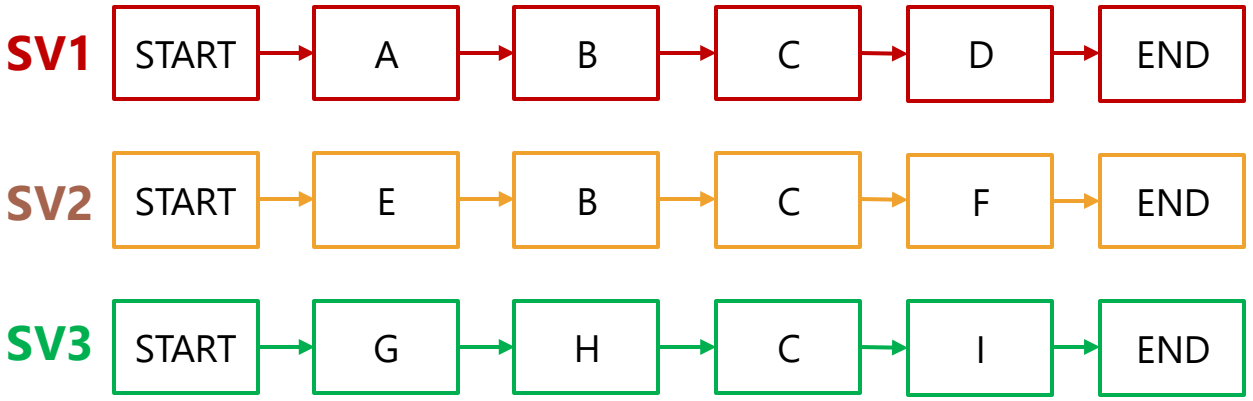
MSV Generation: Direct Merging

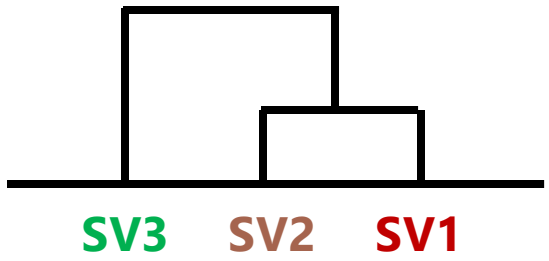
- Directly merge all the SVs in the cluster
  - Easy to understand each SV but only the commonality of all SVs appear

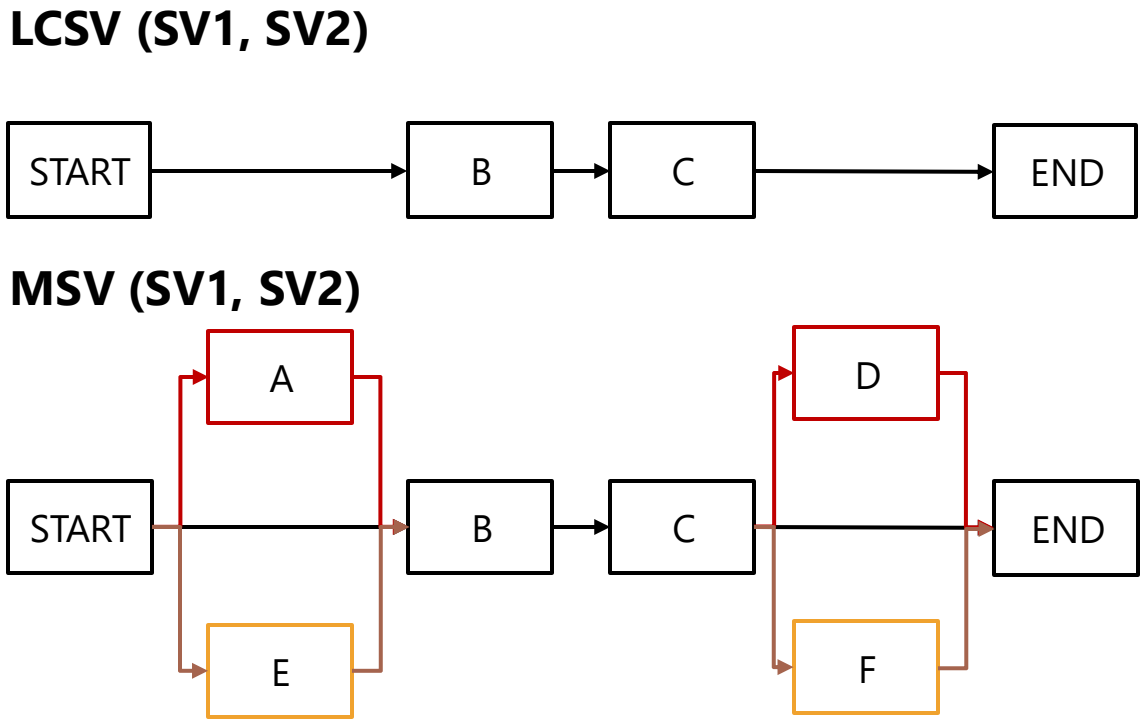


10 MSV Generation: Distance-based Merging (1/2)

- Merge SVs in order of proximity on the dendrogram obtained during clustering

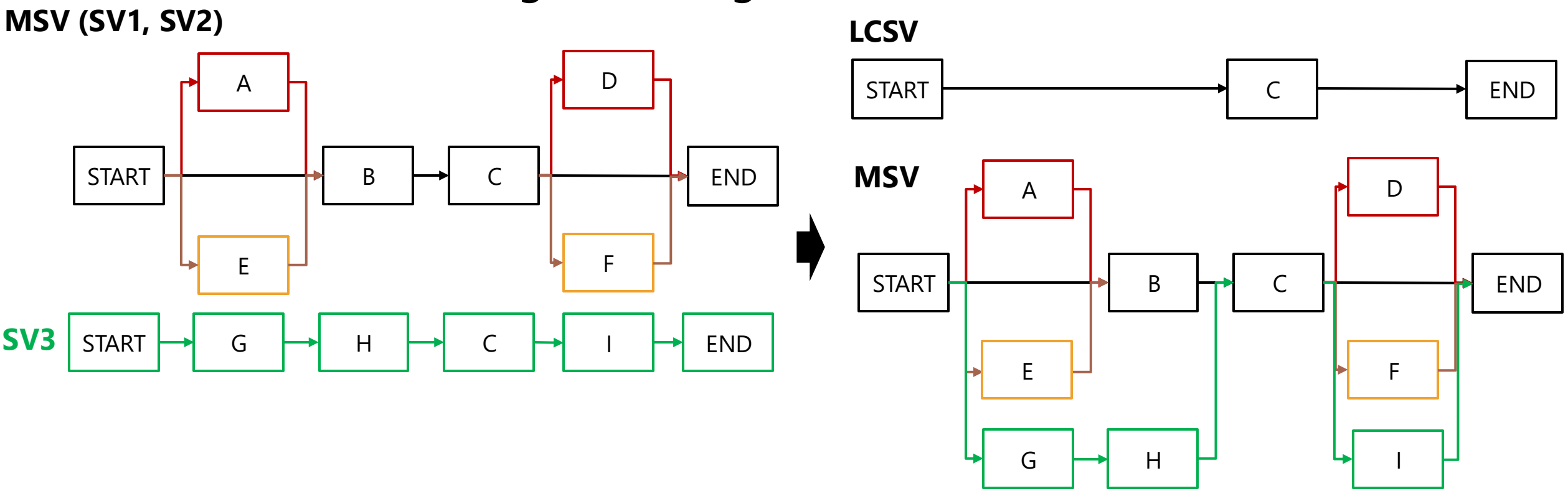






11 MSV Generation: Distance-based Merging (2/2)

- Merge SVs in order of proximity on the dendrogram obtained during clustering



12

Experiment

- Dataset
  - Actual EMR pertaining to the fifth wave of COVID-19 in Japan
    - July 1, 2021 to September 30, 2021
    - 23 medical institutions (Medical Institution A, B, …, W)

TABLE I
STATISTIC INFORMATION OF SEQUENCES IN THE DATASET (PART 1/2)

Institution	A	B	C	D	E	F	G	H	I	J	K
#sequences	130	23	11	102	89	79	126	36	29	104	104
#ave_length (raw)	181.7	255.6	45.3	153.7	242.6	284.3	90.2	127.4	211.4	317.5	112.2
#ave_length (after exclusion)	18.7	56.6	5.3	22.0	47.2	31.8	23.2	15.6	37.2	51.3	35.3

TABLE II
STATISTIC INFORMATION OF SEQUENCES IN THE DATASET (PART 2/2)

Institution	L	M	N	O	P	Q	R	S	T	U	V	W
#sequences	30	179	228	182	29	57	283	55	69	22	28	22
#ave_length (raw)	275.7	282.4	272.5	181.0	73.7	310.0	139.2	174.5	189.4	340.0	220.3	137.1
#ave_length (after exclusion)	57.8	40.6	46.2	23.0	16.4	66.0	13.3	21.3	36.2	51.1	38.6	18.4

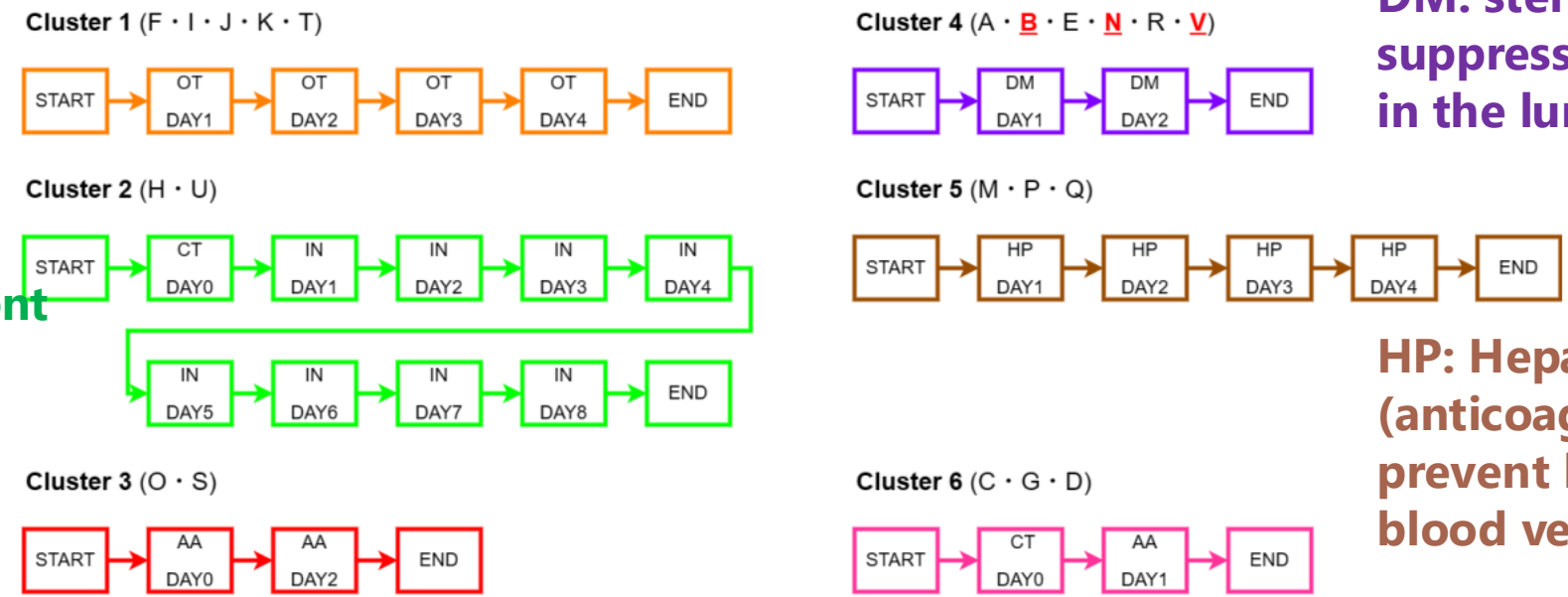
13

Experimental Results
LCSV Generation

- Generate LCSV at each cluster based on the clustering results from the Ward method
- Several essential treatment orders appeared

OT: oxygen administration

IN: medications for diabetes treatment

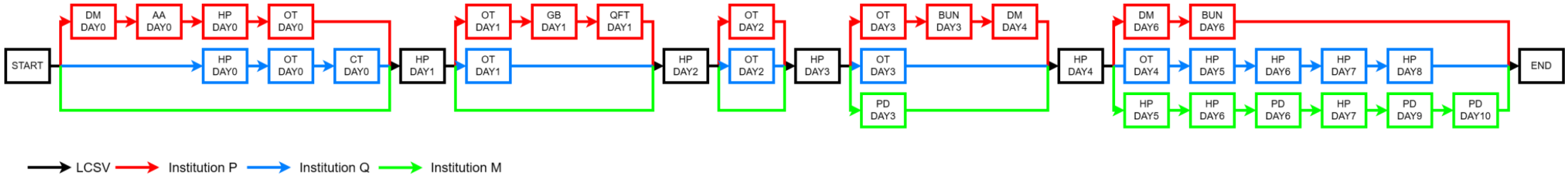


DM: steroid medication suppressing inflammation in the lungs

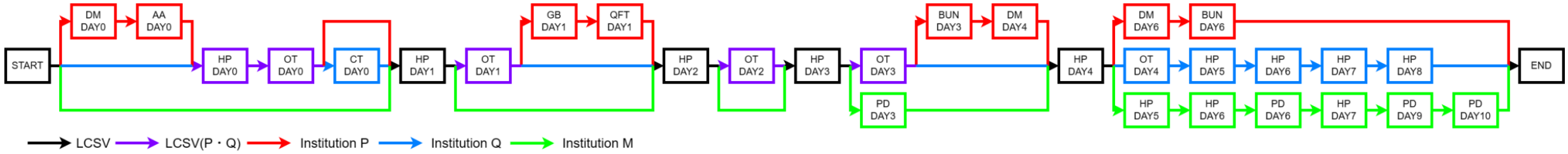
HP: Heparin injection (anticoagulant) to help prevent harmful clots in blood vessels

14 Experimental Results MSV Generation (Cluster 5)

- Direct merging method: Individual SVs are easier to discern, but it results in a larger number of nodes



- Distance-based method: Effectively represents **common nodes** for a subset of compared institutions' SVs



Conclusion

- Proposed a method to understand the commonalities and differences in frequent medical order patterns among three or more medical institutions
  - Exclude general but not essential items during mining
  - Perform hierarchical clustering using a defined similarity metric between SVs
  - Generate MSV by combining the LCSVs and SVs within clusters using direct merging and distance-based methods
- The proposed method was shown to be effective using a real dataset from 23 medical institutions

THANK YOU!

A Clustering-based Sequence Variants Analysis Method for Electronic Medical Records of Multimedical Institutions

APPENDIX

18

Experimental Results
Abbreviation of Medical Orders

Abbreviation	Medical order name	Description
CT	Computed tomography scan	A test to capture the inside of the body
ECG	Electrocardiogram	A test to check the condition of the heart
UE	Urinalysis	A test to analyze the components of urine
TP	Syphilis test	A test to check for syphilis infection
BNP	Brain natriuretic peptide	A test to assess cardiac stress
BUN	Blood urea nitrogen	A test to evaluate kidney function
GB	Antivirus antibody in each globulin class	A test to examine antibodies for various viruses
QFT	Interferon-gamma release assays	A test to check for tuberculosis infection
AST	Aspartate aminotransferase	A test to examine abnormalities in the liver or heart
HbA1c	Hemoglobin A1c	A test to evaluate blood sugar levels
OT	Oxygen administration	Oxygen administration
DM	Dexamethasone	Steroid medication
PD	Prednisolone	Steroid medication
HP	Heparin	Anticoagulant
AA	Acetaminophen	Antipyretic analgesic
IN	Insulin	Medication for diabetes treatment
SEN	Sennoside	Laxative treatment

19

Experimental Results General Items Exclusion Effect

- Excluded common test items whose *Max(TO/S)* results are greater than the Hepatitis screening (2.1)

Medical order	Max(TO/S)
SpO2	11.0
Blood glucose measurement	10.2
Respiratory monitoring	9.5
X-ray test	9.5
Cold evaporator	9.3
Isotonic sodium chloride solution syringe	8.3
Peripheral blood general test	7.5
Glucose	6.8
Total bilirubin	6.5
C-reactive protein	6.4
Intravenous drip	6.2
Ferritin	5.6
Blood gas analysis	5.2
D-dimer	5.1
Prothrombin time (PT)	4.0
Sodium and chlorine	3.7
Direct measurement of arterial pressure	3.6
Arterial blood sampling	3.3
Activated partial thromboplastin time (APTT)	3.3
Lactic acid	3.2
Total protein	3.2
Central venous injection	3.0
Central venous pressure monitoring	3.0
Narcotic analgesic	2.9
Bacterial culture and identification test	2.7
End tidal Co2 monitoring	2.6
Inspiratory distribution	2.4
Bacterial microscope test	2.4
Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)	2.3
von Willebrand factor antigen	2.2
Hepatitis screening	2.1

Medical order	Max(TO/S)
Urinalysis	1.9
Syphilis test	1.6
Electrocardiogram	1.6
Computed tomography scan	1.5
Procalcitonin	1.5
Blood urea nitrogen	1.4
Creatinine	1.4
Aspartate aminotransferase	1.4
Nasopharyngeal swab collection	1.3
Krebs von den Lungen-6	1.0
Brain natriuretic peptide	0.9
Interleukin	0.9
ABO blood type	0.9
Hemoglobin A1c (HbA1c)	0.9
Free thyroxine (FT4)	0.8
Thyroid-stimulating hormone (TSH)	0.8
Thymus and activation-regulated chemokine (TARC)	0.7
LDL cholesterol	0.6
Creatine kinase MB (CK-MB)	0.6
Anti-virus antibody in each globurlin class	0.6
Interferon-gamma release assays	0.5
Pulmonary Surfactant Protein	0.5
Indwelling catheter	0.5
Soluble interleukin-2 receptor	0.5
Fibrinogen	0.2

Order excluded by Max(TO/S)

Order excluded by low relationship with Covid-19 treatment