Health Consumer Usage Patterns in Management of CVD using Data Mining Techniques

Devipriyaa Nagappan, Jim Warren, Pat Riddle

University of Auckland

Motivation

- Examine the potential to use a string pattern matching approach to health consumer trajectories as a basis for analysis of chronic conditions
  - Rather than basing analysis on numeric features, view a healthcare history as a sequence of events of different types (i.e. a string of tokens)
  - Different sorts of healthcare journeys can be clustered and we can look for association to different outcomes

- We’re particularly interested to apply this to cardiovascular disease (CVD)
  - We have really good data about CVD risk management through the Vascular Intelligence using Epidemiology and the Web (VIEW) programme

Background

- Some major decisions for this ‘syntactic’ approach
  - What are our tokens? (defining the events of interest)
  - What’s our string similarity measure (and how do we cluster)?

- Particularly inspired by
  - Yiye Zhang, Rema Padman and Larry Wasserman, “On Learning and Visualizing Practice-based Clinical Pathways for Chronic Kidney Disease” AMIA Annu Symp Proc. 2014, 1980–1989.

Zhang et al approach to state formation

- State ‘token’ is a combination of visit type (e.g. new patient or follow-up), diagnosis (limited to CKD stage and a few comorbidities) and procedure (of 27)
- String for a patient is a series of distinct tokens ordered by visit date

Longest common subsequence (LCS) distance

- *LCS* between two strings *x*, *y* is the length of longest subsequence present in both of them
  - A subsequence is a sequence that appears in the same relative order, but not necessarily contiguous
  - **Examples:**
    LCS for input Sequences “ABCDGH” and “AEDFHR” is “ADH” of length 3.
    LCS for input Sequences “AGGTAB” and “GXTXAYB” is “GTAB” of length 4.
- $dLCS(x, y) = |x| + |y| - 2LCS(x, y)$

- Track record in biomedicine including protein sequence analysis

Our approach for this paper

- Synthetic data set
  - VIEW has a lot of ‘real world’ details
  - Wanted to establish a baseline with structurally similar data (and where we know ‘the answer’ to some degree)
- Explore different clustering methods
  - Effectiveness, efficiency
- Examine the clusters
  - Do they describe?
  - Do they predict?

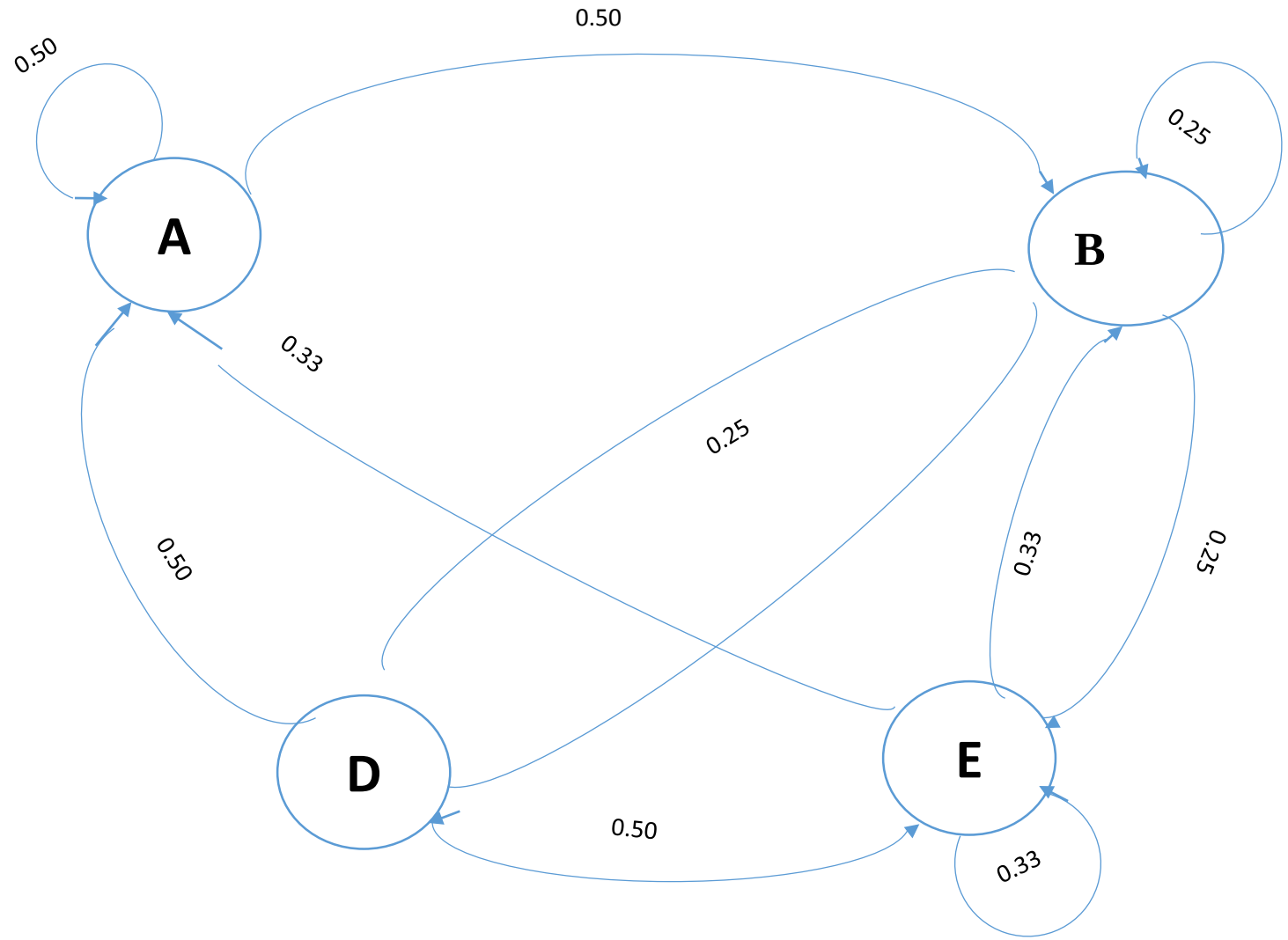
Simulated CVD hospitalisation & recovery

- Created a population with a distribution of risk factors
  - E.g. diabetes, higher risk ethnicities (M&P versus European), smoking status etc.
  - Assigned ‘risk score’ for each case in line with Framingham risk
  - Stratified each case to low, moderate or high risk based on score
- Generated 10,000 individuals
- Simulated 36 months of state transition

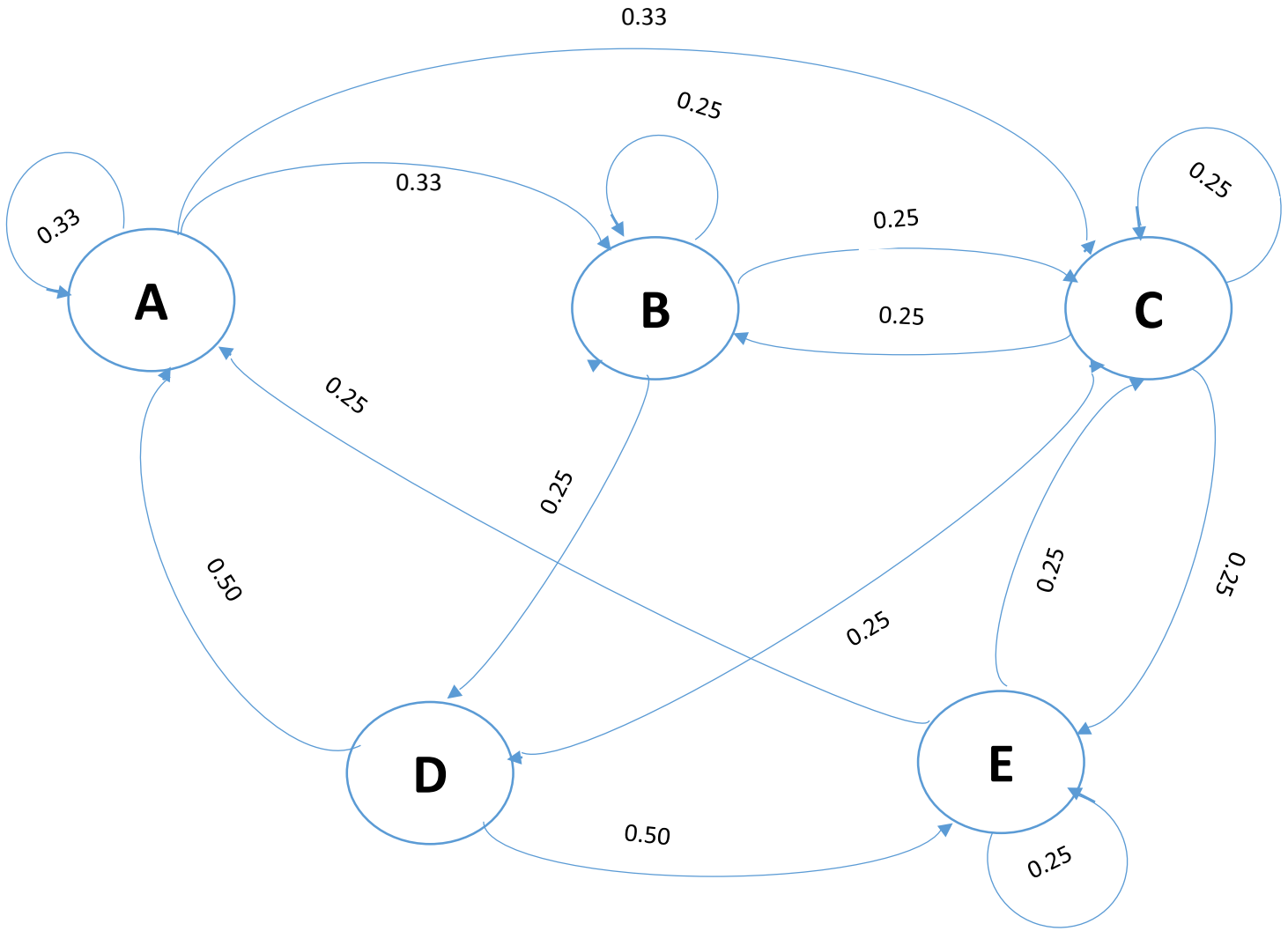
States for simulated data

State of Events	Denoting characters
1. Not-Admitted	A
2. Admitted	B
3. Intensive care unit (ICU)	C
4. Discharged	D
5. Discharged with home care	E
6. Mortality	F

State transition for moderate-risk group



State transition for high-risk group



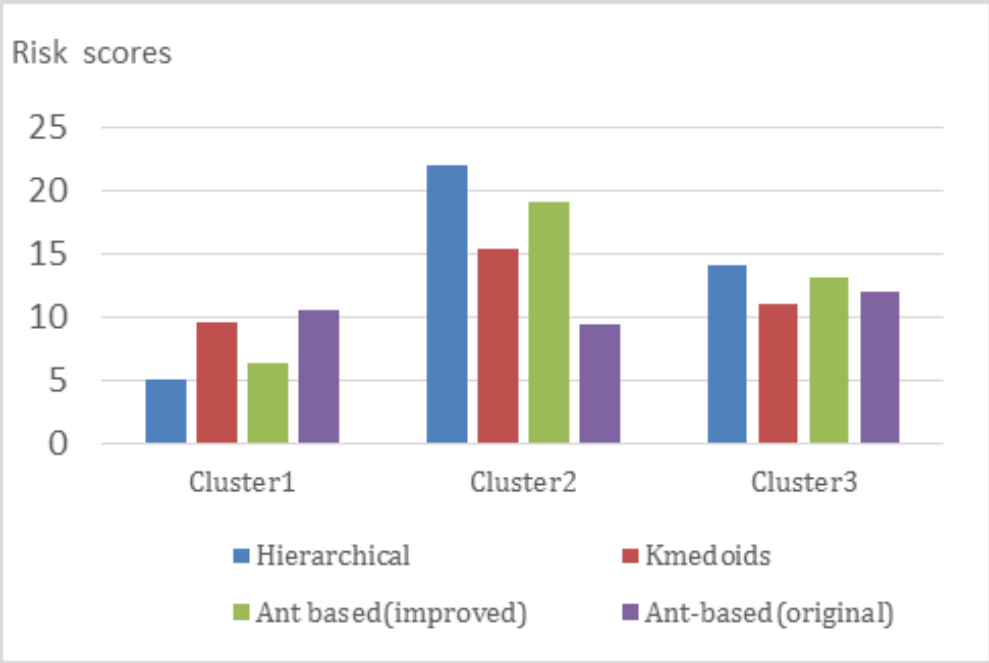
Clustering

- For reference used Hierarchical clustering (deterministic) and k-medoids (non-deterministic)
- Hierarchical requires O(n2) dLCS comparisons
  - Might be a problem for big populations with long sequences
- k-medoids is like k-means (picking random cases to build the k clusters around) but suitable for dLCS
  - Minimizes a sum of pairwise dissimilarities instead of a sum of squared Euclidean distances
- Also wanted to try alternative… ant-based clustering (ABC)

Ant-based clustering

- Metaphorical ‘ant’ agents pick up and drop items on an abstract 2x2 (actually wrapped at edges) grid
- Pick up and drop items with probability based on similarity of case to neighbourhood
  - Likely to pick up a case that has high dis-similarity scores with its neighbours
  - Likely to drop a case that has low dis-similarity score with its neighbours
- Ants wander randomly or heuristically (e.g. following trails, or moving toward cluster centres)
- Resolve by merging nearby cases into clusters

Results - clusters



Cluster 1						
Algorithms	A	B	C	D	E	F
K-medoids	22.6 ± 9.02	4.8 ± 2.66	2.01 ±1.8	3.38 ± 2.4	3.87 ± 4.3	1.15 ± 1.28
Hierarchical	33.7 ± 0.6	0.54 ± 0.05	0.04 ± 0.008	0.54 ± 0.03	0.6 ± 0.24	0
Ant-based(improved)	28.69 ± 2.99	2.7 ± 0.95	0.78 ± 0.49	1.15 ± 0.38	2.26 ± 0.84	0.34 ± 0.37
Ant-based(Original)	19.6 ± 8.02	4.05 ± 3.66	2.05 ± 4.6	3.18 ± 4.38	3.87 ± 5.38	3.15 ± 1.42
Cluster 2						
Algorithms	A	B	C	D	E	F
K-medoids	7.68 ± 5.6	6.6 ± 4.6	7.57 ± 5.6	3.35 ± 2.16	7.05 ± 2.07	1.15 ± 1.5
Hierarchical	5.83 ± 0.6	5.53 ± 0.41	7.4 ± 0.035	4.1 ± 0.07	6.9 ± 0.02	4.09 ± 0.13
Ant-based(improved)	5.68 ± 0.63	8.7 ± 1.05	7.07 ± 1.07	3.68 ± 0.17	6.57 ± 0.48	4.03 ± 1.25
Ant-based(Original)	15.6 ± 5.8	4.5 ± 3.66	4.9 ± 3.27	4.45 ± 4.13	7.9 ± 2.19	1.32 ± 1.3
Cluster 3						
Algorithms	A	B	C	D	E	F
K-medoids	10.68 ± 5.6	8.9 ± 2.05	3.83 ± 2.39	2.73 ± 1.7	5.21 ± 2.59	1.7 ± 2.45
Hierarchical	7.7 ± 0.12	11.3 ± 0.04	2.63 ± 0.24	3.58 ± 0.2	7.4 ± 0.3	0.53 ± 0.003
Ant-based(improved)	11.6 ± 5.9	10.27 ± 2.64	4.83 ±1.15	2.59 ± 0.8	7.16 ± 1.4	1.11 ± 0.94
Ant-based(Original)	20.6 ± 8.05	4.86 ± 3.79	3.49 ± 2.06	5.73 ± 3.7	3.21 ± 2.1	1.03 ± 2.05

Results - performance

- Silhouette index, Dunn Index, DB Index
  - 3 clusters best
  - Hierarchical best, our variant of ABC second best
- Prediction
  - Attempted to predict final 6 tokens from first 30
    - Using closest cases in cluster, and using HMM and RNN
    - 40-60% accuracy, not significantly different for each method
- Run-time
  - k-medoids: 600s, ABC: 7400s, Hierarchical: 18000s

Discussion

- State-token based representation of patient history is a promising direction in analysis of chronic condition management
  - An intuitive way to think about a patient journey
  - Wide range of choices to explore in state definition and distance measures
- Ant-based clustering (with appropriate heuristics) may be a promising middle ground between deterministic (hierarchical) and randomly seeded (k-means/medoids) approaches
  - Clusters can describe population groups, provide insights on patient journeys and (using case-base distance similarity) have potential in prediction

Questions

Thank you!

jim@cs.auckland.ac.nz

http://www.cs.auckland.ac.nz/~jim/