Clustering Discrete State Trajectories of Varying Lengths: Health Care Utilization Patterns

Laura Hatfield, PhD

Associate Professor, Harvard Medical School

ISPOR May 22, 2018

Baltimore, MD

Thank you to my collaborators

Brianna Heggeseth

- Williams College -> Macalester College

Megan Schuler

- RAND Boston

Nina Joyce

- Brown University

Haiden Huskamp

- Harvard Medical School

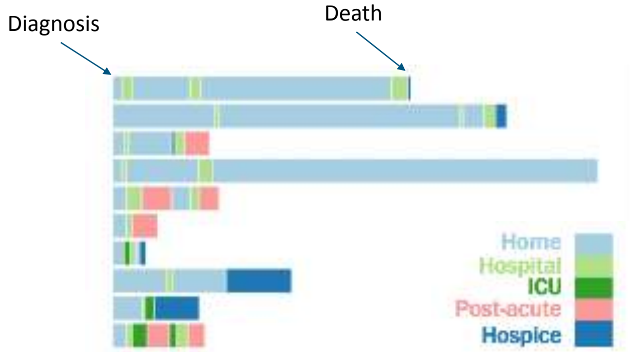
Elizabeth Lamont

- Harvard Medical School

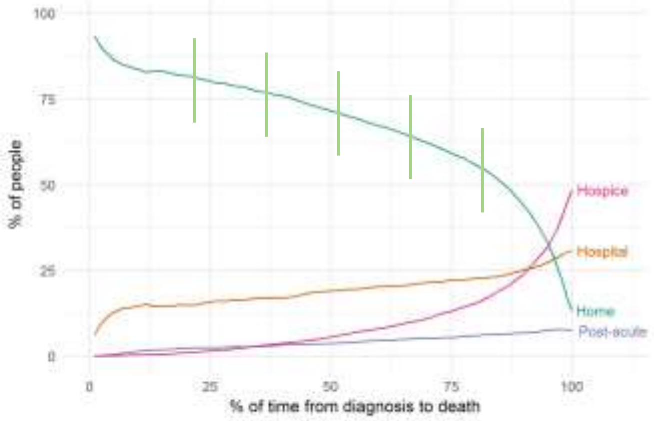
Health care utilization trajectories

Medicare claims
14,257 people
Advanced lung cancer

Individual trajectories: sequences of days in each care setting

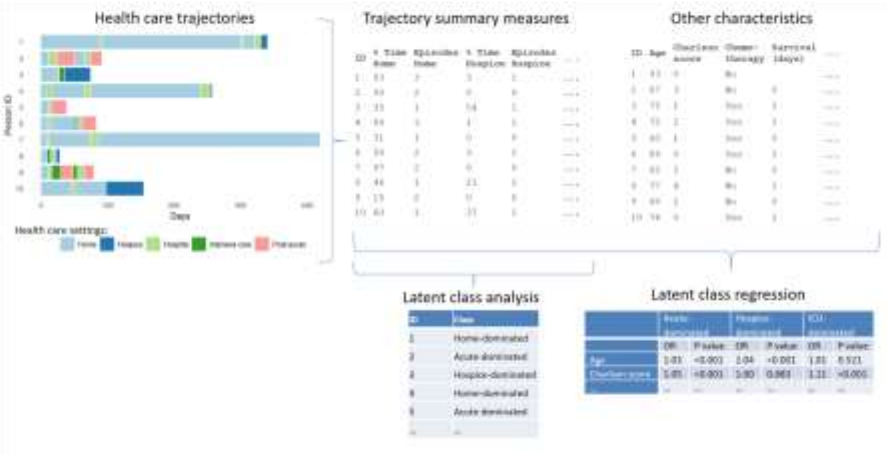


Can we use clustering to **discover** and **illustrate** variation in experiences?



Source: Schuler et al (2017) Health Affairs. doi: 10.1377/hlthaff.2017.0448

Feature extraction + LCA



ID	Class
1	Home-dominated
2	Acute-dominated
3	Hospice-dominated
4	Home-dominated
5	Acute-dominated
…	…

	Acute-dominated		Hospice-dominated		ICU-dominated	
	OR	P value	OR	P value	OR	P value
Age	1.01	<0.001	1.04	<0.001	1.01	0.521
Charlson score	1.05	<0.001	1.00	0.983	1.11	<0.001
…	…	…	…	…	…	…

Latent class analysis

For response pattern $\mathbf{y}$ and class c_k

$$\Pr(\mathbf{Y} = \boldsymbol{y}) = \sum\nolimits_{k=1}^{K} \Pr(C = c_k) \prod\nolimits_{j=1}^{J} \Pr\left(Y_j = y_j \,|C = c_k\right)$$

The class indicators are missing data.

Four distinct classes

66% of people | Mainly at home | 10 months survival

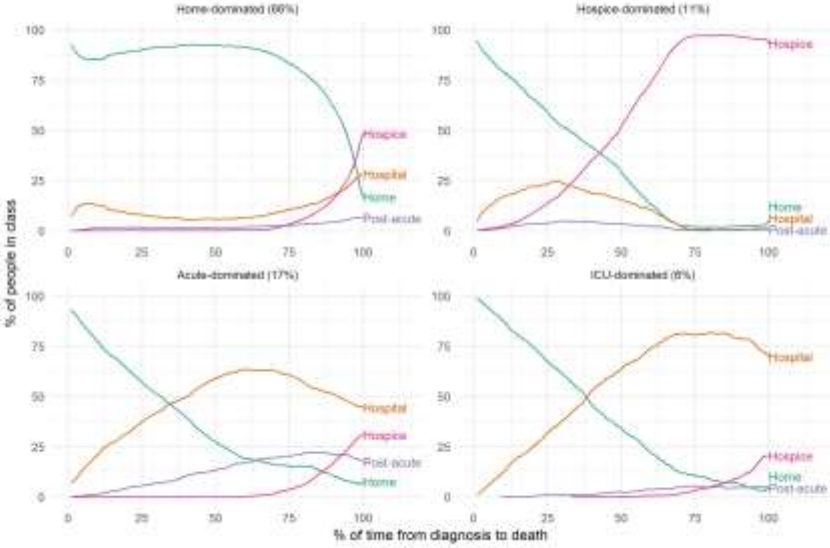
17% of people | Mainly in hospital and post-acute | 1 month survival

11% of people | Mainly in hospice | 4 months survival

6% of people | Mainly in ICU | 1 month survival

Source: Schuler et al (2017) Health Affairs. doi: 10.1377/hlthaff.2017.0448

Classes have distinct trajectories



Source: Schuler et al (2017) Health Affairs. doi: 10.1377/hlthaff.2017.0448

Remaining methods gaps

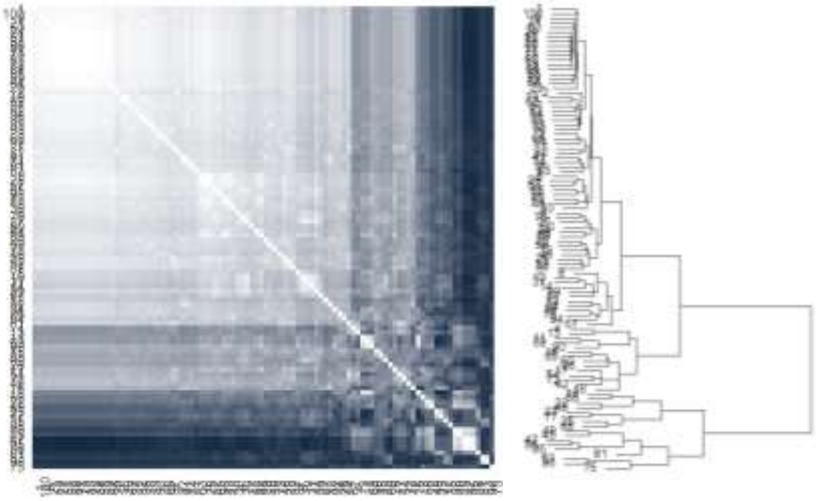
Limitations of feature extraction + LCA	**A new distance measure**
Discards ordering information	Uses sequence information directly
Requires good feature selection	Does not require feature selection by investigator
Sensitive to choice of features	Facilitates standard clustering methods

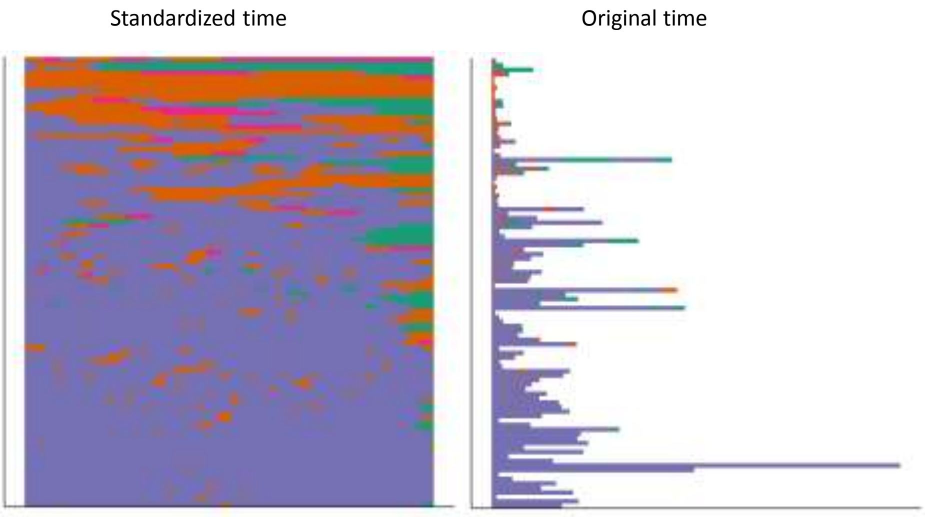
Distance is a weighted combination of

1. moving average of discordant days and
2. length difference

$$d(\boldsymbol{a},\boldsymbol{b}) = w\frac{1}{K}\sum_{k=1}^{K}\frac{\sum\left|\boldsymbol{s}(a_t|\, t \in (k, k+\tau))\ - \boldsymbol{s}\left(b_t|\, t \in (k, k+\tau)\right)\right|}{2\tau} + (1-w)\frac{|l(\boldsymbol{a}) - l(\boldsymbol{b})|}{\max\{l(\boldsymbol{a}), l(\boldsymbol{b})\}}$$

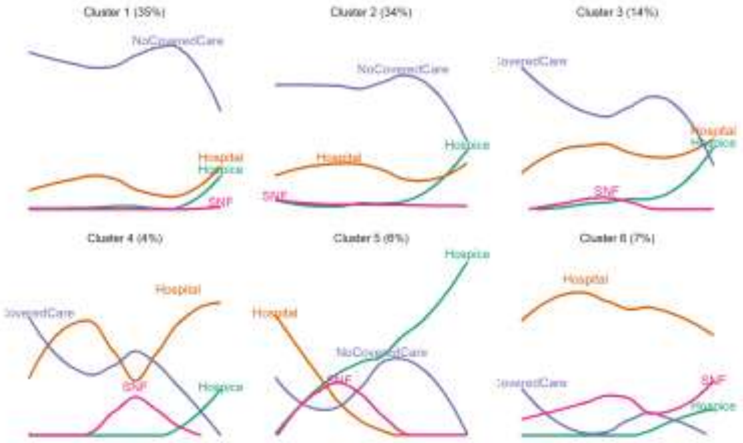
$\boldsymbol{s}(a_t\|t \in (k, k+\tau))$	Vector number of “days” in each state during time window of width τ
τ, w	Bandwidth and weight tuning parameters





Standardized time

Original time



Conclusions

- Clustering can show variation in longitudinal data
- Feature extraction enables use of LCA clustering
- Custom distance measure enables other clustering methods

hatfield@hcp.med.harvard.edu

Thanks!

@laura_tastic

@HPDSLab

healthpolicydatascience.org