Understanding cost variation with clustering

Henry Wilde

Supervised by: Dr Jonathan Gillard and Dr Vincent Knight
In partnership with: Kendal Smith (Head of Financial Flows)

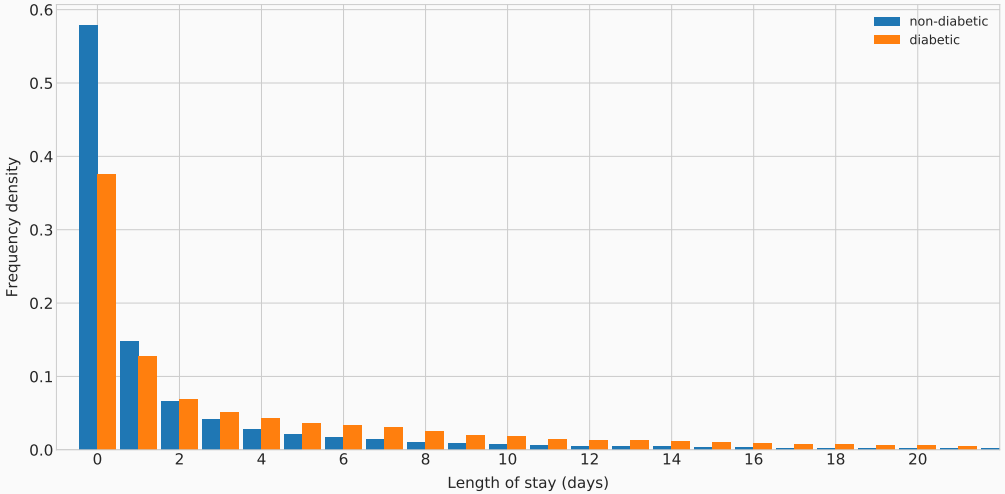
The data

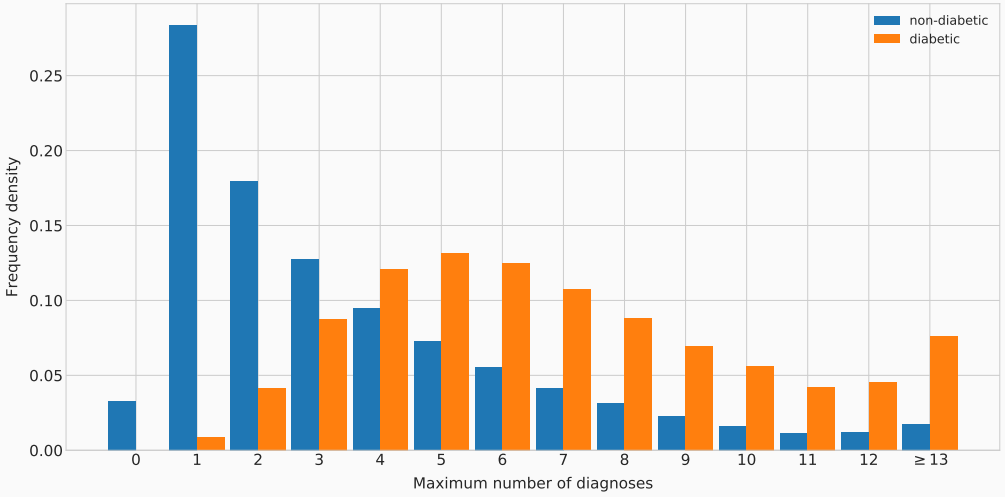
- The Cwm Taf University Health Board
- April 2012 through April 2017
- 2.4 million patient-episode records with 260 attributes

PATIENT ID	SPELL ID	EPISODE ID	Net Cost	Age (years)	HRG	Admission Date	Discharge Date	Length of Stay (days)
ID_123456	M1001	M1001-1	858.14	74	EA05Z	2015-05-06	2015-05-06	0.0
	M1211	M1211-1	333.95	74	FZ38F	2015-07-15	2015-08-01	17.0
		M1211-2	706.09	74	FZ38F	2015-07-15	2015-08-01	17.0
		M1211-3	8671.31	74	RC16Z	2015-07-15	2015-08-01	17.0

Diabetic patient analysis

An overview





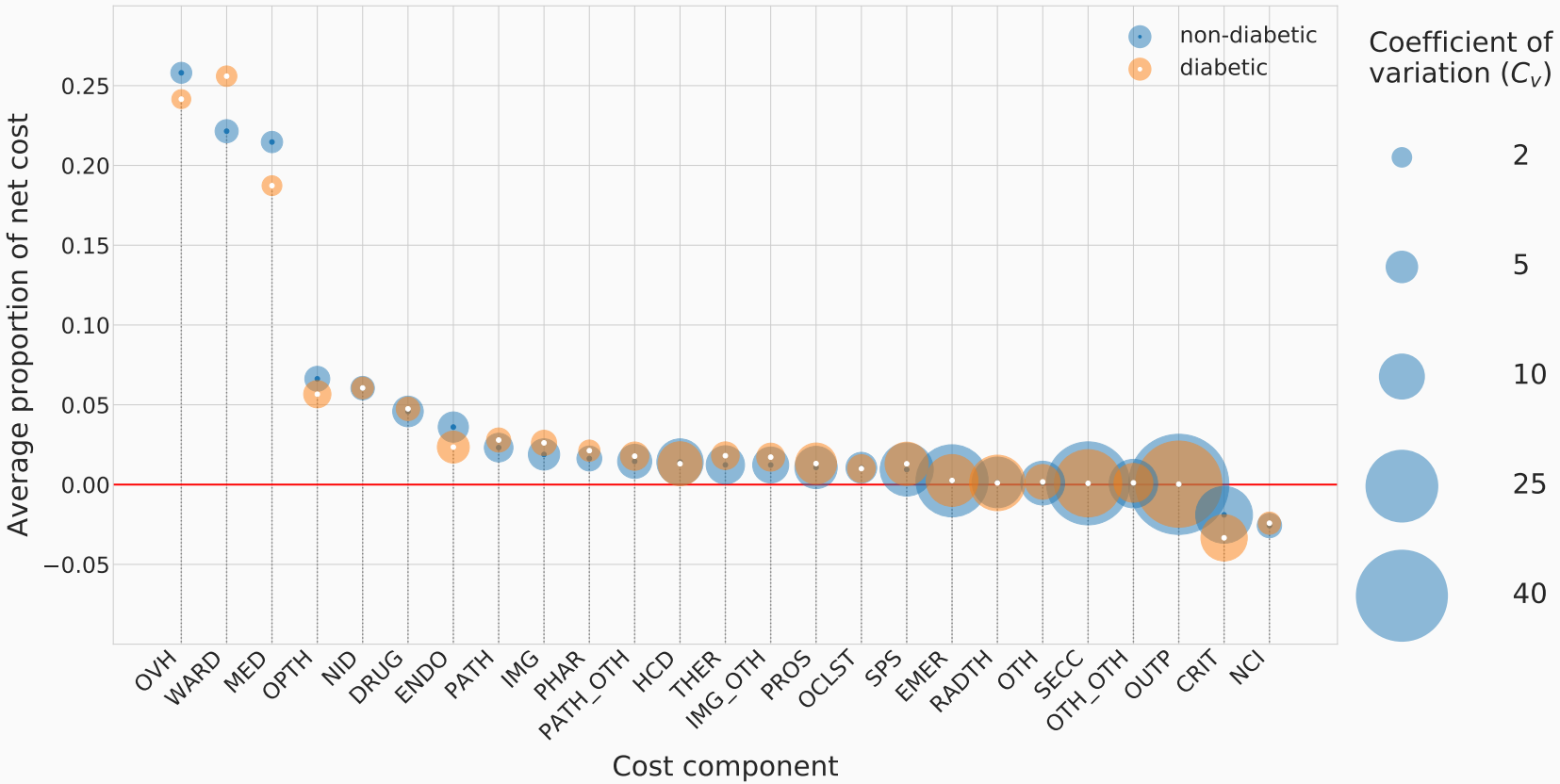
Length of stay

	Non-diabetic	Diabetic
mean	2.57	6.07
std	8.13	12.55
min	0.00	0.00
1%	0.00	0.00
25%	0.00	0.00
50%	0.00	1.00
75%	2.00	7.00
99%	35.00	57.00
max	690.00	705.00

Max. number of diagnoses

	Non-diabetic	Diabetic
mean	2.57	6.07
std	8.13	12.55
min	0.00	0.00
1%	0.00	0.00
25%	0.00	0.00
50%	0.00	1.00
75%	2.00	7.00
99%	35.00	57.00
max	690.00	705.00

Variation and importance



Partitioning the data

Subsets in the data

Traditional methods include[1]:

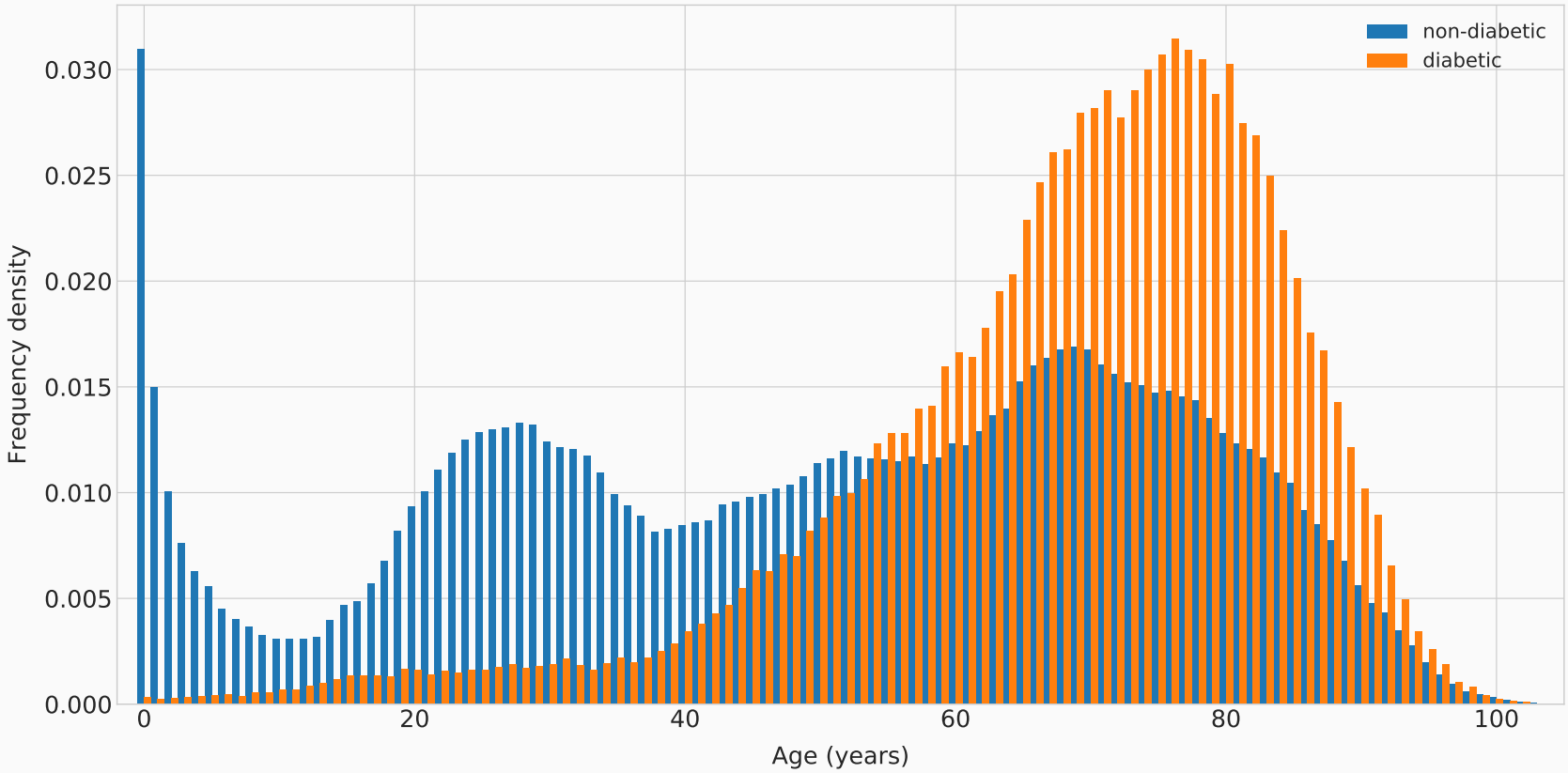
- condition-specific populations
- segmenting by age

However, these methods have the common flaw of under-representing groups of healthcare users[2].

[1] E. Nolte and M. McKee. "Measuring the health of nations: analysis of mortality amenable to health care". In: *BMJ* 327.7424 (2003), p. 1129. DOI: 10.1136/bmj.327.7424.1129.

[2] S. Vuik, E. Mayer, and A. Darzi. "A quantitative evidence base for population health: Applying utilization-based cluster analysis to segment a patient population". In: *Population Health Metrics* 14 (Dec. 2016). DOI: 10.1186/s12963-016-0115-z.

Misrepresentation - age



Misrepresentation - net cost of spell

	Non-diabetic	Diabetic	Diabetic under 30	Diabetic 30 to 65	Diabetic 65 and over
mean	1,647.00	2,648.98	1,335.11	2,110.79	2,940.21
std	3,019.53	4,152.20	2,034.58	3,565.45	4,415.44
min	4.50	10.91	52.48	35.93	10.91
1%	62.55	139.65	125.54	129.63	143.83
25%	338.67	490.64	431.76	404.30	546.11
50%	709.32	1,227.95	840.28	990.53	1,395.78
75%	1,756.90	3,106.44	1,605.21	2,338.76	3,584.18
95%	6,179.79	9,591.06	3,698.95	7,551.34	10,457.59
99%	13,414.48	19,128.45	7,697.49	16,277.28	20,310.51
max	369,168.93	273,450.30	66,963.80	106,860.69	273,450.30

What is clustering?

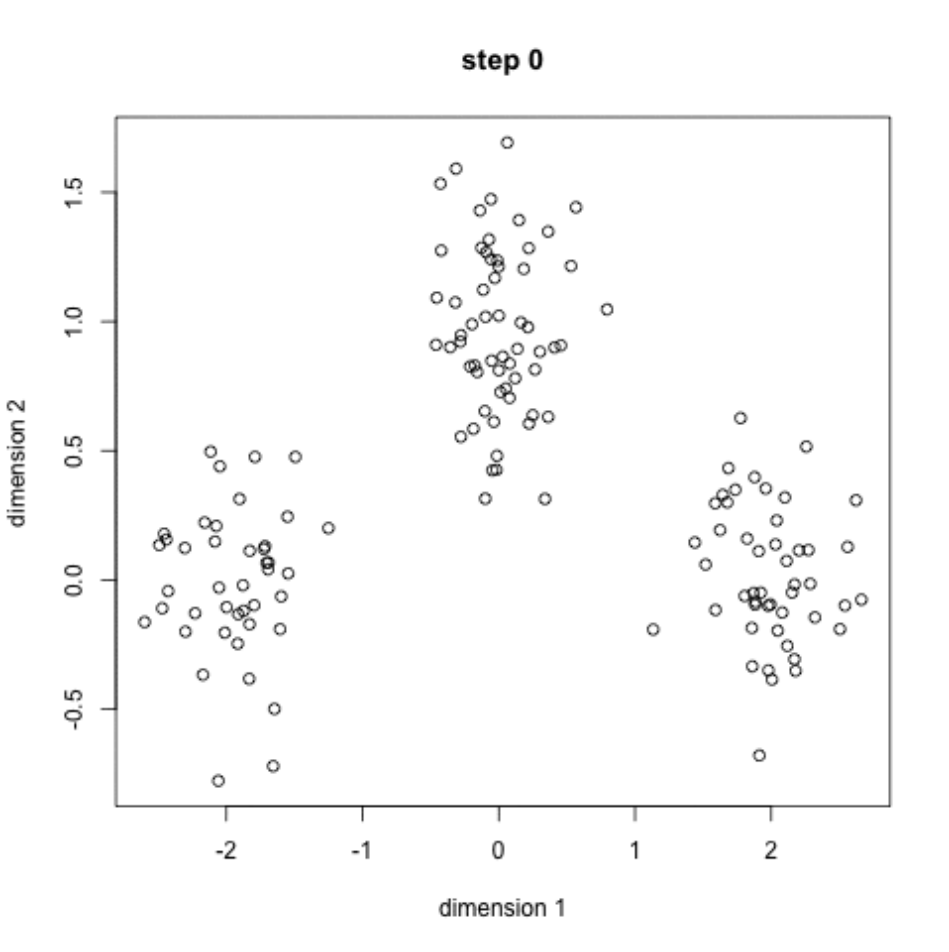


Image taken from https://www.jeremyjordan.me/grouping-data-points-with-k-means-clustering/

Clustering with healthcare data

- Patient pathways

  A. Rebuge and D.R. Ferreira. "Business process analysis in healthcare environments: A methodology based on process mining". In: *Information Systems* 37.2 (2012). Management and Engineering of Process-Aware Information Systems, pp. 99–116. DOI: https://doi.org/10.1016/j.is.2011.01.003

- Utilisation patterns

  S. Vuik, E. Mayer, and A. Darzi. "A quantitative evidence base for population health: Applying utilization-based cluster analysis to segment a patient population". In: *Population Health Metrics* 14 (Dec. 2016). DOI: 10.1186/s12963-016-0115-z

Take home message

- Don't impose a framework
- Data-driven solutions
- Let the data speak for itself