Understanding cost variation with clustering

Henry Wilde

Supervised by: Dr Jonathan Gillard and Dr Vincent Knight

February 11, 2019

The data

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

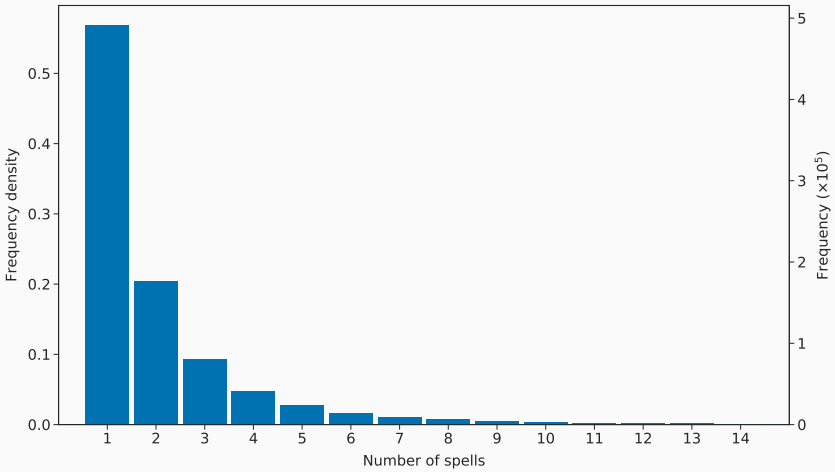
The data

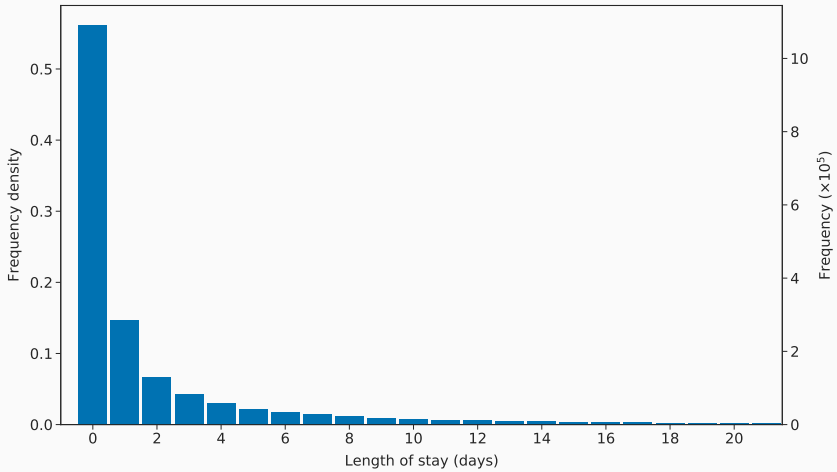
- The Cwm Taf University Health Board
- April 2012 through April 2017
- 2.4 million patient-episode records with 260 attributes

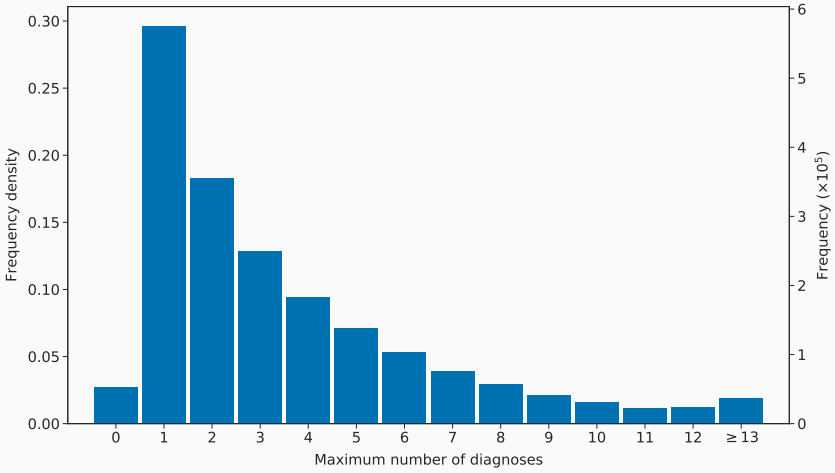
PATIENT ID	SPELL ID	EPISODE ID	Net Cost	Age (years)	HRG	Admission Date	Discharge Date	Length of Stay (days)
ID_123456	M1001	M1001-1	858.14	74	EA05Z	2015-05-06	2015-05-06	0.0
	M1211	M1211-1	333.95	74	FZ38F	2015-07-15	2015-08-01	17.0
		M1211-2	706.09	74	FZ38F	2015-07-15	2015-08-01	17.0
		M1211-3	8671.31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

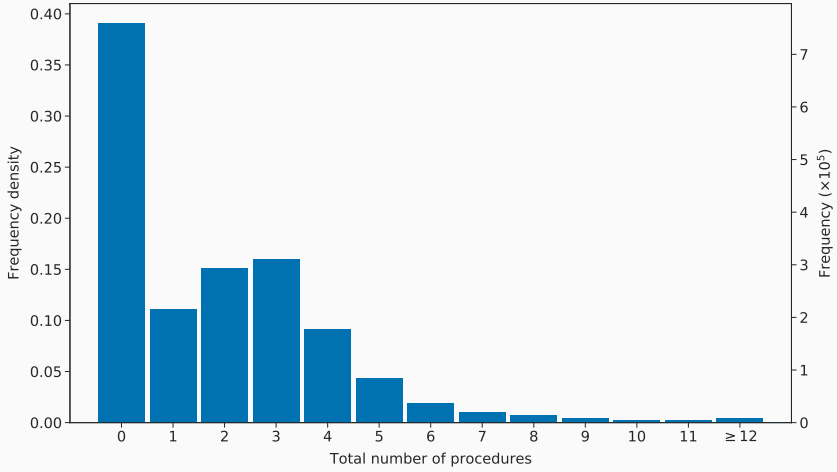
An overview of the data

An overview

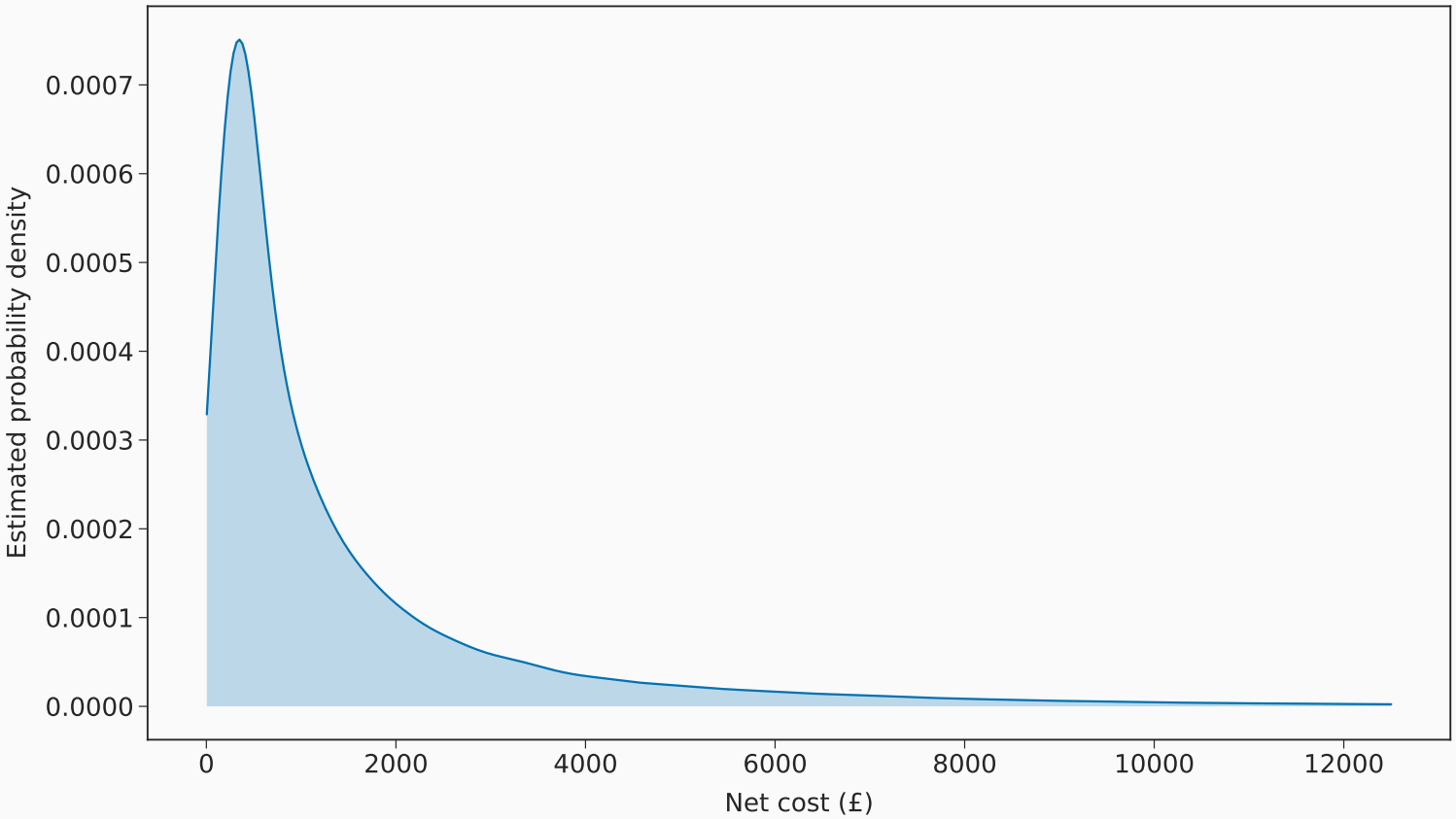

Frequency density
Frequency ($\times 10^5$)
Number of spells
1 2 3 4 5 6 7 8 9 10 11 12 13 14


Frequency density
Frequency ($\times 10^5$)
Length of stay (days)
0 2 4 6 8 10 12 14 16 18 20


Frequency density
Frequency ($\times 10^5$)
Maximum number of diagnoses
0 1 2 3 4 5 6 7 8 9 10 11 12 ≥ 13

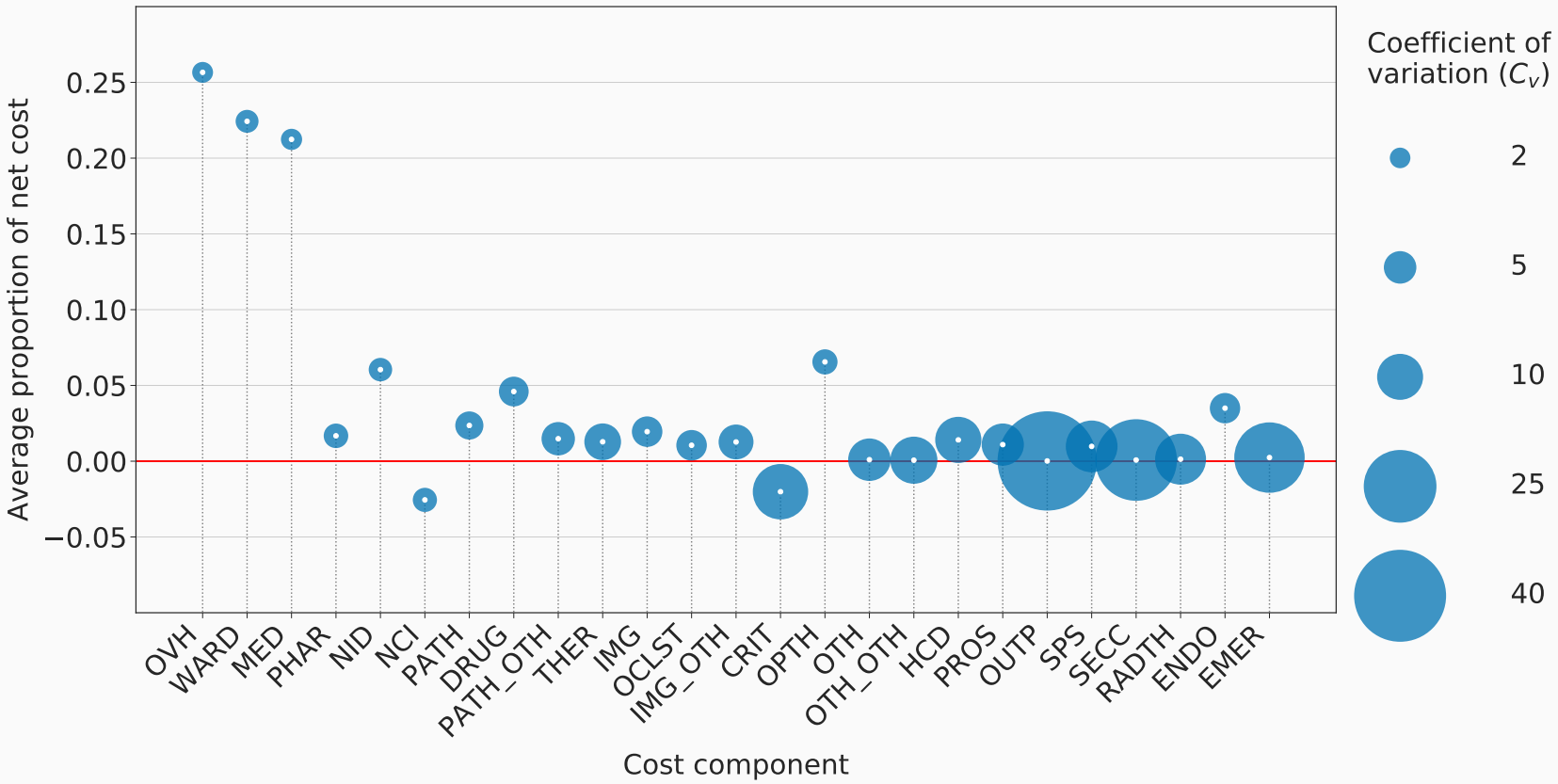

Frequency density
Frequency ($\times 10^5$)
Total number of procedures
0 1 2 3 4 5 6 7 8 9 10 11 ≥ 12

Net costs



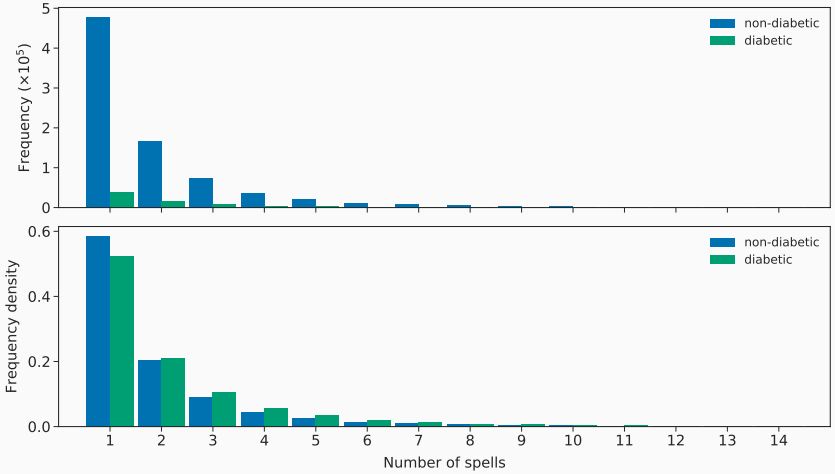
	NetCost
mean	1,737.65
std	3,160.54
min	4.50
1%	62.55
25%	347.07
50%	745.51
75%	1,859.00
95%	6,554.91
99%	14,183.23
max	369,168.93

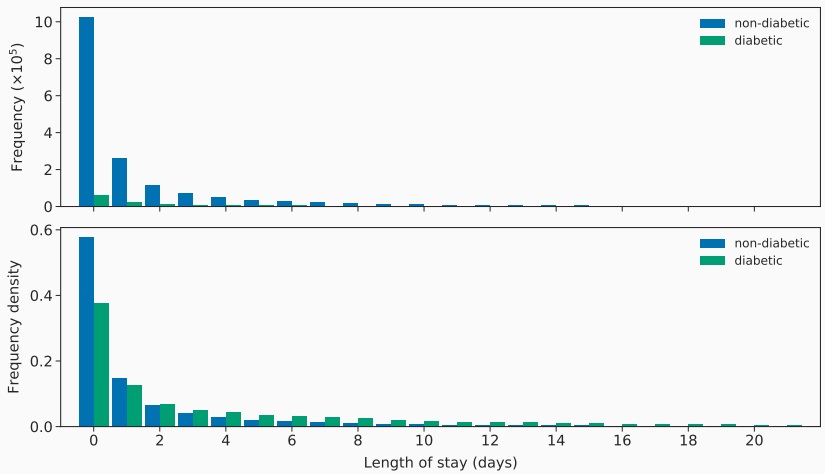
Variation and importance

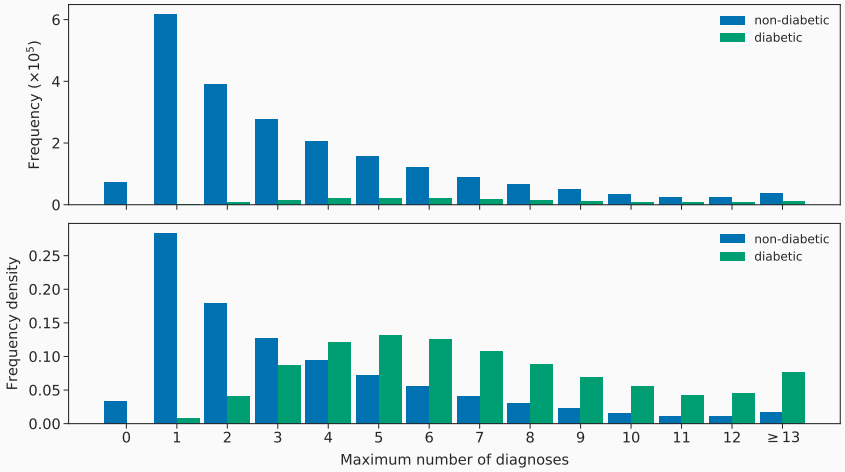


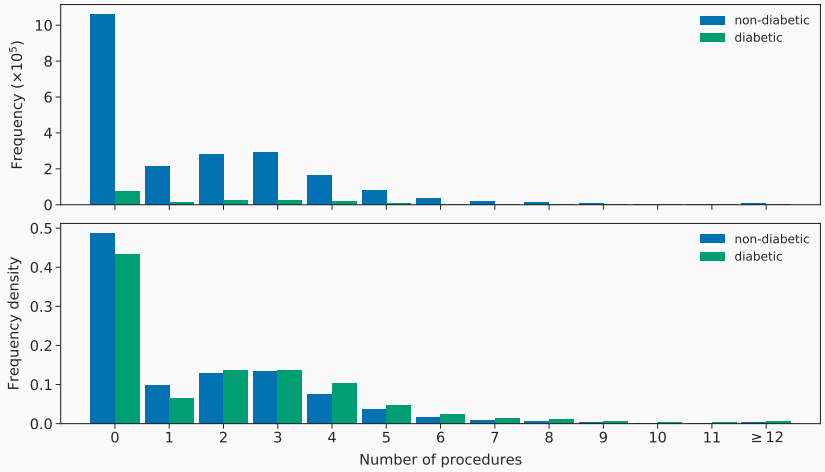
Diabetic patient analysis

An overview

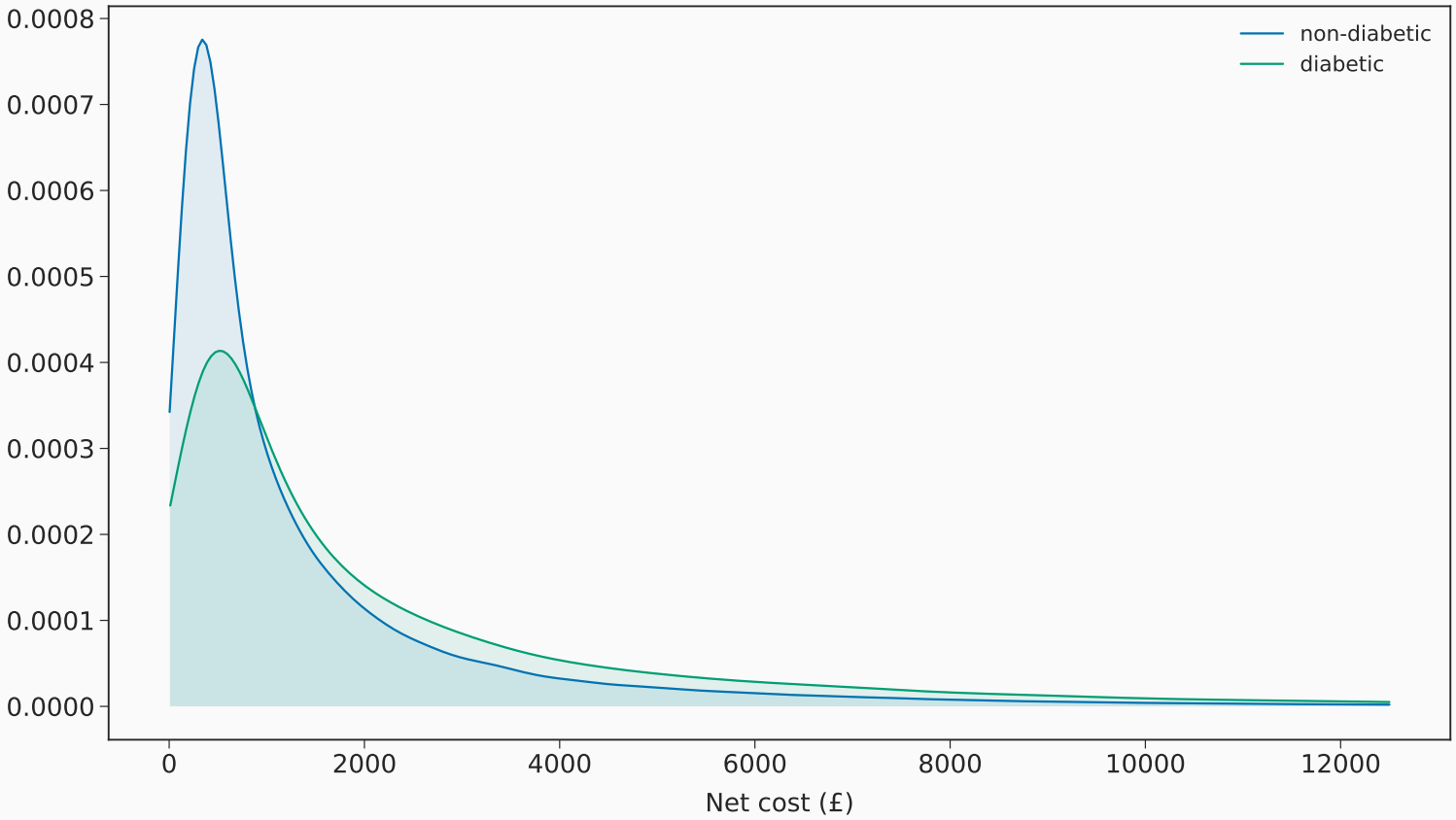

non-diabetic
diabetic
Frequency (×10⁵)
Frequency density
Number of spells


non-diabetic
diabetic
Frequency (×10⁵)
Frequency density
Length of stay (days)


non-diabetic
diabetic
Frequency (×10⁵)
Frequency density
Maximum number of diagnoses

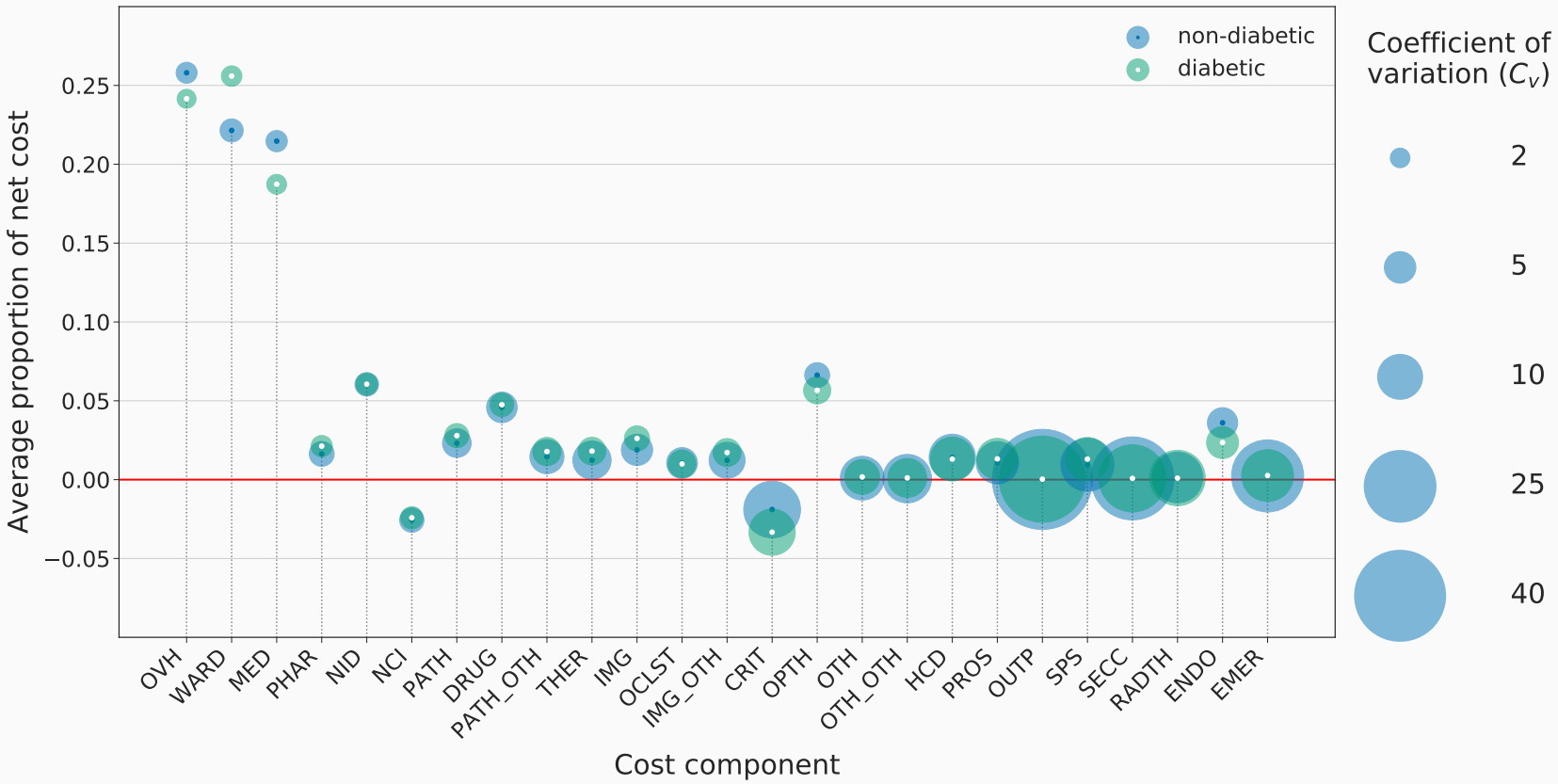

non-diabetic
diabetic
Frequency (×10⁵)
Frequency density
Number of procedures

Net costs



	Non-diabetic	Diabetic
mean	1,647.01	2,648.98
std	3,019.54	4,152.20
min	4.50	10.91
1%	62.55	139.65
25%	338.67	490.64
50%	709.33	1,227.95
75%	1,756.90	3,106.44
95%	6,179.92	9,591.06
99%	13,414.47	19,128.45
max	369,168.93	273,450.30

Variation and importance



Partitioning the data

Subsets in the data

Traditional methods include[1]:

- condition-specific populations
- segmenting by age

[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.

Subsets in the data

Traditional methods include[1]:

- condition-specific populations
- segmenting by age

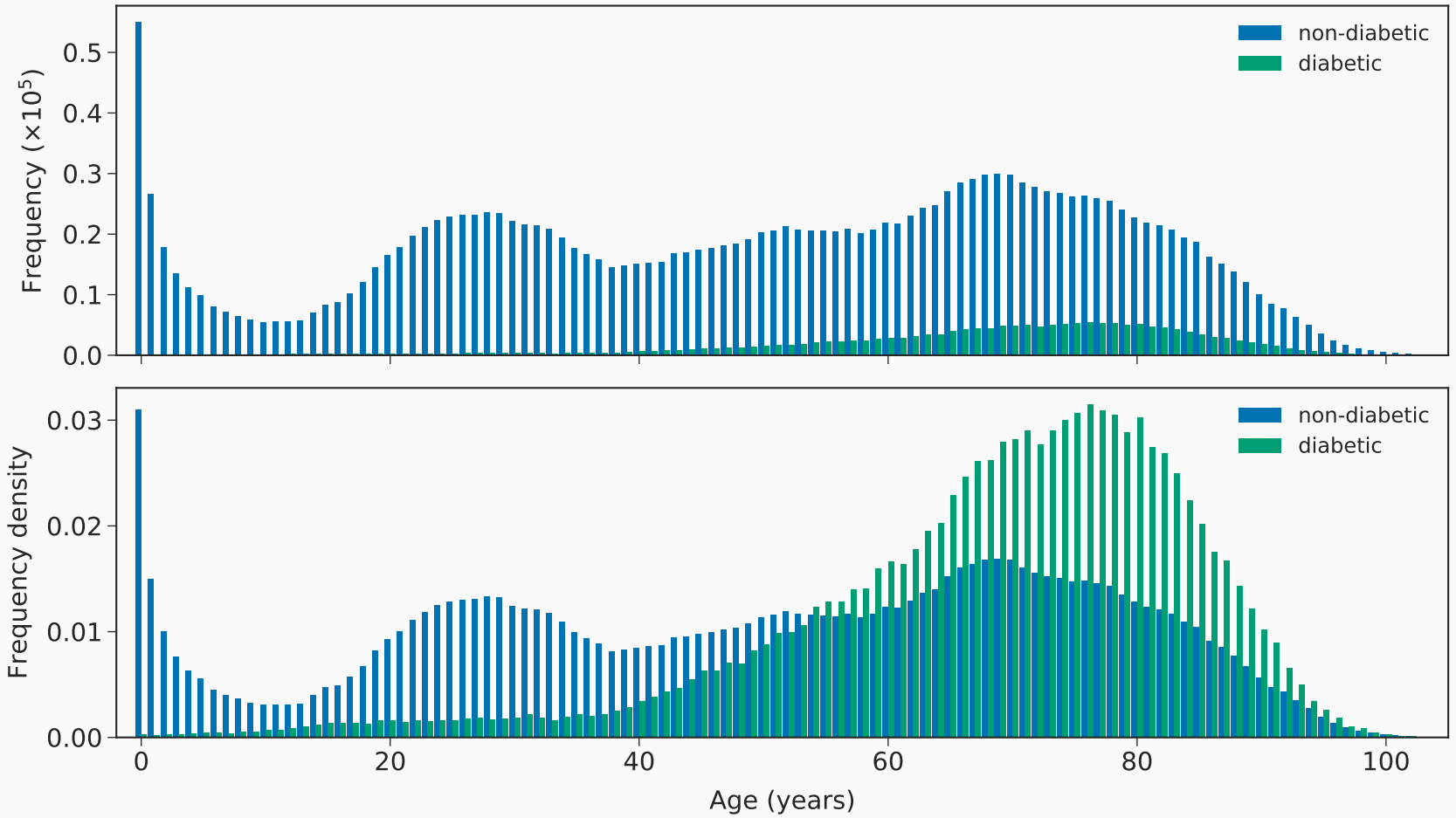
However, these methods have the common flaw of under-representing groups of healthcare users[2].

---

[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



What is clustering?

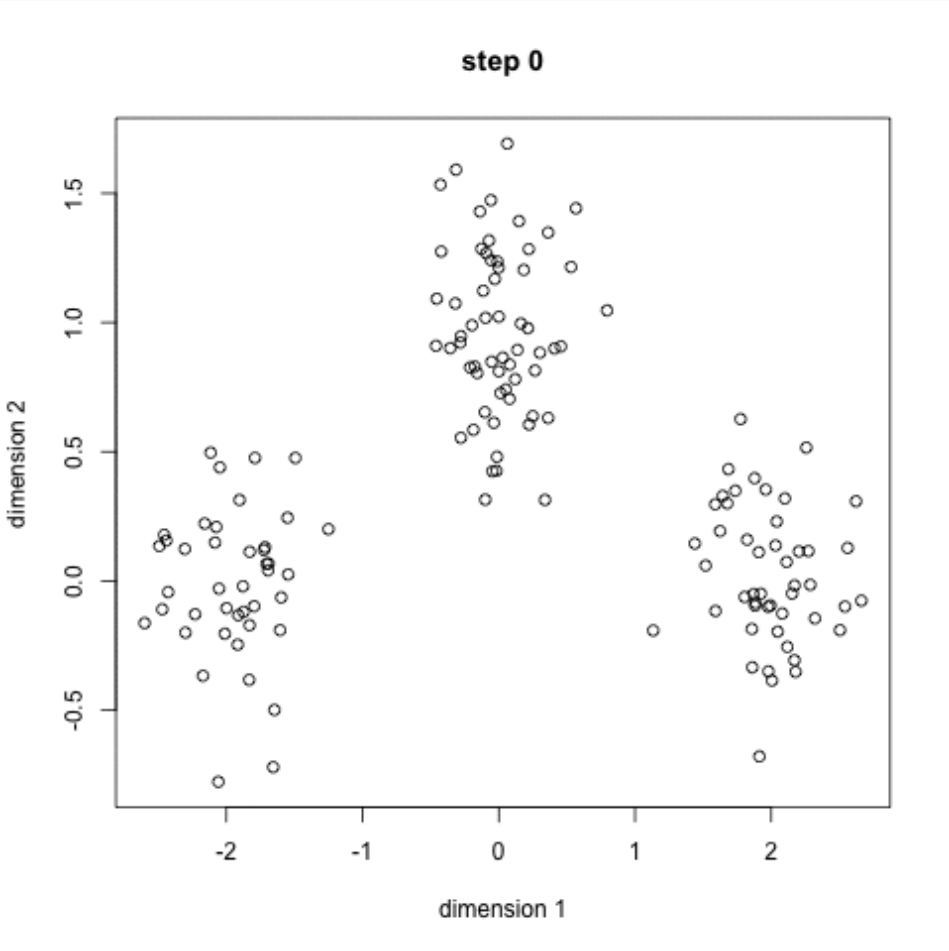


Image taken from `https://www.jeremyjordan.me/grouping-data-points-with-k-means-clustering/`

Clustering with healthcare data

- P. Kalyani. "Approaches to Partition Medical Data using Clustering Algorithms". In: *International Journal of Computer Applications* 49.23 (July 2012), pp. 7–10. DOI: 10.5120/7941-1102
- 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
- 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