Community Fit – Patient Taxonomies

Exploring potential patient taxonomies using a linked dataset.

20/05/2016

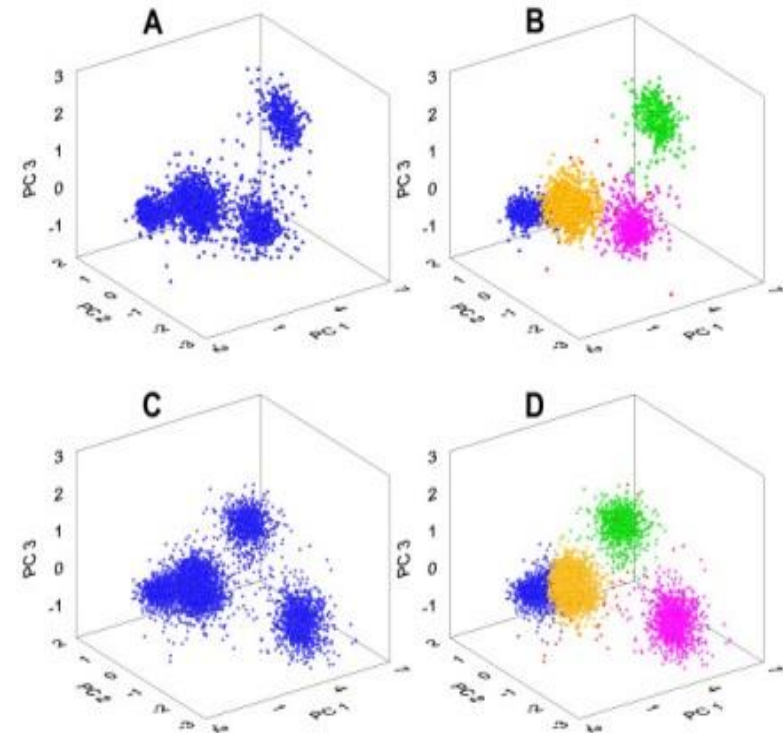# Cluster analysis

Task to be solved

A general question facing researchers in many areas of inquiry is how to organise observed data into meaningful structures. Cluster analysis is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).

A cluster is a group of relatively homogeneous cases or observations. Objects in a cluster are similar to each other. They are also dissimilar to objects outside the cluster, particularly objects in other clusters.

Applications of cluster analysis

- Marketing—cluster analysis is often used to support market segmentation. The knowledge associated with discovery of distinct groups within an organisation's customer base allows for the development of a targeted approach to marketing.

- Healthcare—to examine the heterogeneity of home care clients who use rehabilitation services. Cluster analysis was used to segment a heterogeneous rehabilitation client population into more homogeneous subgroups. The analysis provided an enhanced understanding of client characteristics and needs, and could enable more appropriate targeting of rehabilitation services for home care clients.(1)

Images are from a neuroscience research study. Cluster analysis was applied to data from electrophysiological recordings to sort spiking activity of neurons for investigating brain function.(2) Boxes B and D show clusters of spike waveforms.

(1) Armstrong JJ, Zhu M, Hirdes JP, Stolee P. K-means cluster analysis of rehabilitation service users in the Home Health Care System of Ontario: examining the heterogeneity of a complex geriatric population. *Arch Phys Med Rehabil*. 2012;93(12):2198-205.

(2) Oliynyk A, Bonifazzi C, Montani F, Fadiga L. Automatic online spike sorting with singular value decomposition and fuzzy C-mean clustering. *BMC Neurosci*. 2012;13:96.

Cluster analysis

Aims and goals of cluster analysis:

- To draw out similarities among subsets of the data
- A 'good' clustering means that points in the same cluster should be more similar to each other than they are to points in other clusters
- Different clustering algorithms will give different results
- There are various imperfect methods ('heuristics') for estimating the best number of clusters

The method that we chose to use is the popular k means method – specifically the Hartigan-Wong algorithm. For this algorithm to work, you must initially supply the number of clusters that you want to be returned.

We used a linked dataset, with data from 2014/15 covering multiple sectors to run this analysis.

- Acute
- Community
- Mental health
- Social care

Unfortunately we did not have a complete set of primary care data which would have further enriched the basic dataset.

We used this data to derive clusters of people – groups of people who the data tells us are more similar to one another than to patients in other clusters.

Hartigan, J. A. and Wong, M. A. (1979). A K-means clustering algorithm. *Applied Statistics* **28**, 100–108.

Methods

How do we decide what value of k (the number of clusters) to use?

Part of conducting a cluster analysis using the algorithm we chose (k-means) involves making an initial determination of the number of clusters to aim for.

The initial choice was guided by previous work and discussions of the aims of this work, and initially the number of clusters, k, was set to 15.

As the analysis progressed, the decision was made to separate out people who had contact with a single sector only. These became 4 separate pre-determined 'clusters' (they were not run through the algorithm).

After this, and with further runs, the optimal number of clusters was estimated using the Calinski Harabasz index.

How do we know that the clusters we have identified are stable?

After each attempt at clustering, an evaluation of the stability of the output clusters was made using resampling methods (boostrap).

Valid results were found using different numbers of clusters, but with more than 12 clusters, there were some doubts on the stability of the additional clusters (the clusters were 'dissolved' more often in the bootstrap).

So in the end we present results in 12 + 4 clusters – the 12 clusters run through the algorithm and the 4 'clusters' of single sector activity we earlier separated from the main analysis.

Variables used

- Total cost
    - Acute
    - Community
    - Mental health
    - Social care
- Number of points of delivery the person had contact with
    - Acute
    - Community
    - Mental health
    - Social care
- Number of days between first and last contact
- Average number of contacts per day between first and last contact
- Proportion of contacts occurring before the mid-point of contact period
- The number of periods lasting more than 14 days between the date of the first and last contact

Note that these were standardised to be between -1 and 1 before running the algorithm.

Each variable has an equal weight in the analysis.

Clustering was applied to c. 210,000 adult patients from Shropshire and Telford with a recorded contact with acute, community, mental health or social care contact within 2014/15 where the patients NHS number and age were recorded.

Potential Applications

We hope that this analysis will provide stakeholders with;

-   An enhanced understanding the common ways in which patients use health and social care services in Shropshire and Telford

-   Analysis to guide thinking about the potential to re-organise care – particularly for those patients having contact with more than one sector

-   Information to support considerations about the packages of community support that might substitute for acute care packages

Results

Cluster characteristics

Cluster	Cluster Name	Cluster size (n)	Cluster size (%)	Average cost £)	Average Age	Average acute cost (£)	Average community cost (£)	Average mental health cost (£)	Average social care cost (£)
Multi-sector patients									
11	Complex frail elderly	2,516	1.2%	18,913	76	5,071	2,186	1,873	9,315
6	Complex needs – mainly managed in community	273	0.1%	32,606	73	5,217	19,185	444	7,430
3	Long stay acute care	3,514	1.7%	13,360	70	11,860	1,027	85	288
4	Younger adults with complex disabilities	234	0.1%	94,599	44	995	1,307	2,469	89,689
2	Intensive / institutional mental healthcare	180	0.1%	51,672	57	2,728	741	42,714	5,230
7	Young adults with simpler mental health needs	5,121	2.4%	3,420	50	1,028	53	1,787	431
5	Well-maintained social care users	1,486	0.7%	13,040	61	548	305	148	11,863
1	Discrete planned care with community follow-up	12,571	6.0%	2,815	62	2,588	191	2	14
8	Community support-occasional acute input	7,940	3.8%	2,751	71	1,574	949	30	186
9	Community & social care – occasional acute input	6,075	2.9%	5,728	74	2,288	965	468	1,713
10	Simple investigations – physical health	20,041	9.5%	642	61	431	172	4	16
12	Simple investigations – physical/mental health	3,775	1.8%	824	55	233	128	423	1
Single-sector patients									
13	Acute contact only	113,689	54.0%	931	50	931	-	-	-
14	Community contact only	30,557	14.5%	267	60	-	267	-	-
15	Mental Health contact only	270	0.1%	2,520	50	-	-	2,520	-
16	Social Care contact only	2,237	1.1%	5,880	52	-	-	-	5,880

Cluster characteristics (2)

	% of patients	% of spend	average annual cost per patient (£'000)
Complex frail elderly	1%	12%	19
Complex needs – mainly managed in community	0%	2%	33
Long stay acute care	2%	12%	13
Younger adults with complex disabilities	0%	5%	95
Intensive / institutional mental healthcare	0%	2%	52
Young mental health needs (non-psychotic)	2%	4%	3
Well-maintained social care users	1%	5%	13
Discrete planned care with community follow-up	6%	9%	3
Community support-occasional acute input	4%	5%	3
Community & social care – occasional acute…	3%	9%	6
Simple investigations – physical health	10%	3%	1
Simple investigations – physical/mental health	2%	1%	1
Acute contact only	54%	26%	1
Community contact only	15%	2%	0
Mental Health contact only	0%	0%	3
Social Care contact only	1%	3%	6

Multi sector patients

NHS

The Strategy Unit

Cluster 11
Complex frail elderly

Average age 76 – oldest cluster
Average cost £19,000
2,500 people

Cluster 11 - Frail elderly, multi-sector care

Frail elderly, multi-sector care

Cluster 6
Complex needs – mainly managed in community

Community dominated cluster
Average cost £33,000
Average age 73
270 people

F   M

Age

Number of people

Cluster 6 - High cost Community

Percentage

Annual cost (£)

0 – 2499   2500-4999   5000-7499   7500-9999   10000+

High cost Community

Percentage of patients

Sector

Acute   Community   MentalHealth   SocialCare

Cluster 3
Long stay acute care

This group is a higher cost variant of cluster 1.
3,500 people
Average age 70
Average cost £13,400

Cluster 3 - Medium cost, complex acute

Medium cost, complex acute

Cluster 4
Younger adults with complex disabilities

Youngest cluster – average age 44
Most expensive cluster, the average cost is £94,600.
Over 90% of the cost is in social care.

Cluster 4 - Very high cost, high intensity social care, young

Very high cost, high intensity social care, young

Cluster 2
Intensive / institutional mental healthcare

This small group – the smallest overall, at 180 people (0.1%) – contains the high cost and high intensity mental health patients.
Average cost of £52,000

Cluster 2 - Very high cost, high intensity mental health

Very high cost, high intensity mental health

Cluster 7

Younger adults with simpler mental health needs

Average age 50 – comparatively young group.
Mental health dominates costs which average £3400.
5,100 people (8%)

Cluster 7 - Younger mental health patients

Younger mental health patients

Cluster 1
Discrete planned care with community follow-up

This cluster contains a lot of elective activity.
6% of all people – (12,600)
Average age 62.
Average cost £2,800 - 90% of the costs related to the acute sector.

Cluster 1 - Discrete planned care input

Discrete planned care input

Cluster 5
Well-maintained social care users

1,500 people
Average age 61
Average costs £13,000
Bi-modal cost distribution

Cluster 5 - Low intensity, long-term social care users

Low intensity, long-term social care users

Cluster 8
Community support-occasional acute inpu

Well managed group.
7,900 people
Average cost £1600 – over an average of 30 contacts.
Average age 71 – an older cluster than most others.

Cluster 8 - Low intensity - community and acute

Low intensity - community and acute

Cluster 9

Community & social care – occasional acute input

Old cluster – average age 74
Average cost - £5700
6,100 people

Cluster 9 - Lower intensity elderly, multi-sector care

Lower intensity elderly, multi-sector care

Cluster 10
Simple investigations – physical health

Average age 61

Average cost £640 – the cheapest non-single sector cluster.

20000 patients – the largest non-single sector cluster.

F   M

Cluster 10 - Low cost, medium term, mostly acute care

Low cost, medium term, mostly acute care

Cluster 12
Simple investigations – physical / mental health

Average cost £800
3800 people
Average age 55
Any social care cost is almost all assessments.

Cluster 12 - Low cost, short-term mostly acute care

Low cost, short-term mostly acute care

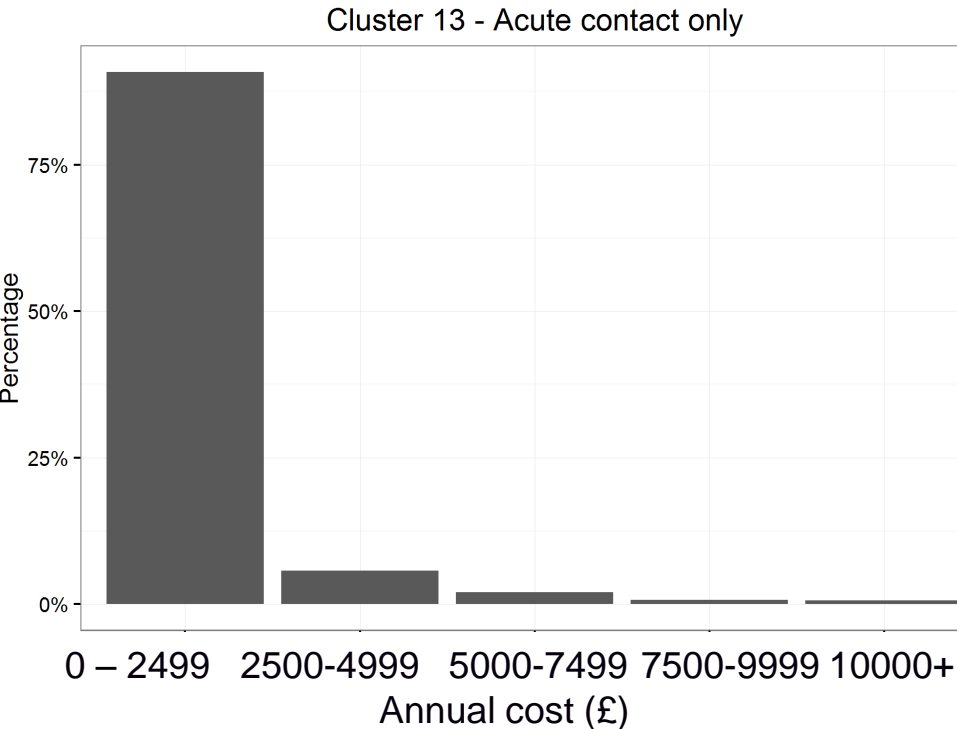Single sector patients

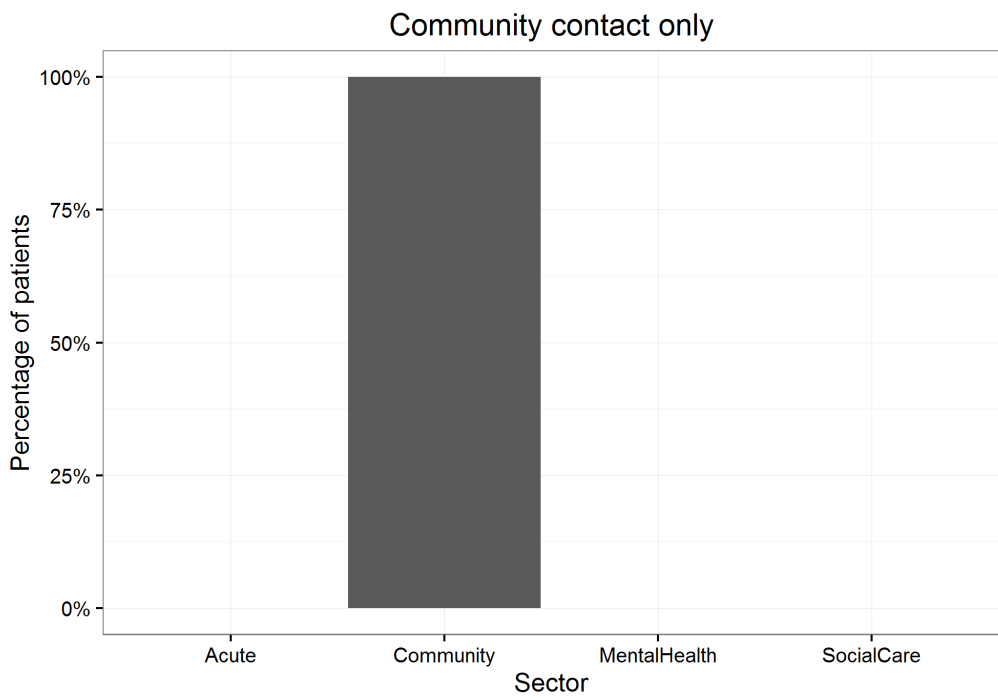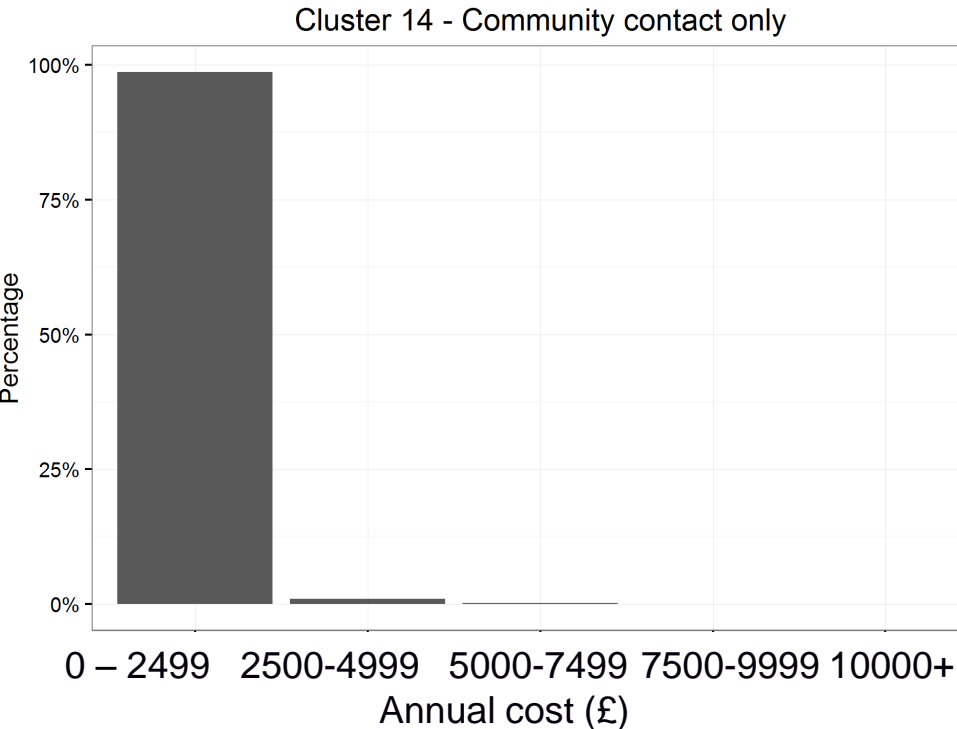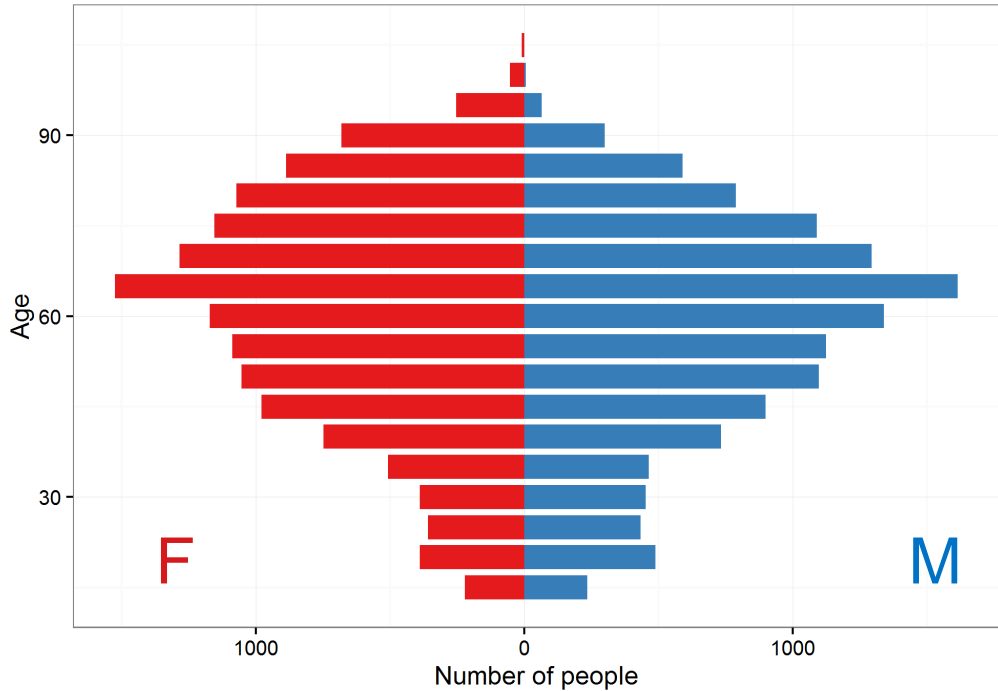NHS

The Strategy Unit

Cluster 13
Acute contact only

Average age 50
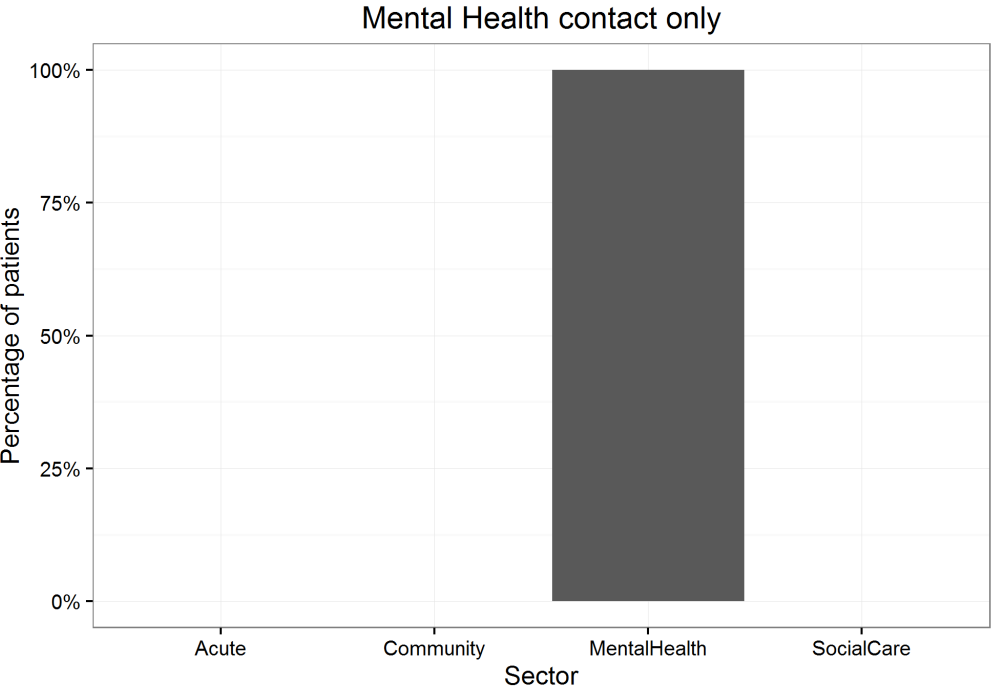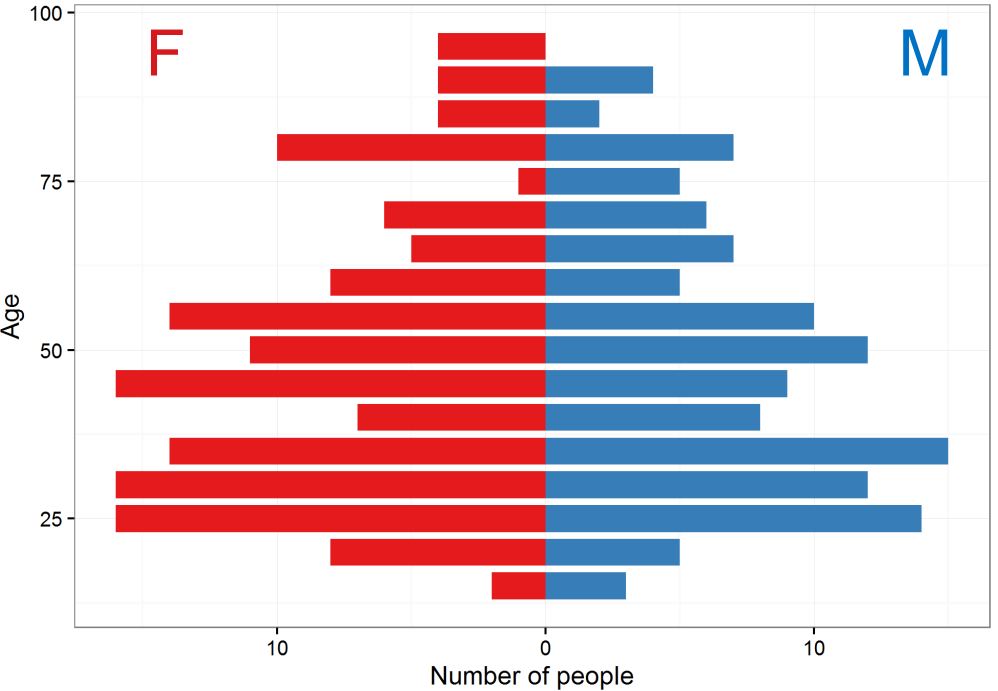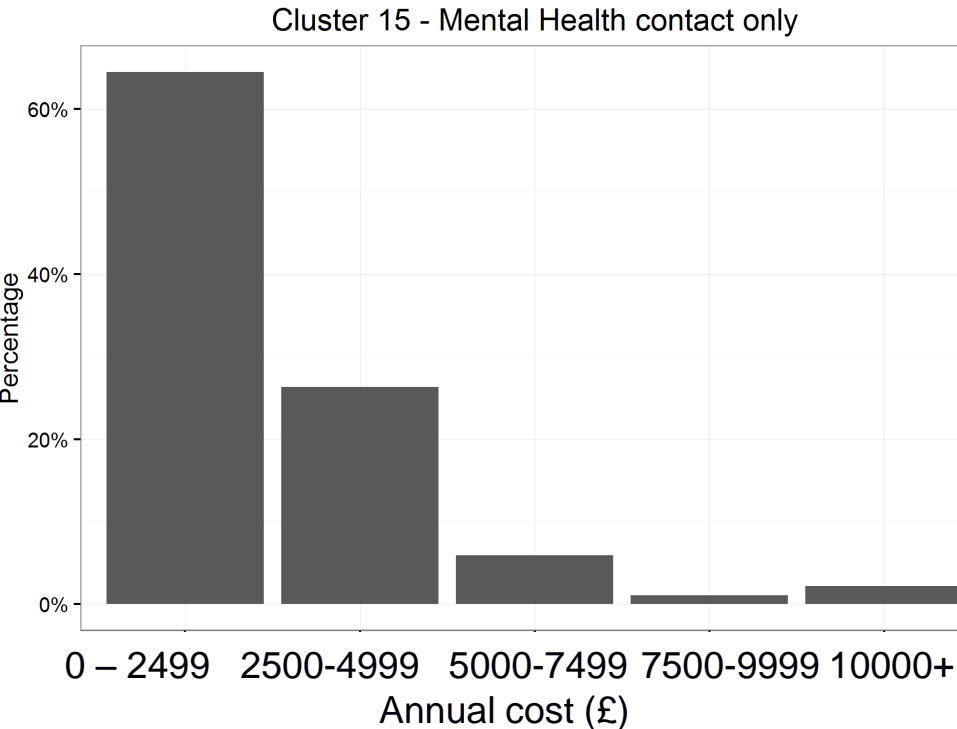Represents 54% of people (114,000)
Average cost £930

Cluster 13 - Acute contact only

Acute contact only

Cluster 14
Community contact only
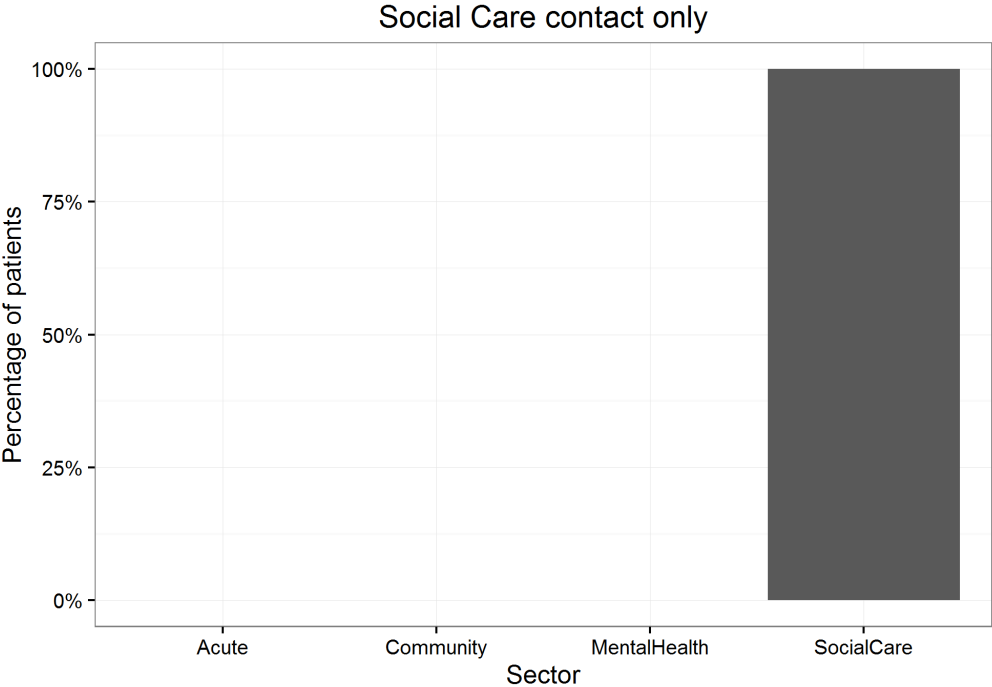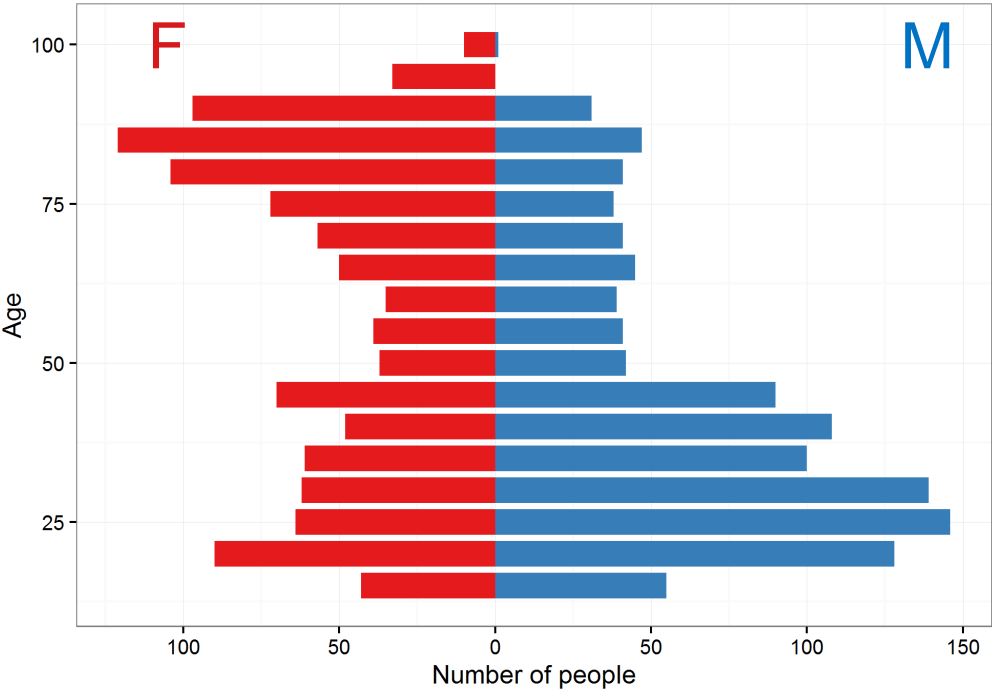
Average cost £270
31,000 people
Average age 60

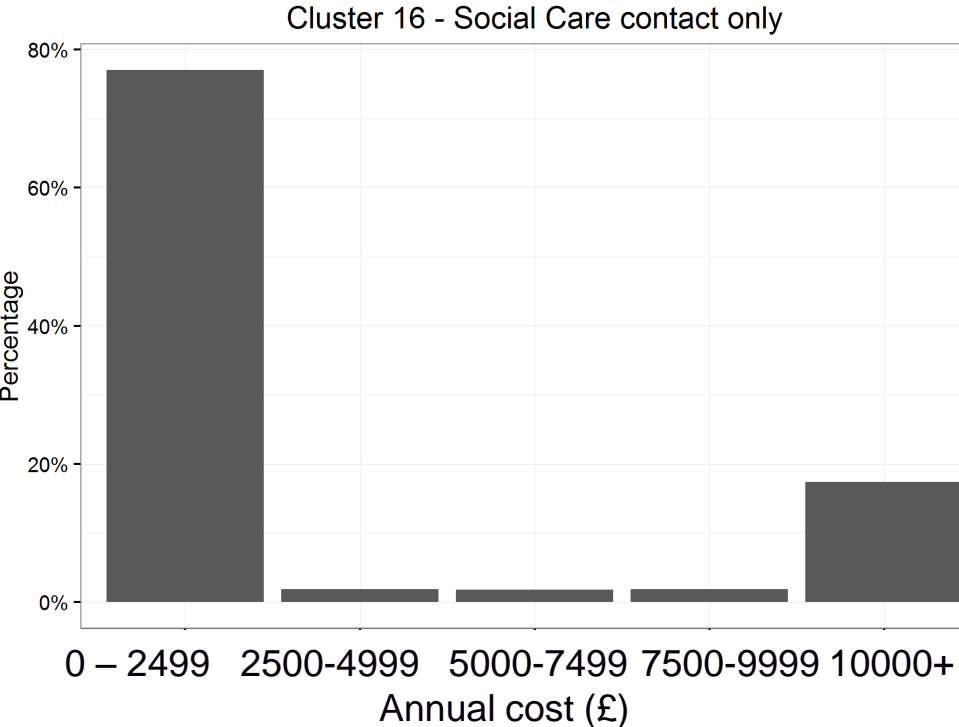Cluster 14 - Community contact only

Community contact only

Cluster 15
Mental health contacts only

270 people
Average cost £2500.
Average age 50.

Cluster 15 - Mental Health contact only

Mental Health contact only

Cluster 16
Social care contact only

2,200 people.
Average cost £5900 – but distribution is
concentrated in low and high cost activity.
Average age 52.

Cluster 16 - Social Care contact only

Social Care contact only

Final thoughts

What are the limitations of this analysis?

We only had a full set of data for the year 2014/15. Some of the variables used in the cluster analysis look at how the intensity of patients care evolves, if we had a full set of data for the period surrounding 2014/15, the refinement of the clusters may be improved.

Clustering using k-means requires making judgements about the data prior to classifying it. The method of k-means also performs better when we know that clusters are of a similar size, which we know is not the case here.