University of BRISTOL

NHS
Bristol, North Somerset and South Gloucestershire
Clinical Commissioning Group

Mini Huddle: Characterising COVID-19 Shielded Patients in R

We will show how we used our System-wide linked dataset (PHM dataset) to identify covid-19 vulnerability early in the pandemic, and then adapted our approach to identify and study the shielded patient population through description and segmentation.

Our system covers one million people Bristol, North Somerset and South Gloucestershire, an area including a major city, rural areas and coastal towns.

Created by
Charlie Kenward (Clinical Lead for Research and Improvement)
Jennifer Cooper (Senior Research Associate at BNSSG CCG/Bristol University)

Agenda

- Our team and Population Health Management Programme
- Context
- Early work
- Segmenting COVID-19 shielded patients in R using cluster analysis to further characterise this population
- Implications
- Questions and discussion

Introduction

- Our team
- Population Health Management Programme
- There’s a pandemic on
- Initial request to identify vulnerability to severe Covid-19
- Adapting for the shielded patient list; a PHM action research approach

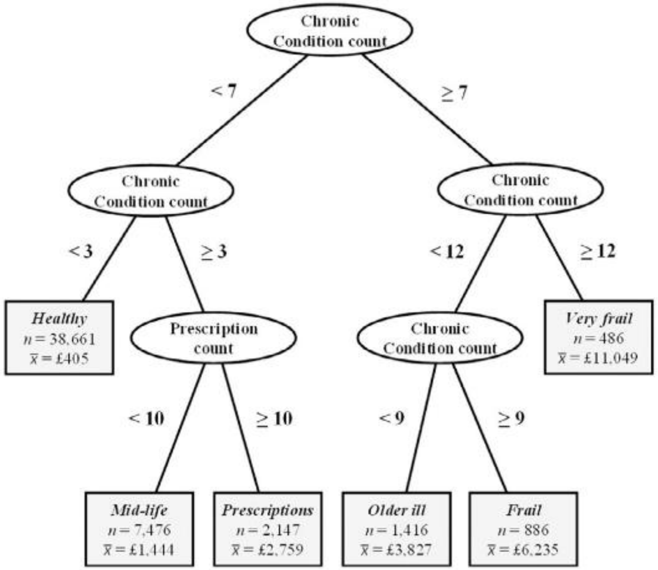
Research Aim

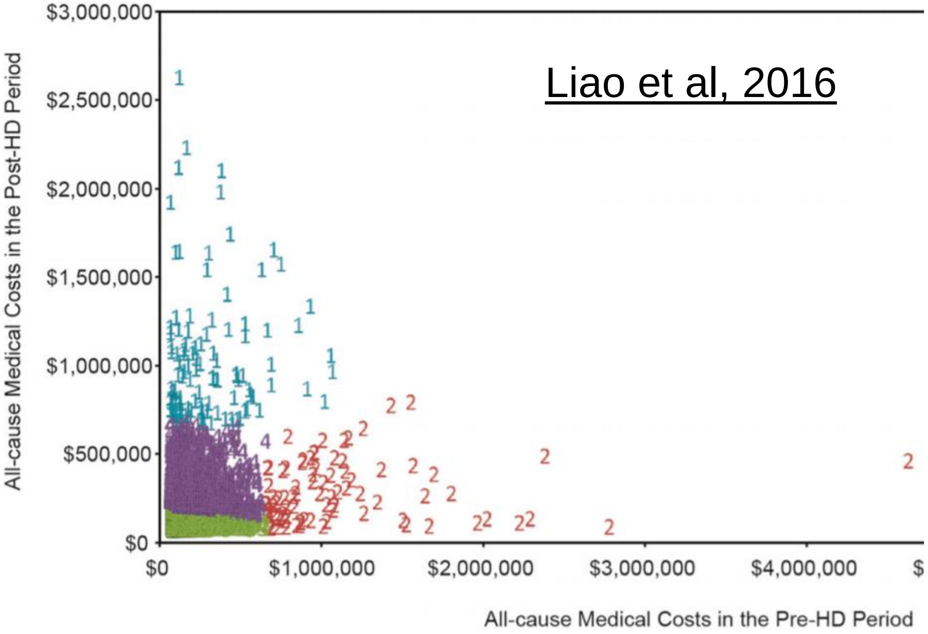
The aim of this research was to use Population Health Management (PHM) methods to identify and characterise a high-risk population for which shielding is required, for the purposes of managing ongoing health needs and mitigating potential shielding-induced harm.

Set in a large healthcare system in South West England during the early stages of the COVID-19 pandemic, this research made use of a System-wide linked dataset containing healthcare activity and clinical, demographic and social attributes for one million individuals.

Methods: Cluster Analysis

- Shielded patients not a homogenous group.
- To further characterise the shielded population we can use ‘population segmentation’ approaches.
- Methods for population segmentation: Decision trees (CART), cluster analysis, judgemental splits, prescribed binning criteria (*Wood, Murch & Betteridge, 2019*)


Chronic Condition count
< 7
≥ 7
Chronic Condition count
Chronic Condition count
< 3
≥ 3
< 12
≥ 12
Healthy n = 38,661 x̄ = £405
Prescription count
Chronic Condition count
Very frail n = 486 x̄ = £11,049
< 10
≥ 10
< 9
≥ 9
Mid-life n = 7,476 x̄ = £1,444
Prescriptions n = 2,147 x̄ = £2,759
Older ill n = 1,416 x̄ = £3,827
Frail n = 886 x̄ = £6,235

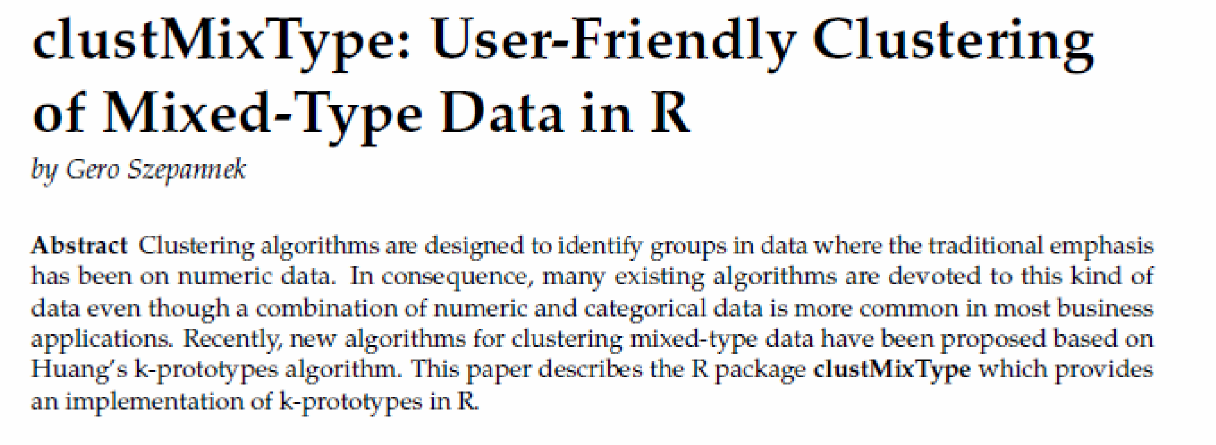


Liao et al, 2016

Methods: Cluster Analysis

- Unsupervised machine learning approach
- Exploratory tool when no a priori hypotheses or outcomes to model
- Groups people into similar groups referred to as ‘clusters’
- Several clustering algorithms: k-means, k-modes, k-prototypes, hierarchical clustering
- k-prototypes can be used for mixed data types
- clustMixType package in R employs Huang’s k-prototypes algorithm (Huang, 1998)

Cluster Analysis Using R: clustMixType Package

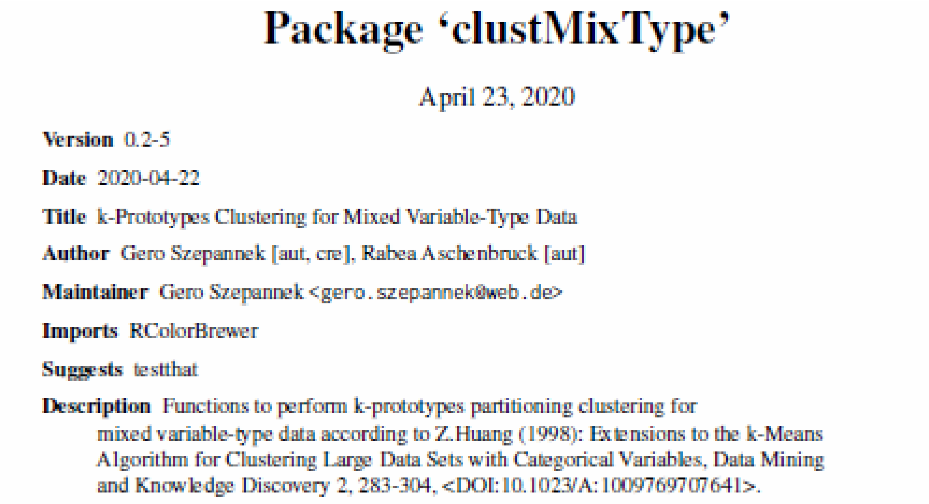


clustMixType: User-Friendly Clustering of Mixed-Type Data in R

by Gero Szepannek

Abstract Clustering algorithms are designed to identify groups in data where the traditional emphasis has been on numeric data. In consequence, many existing algorithms are devoted to this kind of data even though a combination of numeric and categorical data is more common in most business applications. Recently, new algorithms for clustering mixed-type data have been proposed based on Huang's k-prototypes algorithm. This paper describes the R package **clustMixType** which provides an implementation of k-prototypes in R.

- https://journal.r-project.org/archive/2018/RJ-2018-048/RJ-2018-048.pdf



Package 'clustMixType'

April 23, 2020

Version 0.2-5

Date 2020-04-22

Title k-Prototypes Clustering for Mixed Variable-Type Data

Author Gero Szepannek [aut, cre], Rabea Aschenbruck [aut]

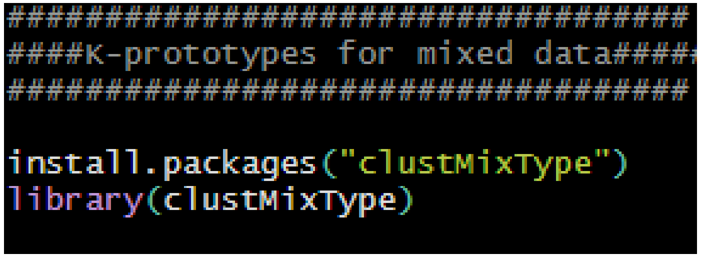
Maintainer Gero Szepannek <gero.szepannek@web.de>

Imports RColorBrewer

Suggests testthat

Description Functions to perform k-prototypes partitioning clustering for mixed variable-type data according to Z.Huang (1998): Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Variables, Data Mining and Knowledge Discovery 2, 283-304, <DOI:10.1023/A:1009769707641>.

- https://cran.r-project.org/web/packages/clustMixType/clustMixType.pdf

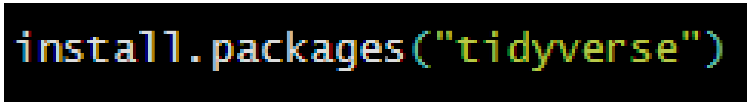


```
###################################
####K-prototypes for mixed data####
###################################

install.packages("clustMixType")
library(clustMixType)
```

Methods: Cluster Analysis

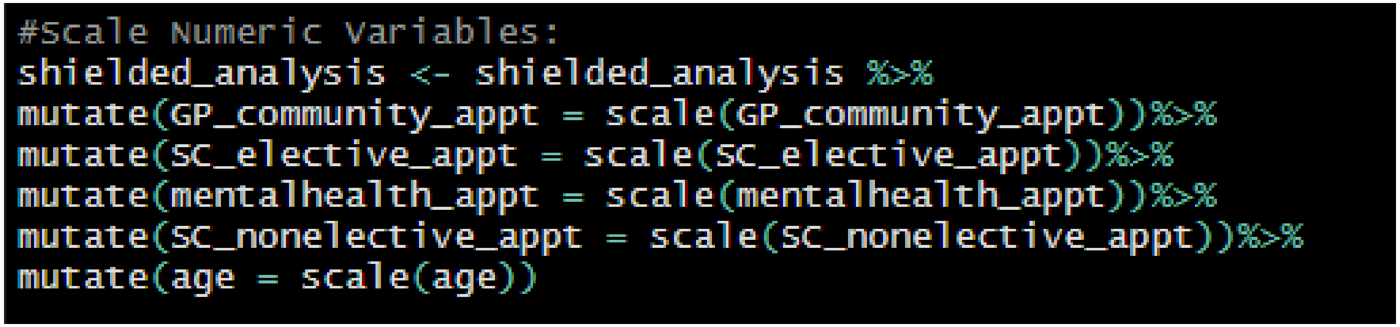
- Consider: selection of variables, data format, number of clusters
- Selection of clustering variables: demographic, clinical, and social attributes to gain a high-level understanding of the clusters and make the clusters actionable in terms of possible intervention. (n=29,454)
- Pre-process data before putting input into cluster algorithm



```
install.packages("tidyverse")
```

-Ensure variables are either numeric type or factor type in R: as.factor, as.numeric

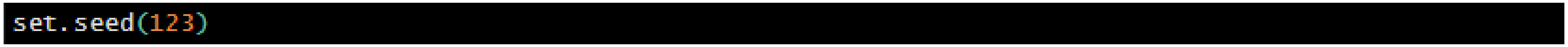
-(Centre) and scale numeric variables



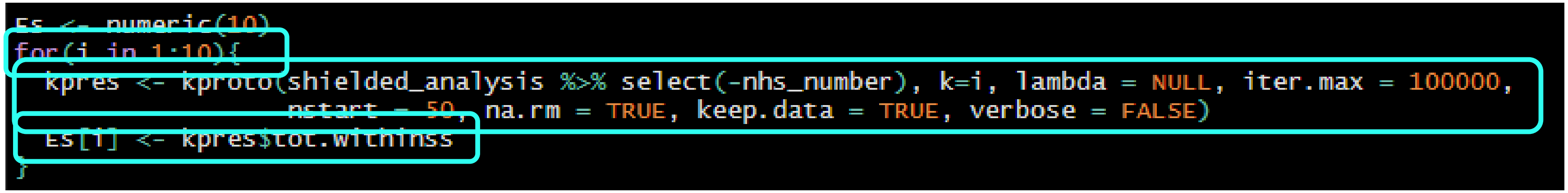
```
#Scale Numeric Variables:
shielded_analysis <- shielded_analysis %>%
mutate(GP_community_appt = scale(GP_community_appt))%>%
mutate(SC_elective_appt = scale(SC_elective_appt))%>%
mutate(mentalhealth_appt = scale(mentalhealth_appt))%>%
mutate(SC_nonelective_appt = scale(SC_nonelective_appt))%>%
mutate(age = scale(age))
```

Determining the Number of Clusters

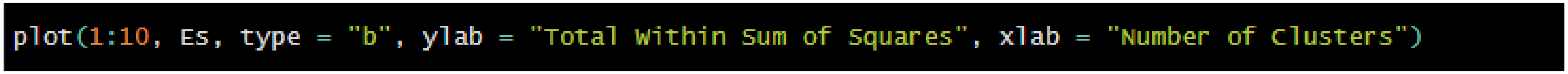
- The number of clusters (or segments) was determined both empirically from the data and by interpretation of the segments and clinical context



```
set.seed(123)
```

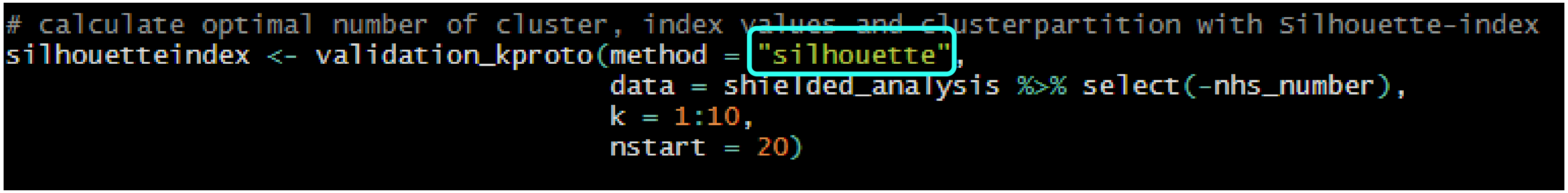


```
Es <- numeric(10)
for(i in 1:10){
  kpres <- kproto(shielded_analysis %>% select(-nhs_number), k=i, lambda = NULL, iter.max = 100000,
                  nstart = 50, na.rm = TRUE, keep.data = TRUE, verbose = FALSE)
  Es[i] <- kpres$tot.withinss
}
```



```
plot(1:10, Es, type = "b", ylab = "Total Within Sum of Squares", xlab = "Number of Clusters")
```

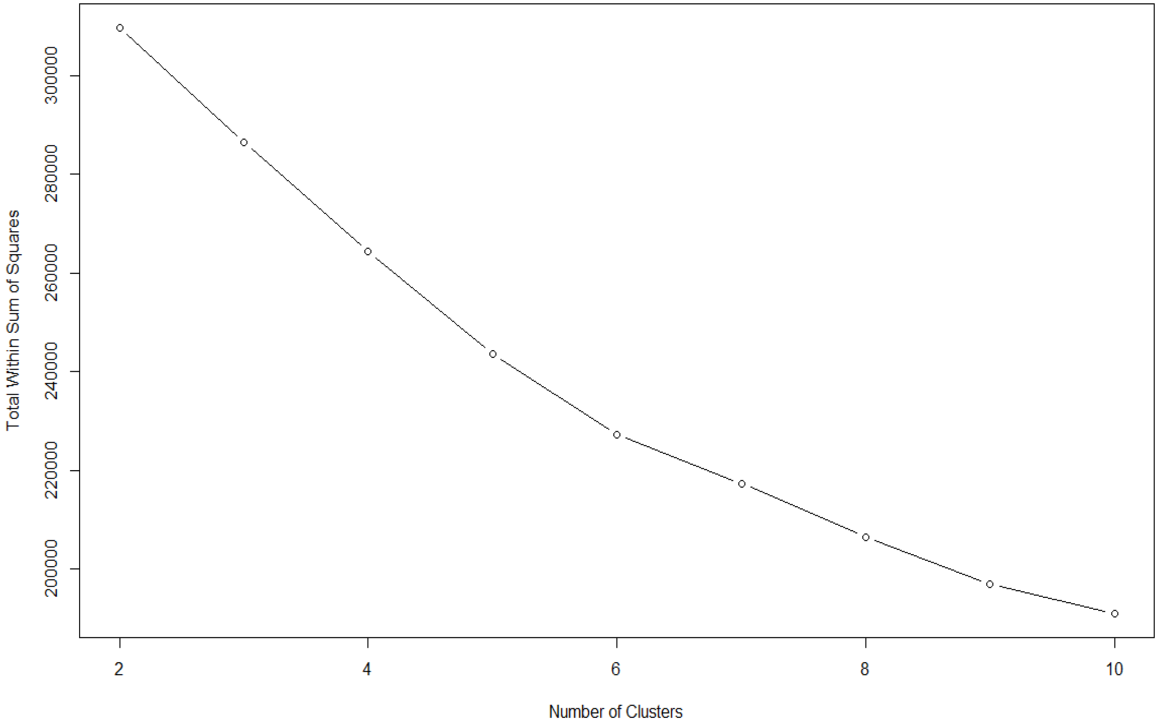
- Use of Indexes: c-index, dunn, gamma, gplus, mcclain, ptbiserial, silhouette, tau



```
# calculate optimal number of cluster, index values and clusterpartition with Silhouette-index
silhouetteindex <- validation_kproto(method = "silhouette",
                                     data = shielded_analysis %>% select(-nhs_number),
                                     k = 1:10,
                                     nstart = 20)
```

Shaping better health

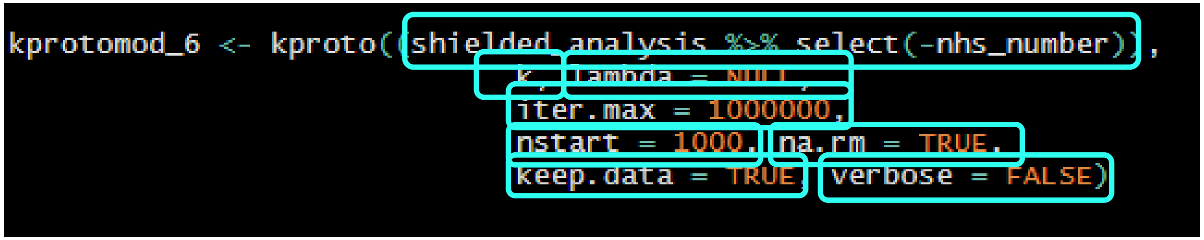
Determining the Number of Clusters



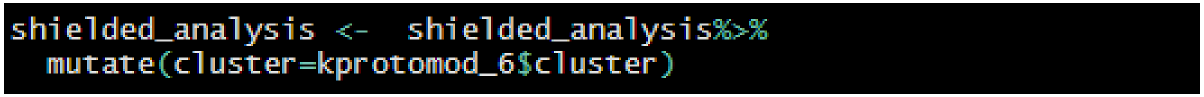
The k-prototypes clustering method was applied to 6,7,8 and 9 clusters and these were reviewed by two clinicians and assessed for their usefulness in terms 5 criteria (Chong, Lim and Matchar (2019))

Number of Clusters	Within Sum of Squares	Silhouette Index
2	309,757.1	0.160
3	286,488.8	0.174
4	264,354.7	0.190
5	243,689.9	0.187
6	227,313.3	0.192
7	217,342.0	0.196
8	206,578.7	0.178
9	196,894.5	0.193
10	191,006.9	0.184

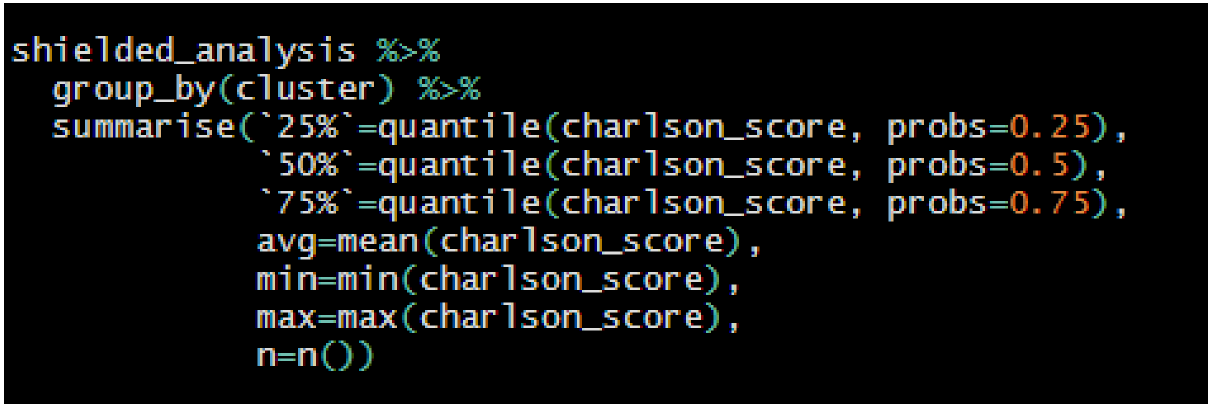
Implementing K-protoypes model for 6 clusters



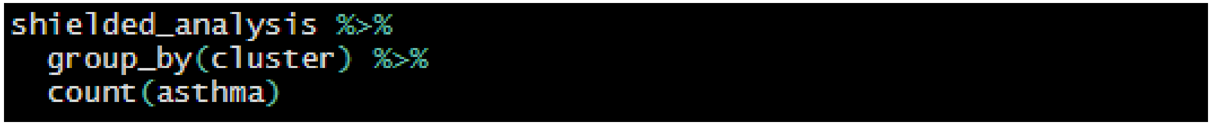
```
kprotomod_6 <- kproto((shielded_analysis %>% select(-nhs_number)),
                      k, lambda = NULL,
                      iter.max = 1000000,
                      nstart = 1000, na.rm = TRUE,
                      keep.data = TRUE, verbose = FALSE)
```



```
shielded_analysis <-  shielded_analysis%>%
  mutate(cluster=kprotomod_6$cluster)
```



```
shielded_analysis %>%
  group_by(cluster) %>%
  summarise(`25%`=quantile(charlson_score, probs=0.25),
            `50%`=quantile(charlson_score, probs=0.5),
            `75%`=quantile(charlson_score, probs=0.75),
            avg=mean(charlson_score),
            min=min(charlson_score),
            max=max(charlson_score),
            n=n())
```



```
shielded_analysis %>%
  group_by(cluster) %>%
  count(asthma)
```

- Implementing k-prototypes algorithm for 6 clusters
- Adding a cluster number variable onto the data
- Summarising numeric and categorical variables for each cluster

Results: Cluster Analysis

- Segmentation of the high-risk population identified six clusters

Cluster	1. Complex Mental Health (n=170)	2. Older Complex (n=1372)	3. Younger Asthma (n=5327)	4. Recent Cancer (n=6612)	5. Drug Monitoring (n=6892)	6. Low Utilisation COPD (n=9171)	Total Population (n=29,454)
Age, mean	58.4	72.7	48.3	65.3	65.1	71.2	68
Median (IQR)	58 (50.25-70.00)	76 (66.00 – 84.00)	51 (34.00 – 64.00)	68 (57.00 – 76.00)	69 (57.00 – 77.00)	72 (64.00 – 80.00)	(55-77)
Female, n (%)	105 (61.76)	871 (63.48)	3892 (74.32)	2665 (40.31)	4850 (70.37)	2891 (31.52)	15274 (51.86)
Primary and community care contacts, median (IQR)	10 (5-20)	49 (25-83)	4 (2-9)	4 (2-9)	4 (2-8)	4 (2-9)	5 (2-10)
Mean contacts per 1,000 population	16,006	61,847	6,721	6,808	6,493	6,733	9,313
Secondary care elective consultations and admissions, median (IQR)	6 (2-13)	11 (5-20)	5 (1-11)	16 (8-28)	7 (3-13)	3 (0-7)	6 (2-14)
Mean contacts per 1,000 population	10,288	15,136	7,860	21,332	9,020	4,818	10,561
Cardiovascular condition (current), n (%)	27 (15.88)	607 (44.24)	435 (8.31)	1018 (15.4)	1093 (15.86)	2569 (28.01)	5749 (19.52)
Cancer diagnosed in the past 5 years, n (%)	25 (14.71)	230 (16.76)	358 (6.84)	5813 (87.92)	224 (3.25)	431 (4.7)	7081 (24.04)
Mental Health, n (%)	134 (78.82)	308 (22.45)	1181 (22.55)	766 (11.58)	1067 (15.48)	1516 (16.53)	4972 (16.88)
Asthma, n (%)	54 (31.76)	322 (23.47)	3970 (75.81)	358 (5.41)	507 (7.36)	1358 (14.81)	6,569 (22.3)
COPD, n (%)	78 (45.88)	800 (58.31)	1053 (20.11)	277 (4.19)	404 (5.86)	7677 (83.71)	10,289 (34.93)
Drugs that require monitoring, n (%)	67 (39.41)	243 (17.71)	640 (12.22)	351 (5.31)	5396 (78.29)	403 (4.39)	7100 (24.11)

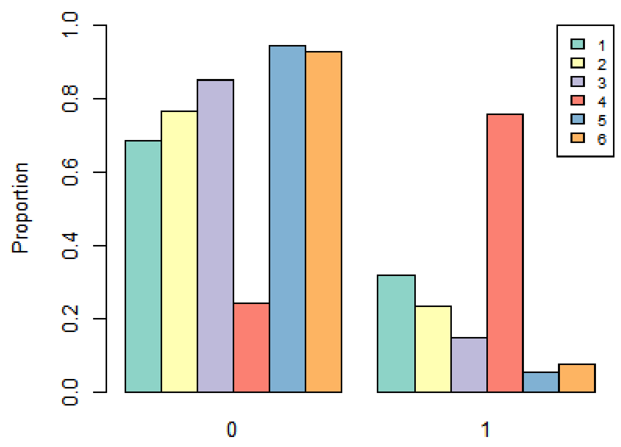
Non-Clustering Variables

Cluster	1. Complex Mental Health (n=170)	2. Older Complex (n=1372)	3. Younger Asthma (n=5327)	4. Recent Cancer (n=6612)	5. Drug Monitoring (n=6892)	6. Low Utilisation COPD (n=9171)	Total Population
Smoking, n (%)	63 (37.06)	211 (15.38)	888 (16.96)	626 (9.47)	679 (9.85)	2317 (25.26)	4784 (16.24)
-Non smoker							
-Current smoker							
IMD decile, median (IQR)	4 (2-7)	5 (3-8)	5 (3-8)	7 (4-9)	7 (4-9)	5 (2-8)	6 (3-8)
Learning disabilities and autism, n (%)	7 (4.12)	20 (1.46)	105 (2.00)	28 (0.42)	32 (0.46)	38 (0.41)	230 (0.78)
Housebound, n (%)	19 (11.18)	328 (23.91)	56 (1.07)	104 (1.57)	181 (2.63)	393 (4.29)	1081 (3.67)
Has a carer, n (%)	10 (5.88)	134 (9.77)	75 (1.43)	137 (2.07)	148 (2.15)	341 (3.72)	845 (2.87)
Is a carer, n (%)	3 (1.76)	68 (4.96)	162 (3.09)	210 (3.18)	271 (3.93)	343 (3.74)	1057 (3.59)
Charlson Score, median (IQR)	3 (2-5)	6 (4-7)	2 (1-4)	5 (3-6)	4 (2-5)	5 (3-6)	4 (3-6)

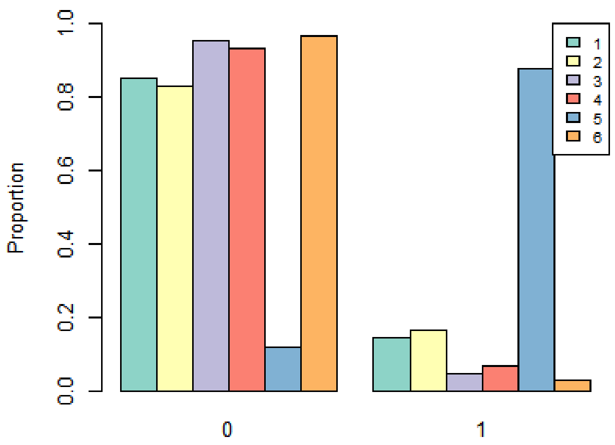
Results: Cluster Analysis

- Cluster profiles can be produced from the package using ‘clprofiles’ command

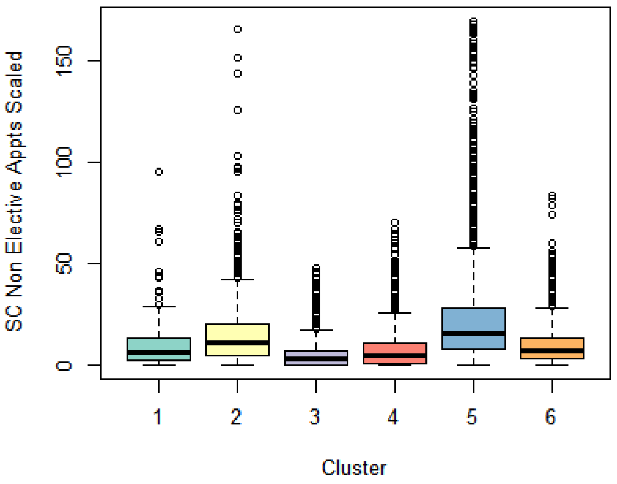
Asthma



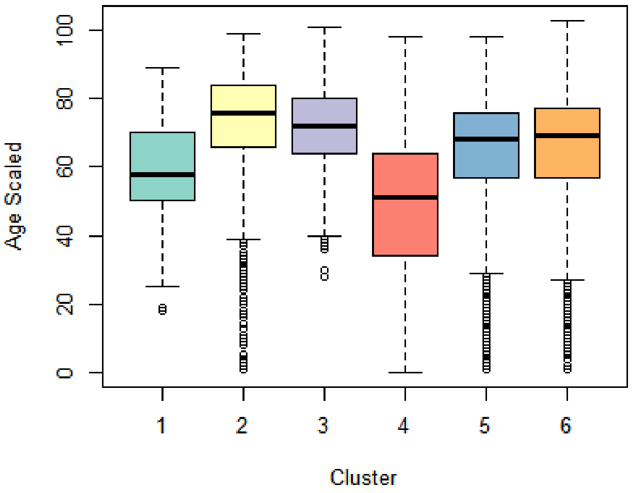
Cancer diagnosed in the last 5 years



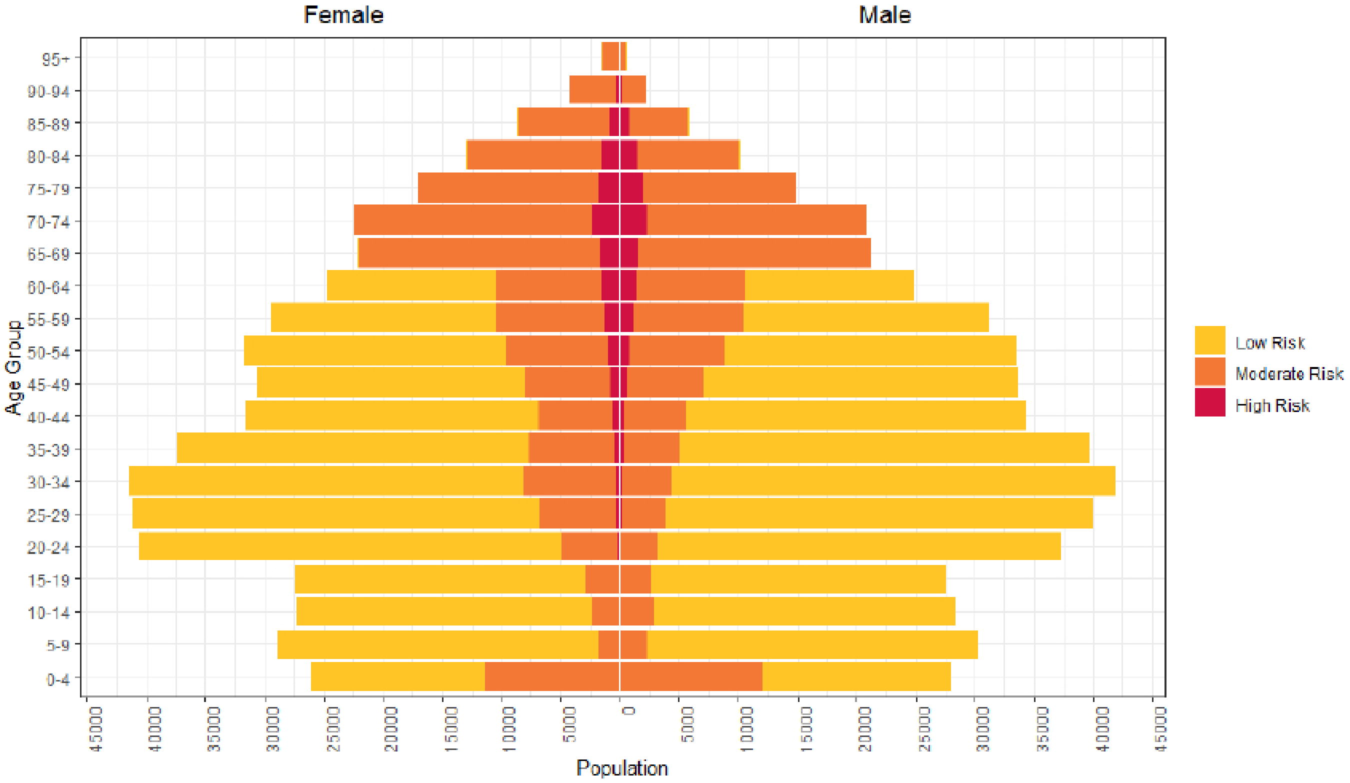
Secondary Care Non Elective Appointments



Age



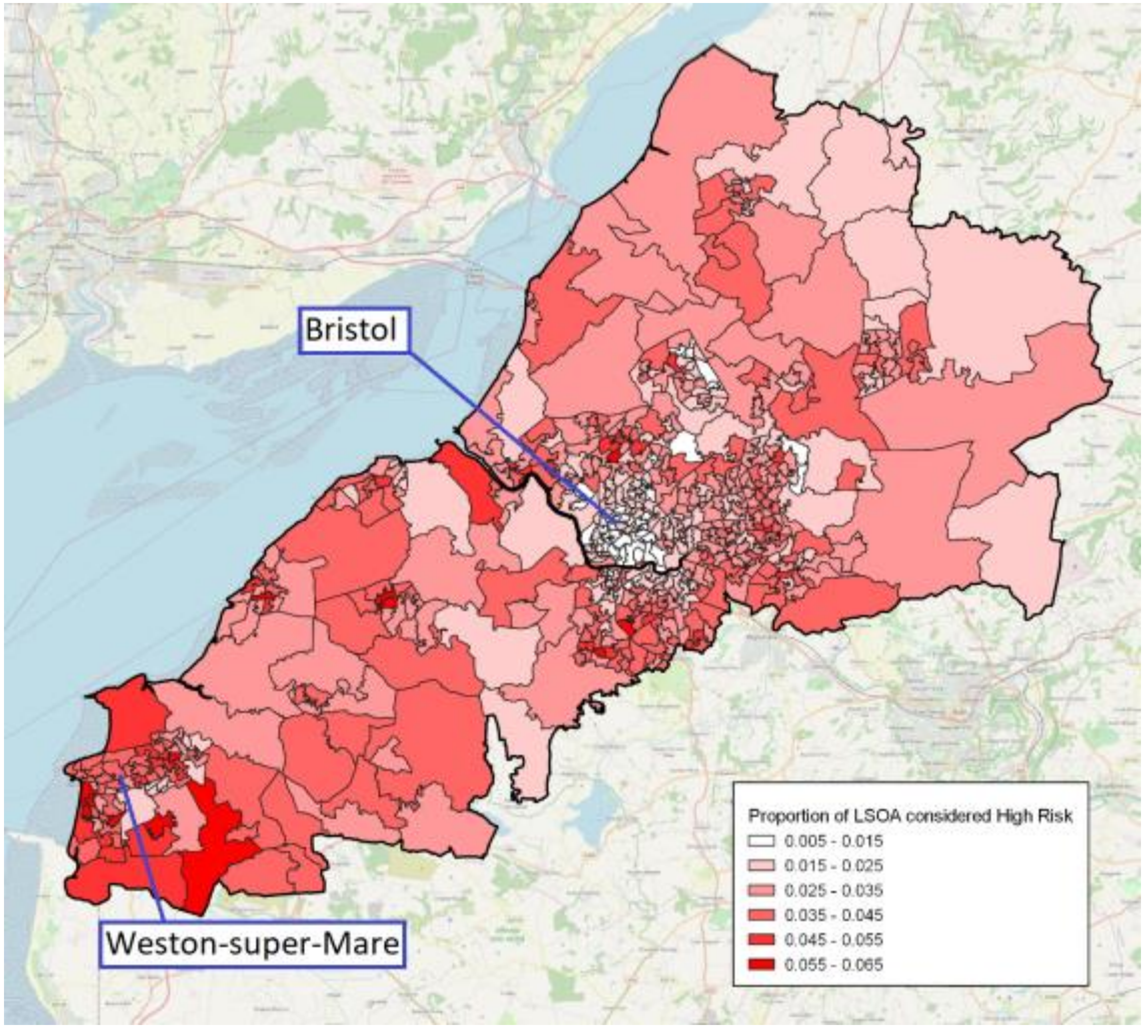
Covid-19 risk populations



Population pyramid showing absolute numbers of the population in 5-year age bands stratified by high-risk (red), moderate risk (orange) and low risk (yellow).

Shaping better health

Where the shielded population R



Geographical map of the Bristol, North Somerset and South Gloucestershire healthcare system illustrating the concentration of high-risk individuals at Lower Super Output Areas (LSOA) level.

Discussion

- Key findings
- Strengths and limitations
- Implications
- Future work
- Conclusions

Selected References Cluster Analysis

- Szepannek G. clustMixType: User-Friendly Clustering of Mixed-Type Data in R. *The R Journal* 2018;10(2)
- Szepannek GA, R. clustMixType: k-Prototypes Clustering for Mixed Variable-Type Data. April 23 2020. *The Comprehensive R Archive Network* 2020
- Huang Z. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery* 1998;2(3):283-304. doi: 10.1023/A:1009769707641
- Chong JL, Lim KK, Matchar DB. Population segmentation based on healthcare needs: a systematic review. *Syst Rev* 2019;8(1):202. doi: 10.1186/s13643-019-1105-6 [published Online First: 2019/08/15]
- Wood RM, Murch BJ, Betteridge RC. A comparison of population segmentation methods. *Operations Research for Health Care* 2019
- Yan S, Kwan YH, Tan CS, et al. A systematic review of the clinical application of data-driven population segmentation analysis. *BMC Med Res Methodol* 2018;18(1):121. doi: 10.1186/s12874-018-0584-9 [published Online First: 2018/11/06]
- Vuik SI, Mayer E, Darzi A. Enhancing risk stratification for use in integrated care: a cluster analysis of high-risk patients in a retrospective cohort study. BMJ Open 2016;6(12):e012903. doi: 10.1136/bmjopen-2016-012903
- Everitt BS, Landau S, Leese M, et al. Cluster analysis. 5th ed. Chichester: John Wiley & Sons 2011.

Thanks for listening

Any Questions?

Contact Details:
charlie.kenward@nhs.net
jenny.cooper10@nhs.net