A comparison of population segmentation methods

R.M. Wood [1, 2], B.J. Murch [2], R.C. Betteridge [2, 3]

[1] School of Management, University of Bath
Claverton Down, Bath, BA2 7AY, United Kingdom

[2] Modelling and Analytics, UK National Health Service
South Plaza, Marlborough St, Bristol, BS1 3NX, United Kingdom

[3] Data Science, Cerner Limited
37 North Wharf Road, London, W2 1AF, United Kingdom

Correspondence to: Dr Richard Wood (richard.wood16@nhs.net; +44 117 900 2510)

Abstract

This paper presents the first comparison of descriptive segmentation methods for population health management. The aim of descriptive segmentation is to identify heterogeneous segments according to some target observed measure. In healthcare it can be used to understand how utilisation is distributed among a population, and to identify the patient attributes which explain the greatest differences (knowledge of which can help shape segment-tailored services). In reviewing a number of segmentation methods that are both employed on the ground and explored more experimentally within the academic literature, this paper aims to open up a range of options allowing clinicians and managers an informed choice on which approach to use for their situation. Results support the recommendation that decision tree approaches are on-the-whole most suitable, being configurable to local data and providing the best inter-segment discrimination. More basic judgemental splits on patient attributes can be powerful, with the count of chronic conditions being a key variable. Prescribed binning methods such as Bridges to Health are unlikely to achieve high levels of discrimination but do have easily interpretable segments and could be useful for benchmarking. Clustering methods are found to lack discriminative power, which can be attributed to a lack of conceptual appropriateness to the problem.

Keywords: *Population Health, Population Segmentation, Decision Trees, Cluster Analysis, Healthcare Utilization*

1. Introduction

Population health has been defined as “the health outcomes of a group of individuals, including the distribution of such outcomes within the group” [1]. Interest in this field has grown in recent years driven by the combination of rising costs of care with increasingly polarised health needs leading to greater inequality in per capita spend and clinical outcome [2]. Facilitated by the rise of big data and the availability of associated analytical methods, this has led to the growth of *population health analytics* as a discipline concerned with quantitatively approaching matters of population health.

One of the principal investigative areas within the field of population health management – and the subject of this paper – is population segmentation. This involves using information about individuals, such as age and sex, to partition a population into similar groups. Ultimately, the aim is to identify meaningful and interpretable population cohorts which are heterogeneous *between* and homogeneous *within*. It is important to have such discrimination since it allows the greatest differences to be uncovered.

Population segmentation is an important tool in healthcare since it allows managers and clinicians to cut through the complexity of large and unwieldy datasets in making sense of the key patient-related attributes that drive the most significant differences in some targeted measure of interest – clinical outcome, waiting time, or utilisation as measured through activity or spend. These insights can help determine the nature and scale of intervention that may be required for cohorts of the population identified as requiring attention. For example, if, as in [3], cohorts representing more deprived sectors of the population are found to be associated with worse clinical outcome, then efforts can be directed at ensuring better accessibility and engagement with those communities.

By targeting activity or spend an alternative interpretation is required since homogeneity should not be sought or expected in these measures. Instead, managers may use an understanding of the patient attributes which lead to significant variation in order to determine how services can be shaped around the respective cohorts in order to most effectively meet the healthcare need [4]. For example, if older people with diabetes are found to have a very high number of emergency admissions then perhaps more proactive community services tailored to this demographic could be introduced to reduce the burden on acute providers. Projections on future numbers of older and diabetic people could also be used to determine the longer-term size of this cohort and thus ensure any interventions are suitably future-proofed. With increasing disparity in health needs driven by people living longer and with multiple chronic conditions, designing and delivering these types of intervention are vital in moving toward system-wide, integrated care that is being promoted in many developed countries [5]. However, to do this first requires having a fit-for-purpose segmentation method in order to appropriately determine the population cohorts.

The focus of this paper is to compare side-by-side a number of descriptive segmentation methods in order to understand how their respective assumptions and construction determines resulting segment composition (i.e. which individuals are assigned to which segments). In order to ensure a consistent comparison, it is necessary to use common input data and to ensure outputs related to segment contents are reported using common metrics. The data used here is from a health system in south west England and comprises multiple patient-related attributes as well as activity and spend data for primary and secondary care. The outcome measure targeted is total spend, due to the importance of being able to identify excessive or inefficient spending in a climate of increasing financial pressures on healthcare delivery (noting that the approach and messages of this paper would apply equally to other targets including clinical outcome measures such as 30-day re-admission risk and performance measures such as waiting time).

Practically, the issues this paper seeks to address are important ones. Indeed the UK's National Health Service (NHS) has recently committed to population segmentation in its long term plan, stating its desire to "become increasingly sophisticated" in such methods [6]. However, a current lack of both specialist knowledge and awareness of suitable methods – some existing only in pockets of the literature – is, in the authors' experience, pushing many healthcare systems to primitive and locally-insensitive approaches. In reviewing a handful of methods that are both employed on the ground and explored more experimentally within the literature, this paper aims to open up a range of options allowing an informed choice on the method used.

The remainder of the paper is structured as follows. The next section introduces the data and reviews the literature. Results are thereafter presented for sixteen selected methods; applying each to the data and comparing side-by-side through a number of summary measures. Finally, a discussion considers the pros and cons of the various methods and whether different approaches are better suited to answering different questions in different settings.

2. Methods

2.1 Data

The data available for this study consists of healthcare utilisation and personal attributes for 51,072 individuals in a health system of south west England. The utilisation data contains primary and secondary care spend, including prescription costs, for each individual over a 12-month period to June 2018. The personal attribute data contains demographic variables (age, sex and deprivation index), social variables (e.g. whether housebound or has a carer), and clinical variables relating to frailty (e.g. fluid retention or difficulty walking) and the presence of chronic conditions (e.g. diabetes or hypertension).

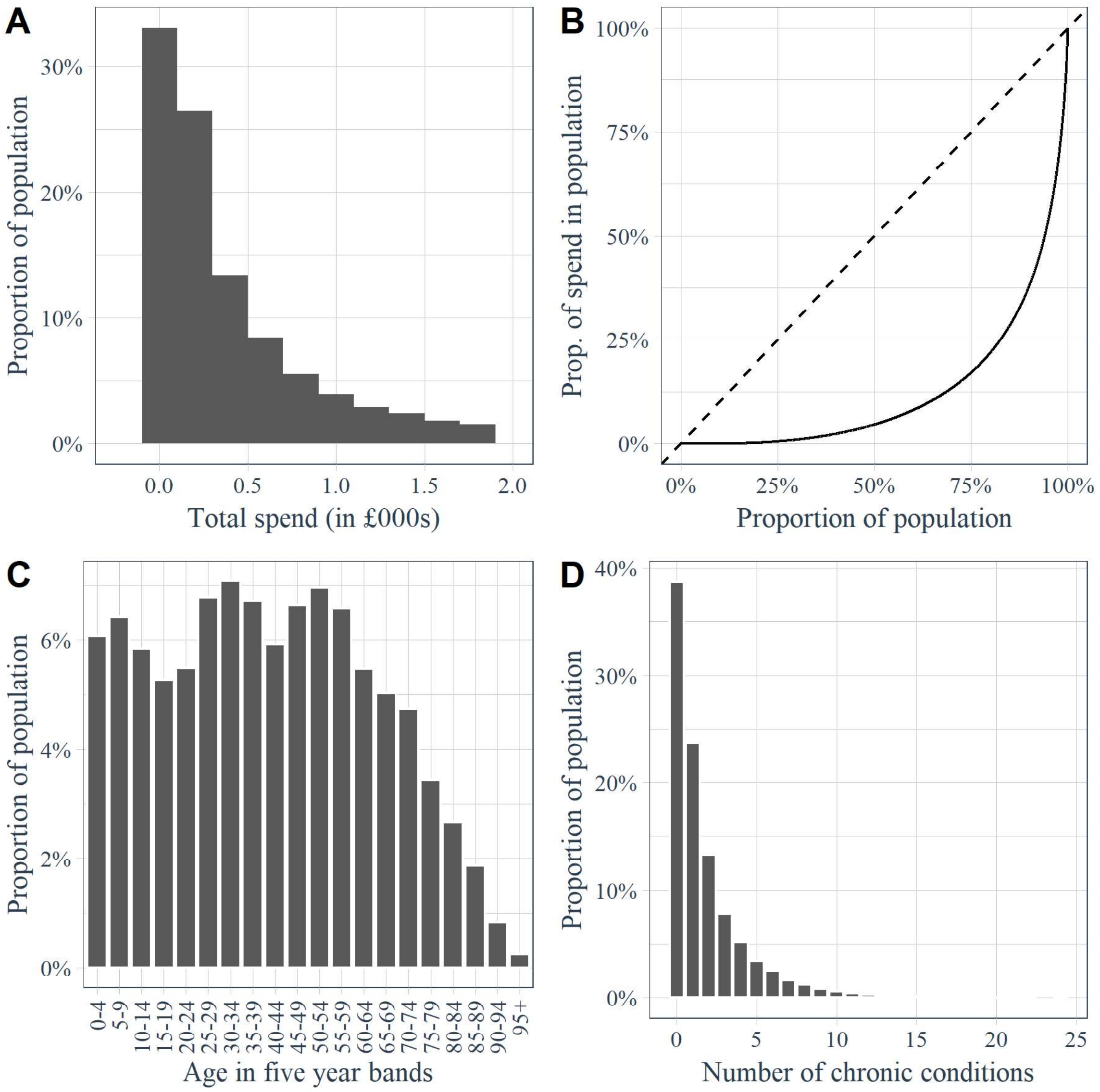


Figure 1 Descriptive summary of the data including A) histogram of spend, B) Lorenz curve for spend, C) histogram of age, D) histogram of chronic condition count

Spend is, as expected, highly right-skewed (skewness 7.89) with a 12 month per capita mean and median of £852 and £240 respectively. The histogram for spends under £2k is plotted in Figure 1A, excluding 5,689 individuals with spend above this threshold (max £64.1k). Figure 1B charts the Lorenz curve, providing evidence of the 80/20 Pareto principle holding since 80% of total spend is generated by 21.3% of the population (a similar observation to other studies, e.g. Bertsimas et al, 2008). Spend inequality is further evidenced by a Gini coefficient of 0.75 and the finding that half of total spend is generated by just under 6% of the population. Figure 1C illustrates the multi-modality of the age distribution and Figure 1D plots chronic condition count, noting that the majority of the population (61.3%) have at least one. Sex is equally represented (50.9% female) and the population is, overall, less deprived than the national population (57.9% of individuals reside in areas ranked in the top half nationally).

2.2 Segmentation methods

To be considered within the comparison of this study, methods must be capable of addressing the outlined aim, i.e. to descriptively partition a population in order to achieve meaningful and interpretable patient cohorts offering credible discrimination in the target measure of healthcare spend. In order to produce meaningful and interpretable patient cohorts, segment membership must be based on patient attributes and thus any methods are excluded whose segments are defined by the target measure (e.g. binning patients according to spend).

Following a review of the literature, four classes of method are determined.

2.2.1 Judgemental splits

This class consists of methods that are most accessible to healthcare practitioners on the ground where cohorts are determined not empirically through available data, but by judgement or the documented experiences of others.

Perhaps the most basic and widely-used approach to segmentation is to make judgemental splits on the values of one or more selected explanatory variables. This is commonplace in many fields, such as marketing where segmentation has become a crucial means of understanding which sectors of the population drive the highest yield and which may be more receptive to advertising. Here, customer age and sex are two of the most typically used explanatory variables where cohorts are defined based on judgemental splits on their values. For example, in investigating hotel guest satisfaction Oh et al [8] partition responses into four segments based on whether male or female and age greater than or less than 55 years. In health, researchers have studied differences in life expectancy by segmenting on explanatory variables such as sex, ethnicity and geographic region [9]. Turning to healthcare utilisation – the topic of this paper – the UK National Health Service recommends that age and condition should be used to segment a population [10], commenting that this is “the preferred approach of many international integrated care providers”.

While this class of approach is simple, it is *non-objective* in the sense that it has not objectively sought discrimination in the target variable. For instance, Oh et al [8] claim their choice of criteria in segmenting guest reviews was “intuitive”, but there may well be a more appropriate split on age which could have led to greater between-segment heterogeneity. There may also have been another variable which, included in the segmentation criteria, could have explained more of the variation in review scores (customer nationality perhaps). On the other hand, there could end up being too many segments. If, for example, suggestions from NHS England [10] are followed then, with the five age splits and three condition splits advised, there would be fifteen segments. Not only could this many be unwieldy in terms of assigning a sensible and relatable meaning to each one, but it would also be questionable whether sufficient discrimination could be afforded among them.

2.2.2 Prescribed binning criteria

Another descriptive segmentation approach commonly used in health services is that here referred to as a prescribed binning criteria method. This involves using an off-the-shelf set of binning rules which channel patients into one of a set of pre-defined cohorts based on their attributes. A popular choice is the Bridges to Health approach described in [11]. This defines eight groups based on expert opinion of distinct healthcare needs. However, as raised in [12], the approach does not clearly articulate the attributes which define membership of each cohort. Instead, Low et al [12] propose a more definitive criterion for their six expert-defined segments. Joynt et al [13] also address this matter, using Bridges to Health as a foundation for their six-segment method. The advantage of these approaches is that they produce a manageable number of segments which, with names such as *Relatively healthy* and *Frail elderly*, have meaningful connotations. The drawback is that they are non-objective, i.e. using expert opinion offers no guarantee of providing good discrimination.

The electronic frailty index (EFI) – a scale from zero to one from which individuals can be segmented into four ordinal cohorts – does, on the other hand, provide an objective approach since its development has statistically targeted treatment need [14]. However, as with the other off-the-shelf methods, there is no guarantee of maintaining any such discrimination when applied to a different demographic in a different region at a different time. This can only be achieved with an objective approach that has been developed and calibrated locally, i.e. one that has been *derived*.

2.2.3 Decision trees

Decision trees are a classification tool whereby following a series of conditional statements leads to one of a number of mutually exclusive outcomes. They are referred to as trees since the successive decision points, each recursively sub-setting the population through splits on the explanatory variables, propagate like branches in a tree.

Statistical methods to learn decision trees can offer an *objective* and *derived* approach to segmentation. The most common choice is through the Classification and Regression Tree (CART) algorithm introduced by Breiman [15]. This established non-parametric method works by binary recursive partitioning, where at each decision point a two-way split on a chosen explanatory variable is performed. Selection of the split, as well as the variable itself (which can be either continuous or categorical) is motivated by maximising homogeneity in the branched nodes (specifically, through Gini impurity). Starting with the whole population, this recursive partitioning is repeated until some pre-specified terminating conditions are met (e.g. depth of tree no more than *n* levels and/or no less than *m* observations in any segment). While this can lead to a discriminative and informative segmentation, it is, in the authors' experience, seldom used on the ground in the UK – presumably due to a lack of analytical capacity [16].

However, in the academic literature there are numerous examples of CART analysis in use. In the healthcare setting, investigators have employed the method to understand population segments that best explain differences in hospital length of stay [17], immunisation take-up [18], and mortality given disease [19] and infection [20]. With regards to segmentation on utilisation, Cairney et al [21] find that splits based on age, sex, marital status, and income best explain the variance in mental health activity. Paediatric spend is targeted in [22], where respiratory condition and previous surgery are found to be explanatory. CART analysis has also been performed to segment claims data on healthcare spend from American insurance companies, such as in [7,23,24].

While CART provides a good fit to healthcare segmentation problem, there are other potentially appropriate methods that have not received the same attention. For instance, conditional inference trees work in a very similar fashion to CART, but – in an attempt to reduce variable selection bias – they perform permutation significance tests at each decision point as opposed to minimising Gini impurity [25]. Information gain is instead chosen to perform splits in the C4.5/5.0 algorithm, albeit with the restriction that the target variable is categorical [26]. And the CHAID method – which uses

Chi-square *p*-value significance testing – extends this restriction to explanatory variables, while relaxing the requirement of binary splits [27]. More recently, tree-based ensemble methods, such as random forests and boosting, have been used to combat over-fitting and bias in studies estimating healthcare spend [23,24]. However, it is important to bear in mind that while these ensemble approaches can be useful in *prediction*, they are not suitable for *classification* (i.e. segmentation) since a final tree is not output which can be used to descriptively group individuals.

2.2.4 Cluster analysis

Clustering concerns the grouping of data points into similar cohorts, referred to as clusters. It represents a class of methods which are *unsupervised*, meaning there is no response variable, only a number of explanatory features that relate to each of the data points (observations). In the context of descriptive population segmentation there is thus no variable – such as spend – which is targeted; although the target variable can be included alongside the other features for which homogeneity is sought. Thus, in this respect, it is different to decision trees which are *objective* by design. With this in mind, it is therefore arguable at a conceptual level whether cluster analysis is an appropriate technique in meeting the afore-mentioned aim of this study. However, given previous efforts in this field alongside current interest in the art of the possible of machine learning approaches, it is thought important that clustering methods should be investigated within the comparison of this study.

While *non-objective*, clustering methods do offer a *derived* approach for population segmentation, configurable to local data. A commonly-used clustering method is *k*-means. In simple terms, this works by first randomly assigning a number between 1 and *k* to each observation and then attempting to reduce the mean distance from each labelled observation to the *k* cluster centroids through iteratively reassigning the label observations to the closest centroids. The method has been used for population segmentation in a variety of settings. In seeking to understand which patient attributes best distinguish between Canadian rehabilitation service users, Armstrong et al [28] use *k*-means to uncover seven types of patient based on age, sex, cognition and functional impairment. In the renal service setting, Liao et al [29] instead cluster on incurred cost profiles, thereafter inspecting patient attributes and attempting to assign interpretable names for each cluster. Vuik et al [30] also cluster on cost profiles, using *k*-means to deduce eight types of patient for which utilisation of primary and secondary care services in the UK is materially different. Low et al [31] investigate a similar problem for Singapore, finding five homogeneous groups. However, they also include a patient attribute (age) alongside the service-level primary and secondary utilisation features used for the clustering.

It should be noted that a requirement for *k*-means, and centroid-based clustering more generally, is to define the value of *k* upfront. Both [30,31] determine this through employing hierarchical clustering to samples of their data. Hierarchical methods do not require an upfront value of *k* and instead deduce this through either a bottom-up (agglomerative) or top-down (divisive) approach. These are computationally expensive methods which are impractical with larger datasets; thus their application to samples of the data in the two mentioned studies. On smaller health datasets it has, however, been applied in full, e.g. in [29] (who also implement *k*-means, favouring this on the grounds of relative segment interpretability).

Moving away from the healthcare segmentation literature, one can find a variety of other clustering methods available for use. *k*-modes, for example, is an analogue of *k*-means where all features are required to be categorical (as opposed to continuous-valued under *k*-means). There is *k*-prototypes, which is a technique accommodating both numeric and categorical data (using Euclidean distance for the former and dissimilarity measures for the latter, this method essentially combines the *k*-means and *k*-modes algorithms). And there is also the *k*-medoids method, implemented through algorithms such as partitioning around medoids (PAM), which is a more robust alternative to *k*-means through its use of total sum of square distance as opposed to mean distance in the calculation of centroids. This makes it less sensitive to outliers but at the expense of increased computational requirements (this is discussed further in Section 4).

2.3 Implementation

Table 1 outlines the sixteen methods evaluated in this study, including details on implementation which has been performed on a standard desktop computer with the statistical software *R*.

For the *derived* decision tree and clustering methods the number of segments has been fixed at six. This is done in order to support consistent comparison (noting discrimination should always increase with segment number) with six being chosen as a balance between adequate discrimination and interpretable segment naming, as well as being the average number of segments used within previous healthcare segmentation efforts (Table 2). A hard constraint governing the calibration of both the decision tree and clustering methods is that no less than 0.5% (256) of observations are contained within any segment. Note that this is lower than the 1% used in [31], owing to the importance here attached to being able to effectively capture the small groups of very high spend patients that are known to exist (Figure 1B).

For the methods which are not able to handle mixed data types some pre-processing is required. For those accepting categorical data only (CHAID, *k*-modes) the continuous-valued variables are converted to ordinal categories. For those best-suited to numeric data (*k*-means, *k*-medoids) the technique factor analysis of mixed data (FAMD) is used to transform the categorical variables into a set of continuous-valued components which can thereafter be used with Euclidean distance measures for clustering (note that while *k*-means and *k*-medoids can accommodate mixed data using dissimilarity measures such as Gower distance, this is found to be computationally impractical).

Table 1 Summary of segmentation methods used in this study

Class	**Ref**	**Method**	**Implementation**
Judgemental splits	1	Age	0-12, 13-17, 18-49, 50-74, 75+ (based on [10])
	2	Sex	Male/Female
	3	Chronic conditions	0-4, 5-9, 10-14, 15+
	4	Age and chronic conditions	9 groups: age 0-17, 18-64, 65+ by CCs 0-4, 5-10, 11+
Prescribed binning criteria	5	Lynn et al, 2007 (Bridges to Health)	See Supplementary Material A for interpreted binning criteria
	6	Low et al, 2017	
	7	Joynt et al, 2017	
	8	Electronic Frailty Index	Four prescribed groups (see [14])
Decision trees	9	CART	[15] method (*R:rpart*)
	10	Conditional inference	[25] method (*R:partykit*)
	11	C5.0	[26] method (*R:C50*)
	12	CHAID	[27] method (*R:chaid*) Pre-processing: binning of numeric vars
Cluster analysis	13	*k*-means	[32] method (*R:stats*) Pre-processing: FAMD (*R:FactoMineR*)
	14	*k*-modes	[33] method (*R:klaR*) Pre-processing: binning of numeric vars
	15	*k*-prototypes	[34] method (*R:clustMixType*)
	16	*k*-medoids (PAM)	[35] method (*R:cluster*) Pre-processing: FAMD (*R:FactoMineR*)

Table 2 Number of segments used in previous population segmentation studies

Class	Reference	Number of segments
Prescribed binning criteria	[11]	8
	[12]	6
	[13]	6
Decision trees	[7]	*Not specified*
	[14]	4
	[21]	6
	[22]	5
	[23]	*Not specified*
	[24]	*Not specified*
Cluster analysis	[7]	*Not specified*
	[28]	7
	[29]	4
	[30]	8
	[31]	5

3. Results

3.1 Descriptive summary

The empirical cumulative distribution functions for segment-level spend are presented for each of the sixteen implemented methods in Figure 2. From these it is possible to understand information on i) the number of segments, ii) the profile of spend within each segment, and iii) the number of patients within each segment.

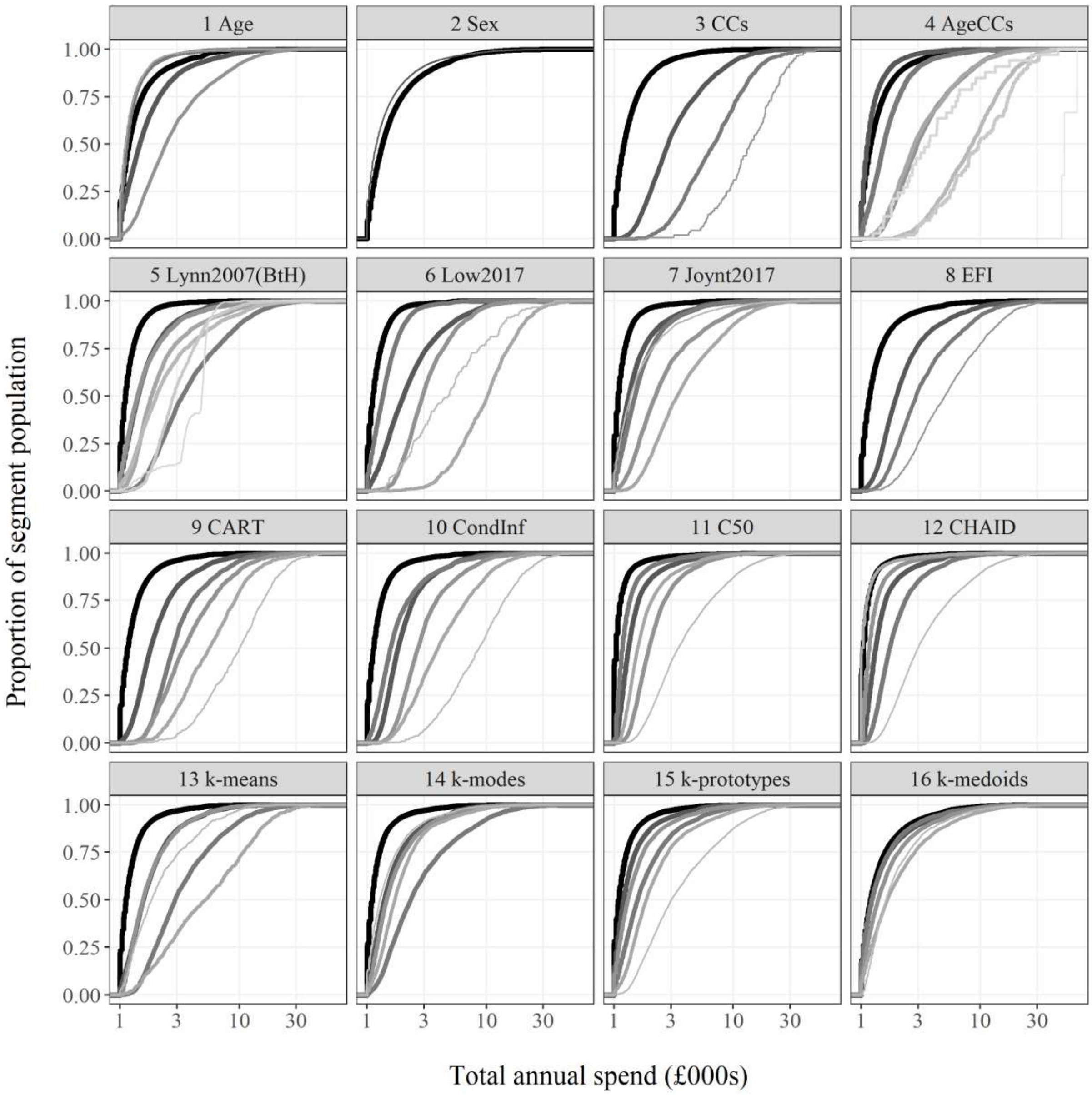


Figure 2 Segment-level distribution functions for spend (note: line shading is used to represent the order of segment size with darker and thicker lines representing larger segments)

There are a number of rules of thumb which can be used in interpreting the plots. First, the closer the lines are, the less discrimination in spend has been achieved. For example, the lines are very close together for the judgemental split on sex (method 2) indicating the spend profiles for males and females are quite similar. The issue of discrimination is addressed further in Section 3.3.

A second rule of thumb is based upon line consistencies, where those with greater shape volatility indicate failure to achieve homogeneity in segment spend. For example, see the light grey line towards the lower-right corner of Bridges to Health (method 5). It may be that this segment, accounting for women with pregnancies and/or terminations (see the *Maternity* segment in Supplementary Material A.1), is conflating patients with essentially different spend profiles and is therefore conceptually inappropriate. It could also be that, with a size of only 314, there are simply too few individuals within this group. This leads on to the third rule of thumb: that the smoother the line, the greater the number of observations. Consider, for example, the right-most line of the judgemental split on age and chronic condition (method 4). The steps between the observations are clearly visible. In fact, this segment, including those up to 17 years with 11 or more chronic conditions, only contains three individuals.

The final rule of thumb relates to the order of the line shading when moving left to right. When there is consistent progression from dark to light, e.g. Electronic Frailty Index (method 8) and CART (method 9), this indicates the prescence of segments which get progressively smaller and more expensive. On the other hand, it can be seen that this does not hold when segmenting by age (method 1). Here the largest segment (containing 41% of observations) is actually ranked third of the five segments in terms of highest spend.

Figure 3 provides further description of the segment-level method outputs for four selected methods (complete results in Supplementary Material B). For the target measure of spend this includes information on the proportions of population and spend within each segment. In the case of perfect spend equality these measures would be equivalent. However – as is known from Figure 1B – this is not the case for the population considered here. Although there are considerable differences in the ratio between these proportions. See, for example Bridges to Health, where in the *Chronic conditions* segment 30% of the population consume a comparable 28% of the spend. Conversely in the *Frailty* segment just 7% of the population consume 31% of the spend. Detail on the proportion of spend according to point of delivery is also included in Figure 3. This illustrates some quite stark differences, e.g. note the substanitally greater proportion of non-elective admission spend in the *15+* segment against the *0-4* segment for the judgemental split on chronic condition count.

Also included in Figure 3 is information on the distribution of some selected explanatory variables relating to patient attributes. The box and whisker plots are useful in understanding any differences in age profile between segments, e.g. take the *Healthy* and *Frail* segments of the *k*-means method. Note, for this method, the plots for *Mid-life* and *Endomitriosis* appear quite similar. However, inspecting the *% male* field indicates that the latter segment is entirely female. This segment can, in fact, be seen as a more costly subset of the former which has been split out due to the high incidence of this and related female-specific conditions. An interesting property is evident for judgemental split on chronic condition count and CART where segment-level age increases with reducing segment size – this is clearly not present for the other methods. Finally, it can be seen from the distribution of chronic condition count that there is some correlation between this variable and per capita spend. This is to expectation and confirmed with an overall Spearman's rho of 0.65 between these variables.

The segment names for the derived methods contained in Figure 3 have been given through an inspection of the segment-level patient attributes. For the CART method, the resulting decision tree can also be conveniently inspected to understand explicitly the explanatory variable bounds (Figure 4A). Naming the large *Healthy* segment is trivial. The *Mid-life* segment has been named due to a mean age of 57 years (IQR 26). The substantial number of unique prescription types (10 and over) has yielded the *Prescriptions* segment. The *Older ill* segment name has been arrived at due to a mean age of 72 years and the presence of illnesses such as chronic kidney disease (42%) and depression (39%). Finally, the *Frail* and *Very frail* segments are named due to having over three quarters with frailty graded as at least moderate according to the EFI measure. The latter segment can be seen as a more severe frailty subset, containing individuals with 12 or more chronic conditions.

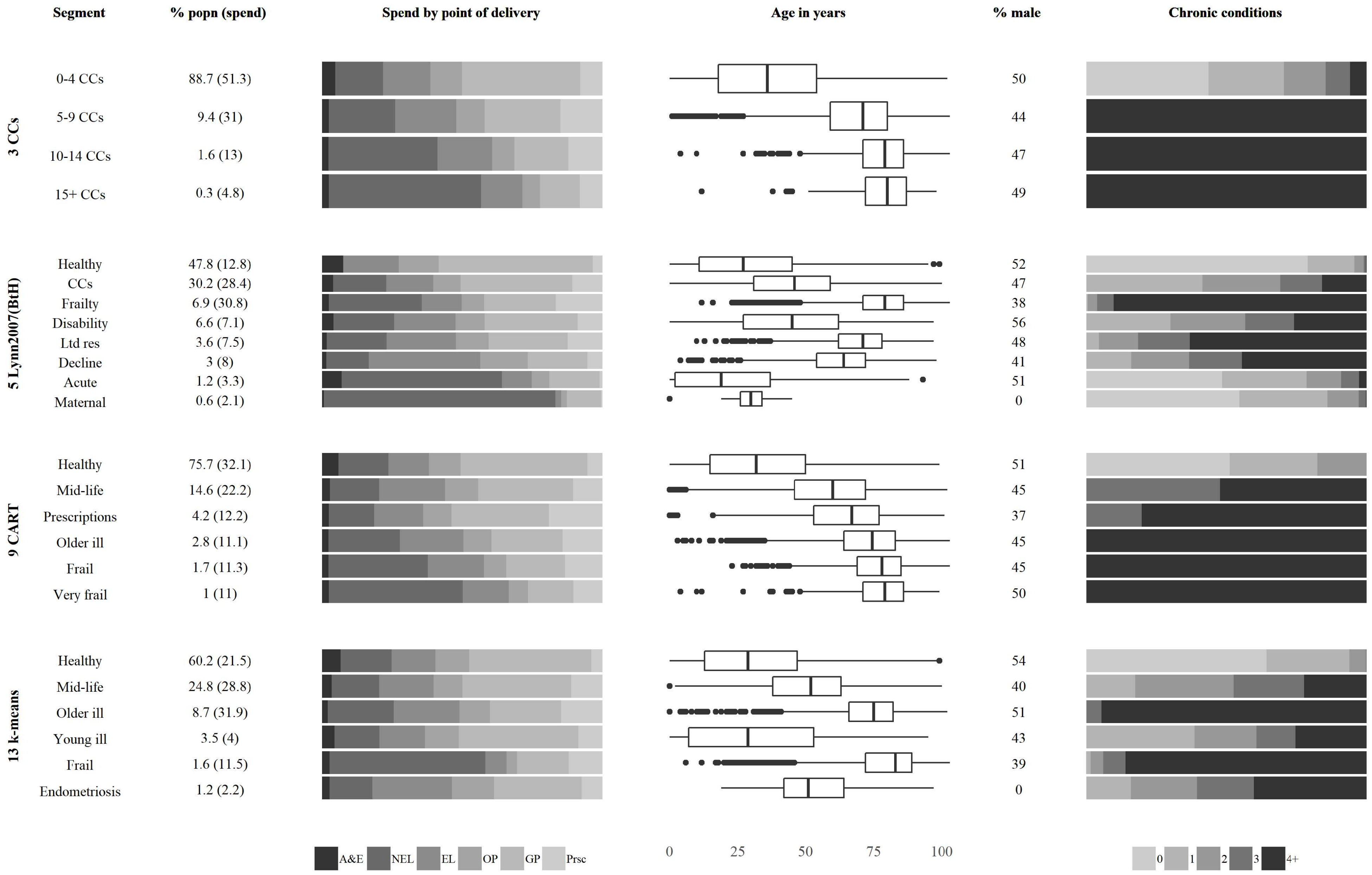


Figure 3 Descriptive summaries of selected segmentation methods (note: segments are ordered by decreasing size for each method; see Supplementary Material B for a description of all sixteen methods)

Figure 4B is known as a *treeplot*, which can be useful, alongside the decision tree in Figure 4A, in readily understanding the size of and spend within deduced segments. Here the size of the blocks represents the number of observations within the segment, while the shading represents the scale of average spend. These can be useful visual aids in summarising segmentations to clinicians and managers.

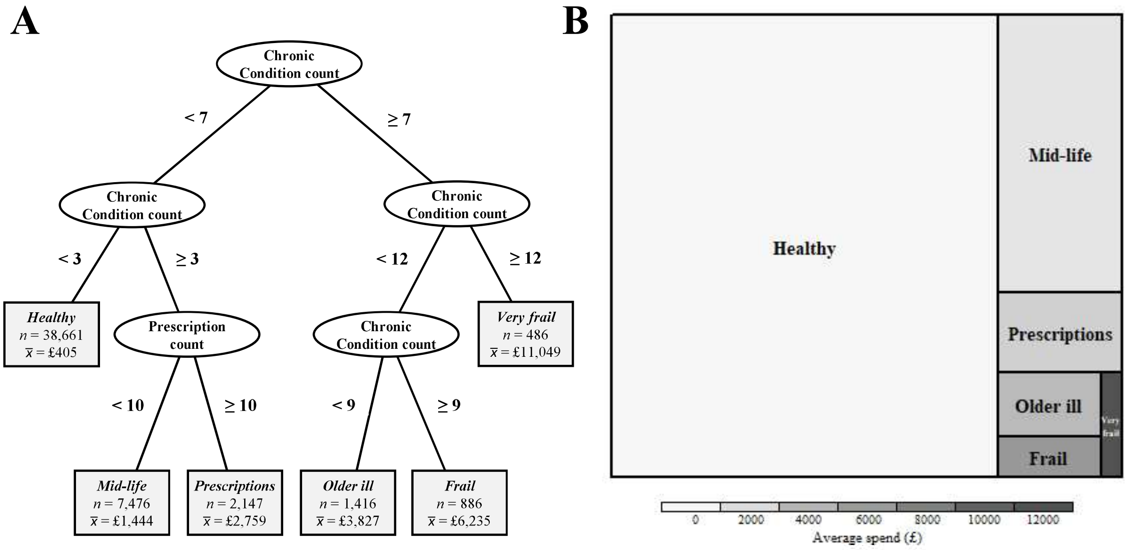


Figure 4 CART method implementation, showing A) selected branching variables, and B) corresponding treemap

3.2 Composition similarity

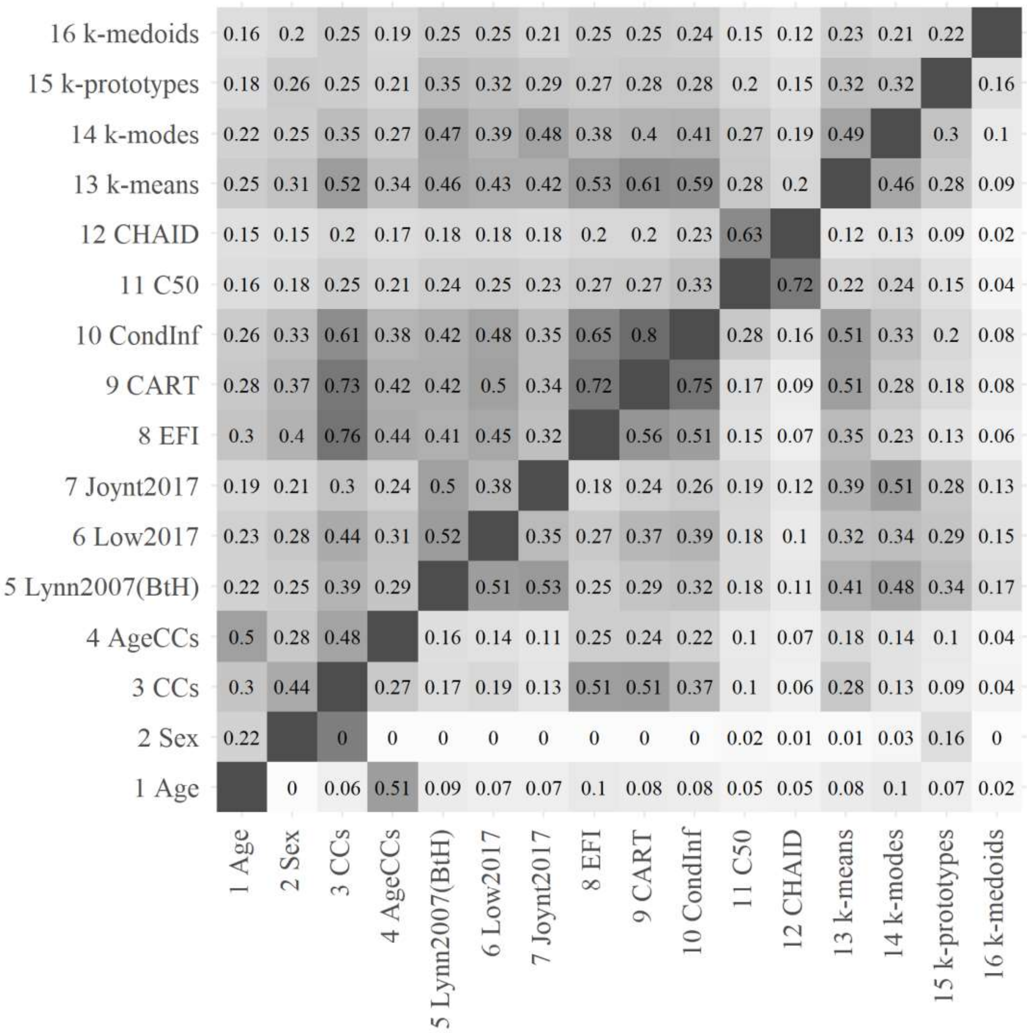
The Jaccard and Rand indices are established measures for assessing the similarity between two segmentations of the same data [36]. Both output a value between zero and one, where zero indicates no agreement between the segmentations and one indicates perfect agreement (i.e. the respective segments contain exactly the same observations).

While both measures are based on counting pairs of observations that are similarly classified, there is a subtle difference in their calculation. With X and Y denoting the two segmentations and defining

A = *# pairs that appear together in the same segment in X and in the same segment in Y*
B = *# pairs that appear in different segments in X and in different segments in Y*
C = *# pairs that are in the same segment in X and in different segments in Y*
D = *# pairs that are in different segments in X and in the same segment in Y*

the Jaccard index is calculated $\frac{A}{A+C+D}$ and the Rand index is calculated $\frac{A+B}{A+B+C+D}$. Note that the adjusted Rand index is used here in order to account for the issue of attaining a value greater than zero based on chance [37].

Pairwise results for both indices (Figure 5) suggest quite substantial differences between the methods, although there is some evidence of intra-class similarity. This is illustrated by the relatively darker shadings of boxes along the diagonal for methods 5-7 (prescribed binning criteria), 9-10 and 11-12 (decision trees) and 13-16 (cluster analysis). Within classes the greatest similarity is between the CART and conditional inference decision tree methods, scoring 0.75 and 0.8 on the Jaccard and Rand indices respectively. Between classes, there is good agreement between the judgemental split on chronic condition count and CART – although this is to be expected given that chronic condition count has been repeatedly selected through the CART branching criteria (Figure 4A). Such findings are supported by the descriptive summaries of Supplementary Material B. Note that overall there is good agreement between the Jaccard and Rand indices, where a Spearman's correlation coefficient of 0.76 is found between the ranked values of each measure.



	1 Age	2 Sex	3 CCs	4 AgeCCs	5 Lynn2007(BtH)	6 Low2017	7 Joynt2017	8 EFI	9 CART	10 CondInf	11 C50	12 CHAID	13 k-means	14 k-modes	15 k-prototypes	16 k-medoids
16 k-medoids	0.16	0.2	0.25	0.19	0.25	0.25	0.21	0.25	0.25	0.24	0.15	0.12	0.23	0.21	0.22	
15 k-prototypes	0.18	0.26	0.25	0.21	0.35	0.32	0.29	0.27	0.28	0.28	0.2	0.15	0.32	0.32		0.16
14 k-modes	0.22	0.25	0.35	0.27	0.47	0.39	0.48	0.38	0.4	0.41	0.27	0.19	0.49		0.3	0.1
13 k-means	0.25	0.31	0.52	0.34	0.46	0.43	0.42	0.53	0.61	0.59	0.28	0.2		0.46	0.28	0.09
12 CHAID	0.15	0.15	0.2	0.17	0.18	0.18	0.18	0.2	0.2	0.23	0.63		0.12	0.13	0.09	0.02
11 C50	0.16	0.18	0.25	0.21	0.24	0.25	0.23	0.27	0.27	0.33		0.72	0.22	0.24	0.15	0.04
10 CondInf	0.26	0.33	0.61	0.38	0.42	0.48	0.35	0.65	0.8		0.28	0.16	0.51	0.33	0.2	0.08
9 CART	0.28	0.37	0.73	0.42	0.42	0.5	0.34	0.72		0.75	0.17	0.09	0.51	0.28	0.18	0.08
8 EFI	0.3	0.4	0.76	0.44	0.41	0.45	0.32		0.56	0.51	0.15	0.07	0.35	0.23	0.13	0.06
7 Joynt2017	0.19	0.21	0.3	0.24	0.5	0.38		0.18	0.24	0.26	0.19	0.12	0.39	0.51	0.28	0.13
6 Low2017	0.23	0.28	0.44	0.31	0.52		0.35	0.27	0.37	0.39	0.18	0.1	0.32	0.34	0.29	0.15
5 Lynn2007(BtH)	0.22	0.25	0.39	0.29		0.51	0.53	0.25	0.29	0.32	0.18	0.11	0.41	0.48	0.34	0.17
4 AgeCCs	0.5	0.28	0.48		0.16	0.14	0.11	0.25	0.24	0.22	0.1	0.07	0.18	0.14	0.1	0.04
3 CCs	0.3	0.44		0.27	0.17	0.19	0.13	0.51	0.51	0.37	0.1	0.06	0.28	0.13	0.09	0.04
2 Sex	0.22		0	0	0	0	0	0	0	0	0.02	0.01	0.01	0.03	0.16	0
1 Age		0	0.06	0.51	0.09	0.07	0.07	0.1	0.08	0.08	0.05	0.05	0.08	0.1	0.07	0.02

Figure 5 Jaccard (upper triangle) and Rand (lower triangle) indices for pairwise combinations of segmentation methods

3.3 Discrimination

It is important that methods achieve sufficient discrimination since this allows the greatest differences in spend to be uncovered. Here weighted variance reduction is used to assess the discrimination achieved through the various segmentation methods. This is a commonly-used measure with the advantages that it is straightforward and can be applied consistently across all considered methods (noting that other measures, such as information gain, can be applied to decision trees but not cluster analysis).

The weighted variance can be calculated for a segmentation through $\sigma_W^2 = \frac{\sum_{i=1}^{N} n_i \sigma_i^2}{\sum_{i=1}^{N} n_i}$, where n_i and σ_i^2 are respectively the number of observations and variance of spend in segment $i \in \{1, \ldots, N\}$. Letting σ_P^2 equal the original, population variance, then the reduction in variance can be calculated as $\frac{\sigma_P^2 - \sigma_W^2}{\sigma_P^2}$, which can be expressed as a percentage.

The reduction in variance for each method is presented in Table 3. These are supported by the spend distribution plots of Figure 2, where greater reductions in variance are consistent with greater separation of distribution curves (i.e. first rule of thumb in Section 3.1). Note, for example, the separation in the curves for *k*-medoids (achieving 2%) and conditional inference tree (achieving 37%). Table 3 shows the prescribed binning criteria of [12] is the most effective at reducing the overall weighted variance. However, it should be noted that this is one of the only implemented methods – in addition to Bridges to Health – that uses activity itself as an explanatory variable of spend. Thus, with such circularity, it should be of little surprise that the method performs so well (for this reason, activity has not been included as a candidate explanatory variable used in any decision tree or cluster analysis implemented here).

As in the segmentation studies of [30,31], hypothesis testing has been performed in order to assess whether there are significant differences between the segments derived through each method. The conditions for using a one-way ANOVA test were violated (due to a lack of normality in the residuals of spend), and so the non-parametric alternative Kruskal-Wallis test was instead used, where the null hypothesis of no significant difference could be rejected (p-value < 0.001) for each of the sixteen methods. A Wilcoxon rank-sum test was used to further inspect the similarities between combinations of segments under each method (Supplementary Material C), showing that the methods, in general, create segments with distinct distributions.

Table 3 Reduction in variance achieved through each segmentation method (* indicates methods in which activity is used as an explanatory variable)

Category	Ref	Method	Variance	Reduction
Baseline	-	Baseline	5,433,321	-
Judgemental	1	Age	4,912,301	10%
	2	Sex	5,418,924	0%
	3	Chronic conditions	3,516,112	35%
	4	Age and chronic conditions	3,505,768	35%
Prescribed binning	5	[11] (Bridges to Health) *	4,275,849	21%
	6	[12] *	3,145,663	42%
	7	[13]	4,376,765	19%
	8	[14] (Electronic Frailty Index)	4,319,655	20%
Decision trees	9	CART	3,352,816	38%
	10	Conditional inference	3,403,747	37%
	11	C5.0	4,017,665	26%
	12	CHAID	4,180,762	23%
Clustering	13	*k*-means	4,088,182	25%
	14	*k*-modes	4,647,659	14%
	15	*k*-prototypes	4,564,590	16%
	16	*k*-medoids	5,331,346	2%

4. Discussion

This paper contributes to the literature by being the first published comparison of descriptive methods which can be used for utilisation-based population segmentation. While other investigators have shown interest in this field, their efforts have been limited to the calibration of individual methods [21,22,31] and those for which segment membership is based on spend and not patient attributes [29,30]. Lacking a side-by-side comparison with consistency in input data and output measures, this has posed an issue for healthcare practitioners keen to understand which method may be best-suited to addressing their population segmentation problem.

Of the comparison studies that have made use of segmentation methods, these have been with the aim of *predicting* cost (e.g. [7,23,24]) which, although useful for estimating individuals' future claims or pricing their insurance policies, is not suited for *describing* the population more thematically. Additionally, these papers select and appraise their methods on the basis of statistical accuracy, whereas in this study a broader variety in class of approach is considered: from those used routinely on the ground in the health service, e.g. judgemental splits and off-the-shelf methods like Bridges to Health; to statistical techniques established in other fields like decision tree learning; and on to machine learning methods such as cluster analysis, which have only recently been applied to population health management.

4.1 Key findings from this study

While better discrimination can be achieved through *derived* methods, it is of surprise just how well judgemental splits perform (Table 3). Splitting by number of chronic conditions achieves a reduction in variance (35%) only narrowly surpassed by the best-performing derived approach, CART decision tree (38%). When inspecting the branching criteria for the CART tree (Figure 4A) it is clear to see why: four of the five branching criteria are based on the variable *number of chronic conditions*.

It is interesting that this variable is so powerful in explaining differences in spend. Indeed, when age is included alongside chronic condition count (method 4) there is no improvement in discrimination. Age is commonly touted as a variable to segment on (e.g. [10]) with some studies including it as the sole explanatory variable in their segmentation approach [31], declaring "age is an important health determinant with profound implications on healthcare needs". While this current study does not refute the importance of age, it does suggest that overall healthiness (as measured by chronic condition count) could be a better indicator of spend. While this is correlated with age it is not perfect (Spearman's rho = 0.57) with there being older persons who are healthy with low spend and younger persons who are unhealthy with high spend.

With regards to the prescribed binning methods, the circularity of basing segment membership on activity has already been raised (Section 3.3) as the reason for why Low et al [12] achieves such high discrimination (42%). The tautology within their concluding remarks "we found that patients in the *complex chronic disease with frequent hospital admissions* segment accounted for the highest hospital admissions" is a compelling reason to avoid circular inclusion of utilisation-related explanatory variables where the considered target measure is itself utilisation. Interestingly, the other off-the-shelf methods, such as Bridges to Health, performed considerably poorer (19-21%) when compared to the basic judgemental split on chronic condition count, raising the question of whether the additional complexities associated with these approaches are worthwhile.

This paper also queries the suitability of cluster analysis. While the authors harboured concerns early-on around conceptual appropriateness to the problem, the recent and growing interest around such methods in this domain and more widely supported their inclusion within this comparison study. However, with performance ranging from just 2 to 25% these concerns were confirmed: derived approaches must be *objective*.

4.2 Considerations for practical use

Achieving good discrimination is important for utilisation-based population segmentation, but there should be other considerations affecting practical use, such as segment interpretability. For instance, while the CHAID decision tree (method 12) achieves a four percentage point better reduction in variance than under the Joynt et al [13] binning (method 7), it may be that the latter is favoured due to the accessible interpretability of its segment names, such as *Healthy* and *Acutely ill*. Whereas under the former some thought could be required to manually deduce what kind of patient each represents from the branching criteria and an inspection of the groups (noting this task is more difficult for clustering methods where segment composition is based on combinations of explanatory variable values).

Understanding the implementation complexities of the various methods is also important. For example, implementing a judgemental split on, e.g., age or sex is fairly trivial, whereas implementing a clustering approach based on FAMD pre-processing would undoubtedly require a specialist skillset (which are often lacking in healthcare [16]). Some effort may also be required in implementing off-the-shelf methods, where published segment descriptions are not always sufficient in determining a set of attribute binning criteria for practical use. In the authors' experience, a lack of understanding around this is encouraging the use of external consultancies when, with some thought (Supplementary Material A), solutions could be found in house.

It is also worth acknowledging some data-related issues affecting on-the-ground usage. Firstly, different methods require the availability of different variables. A judgemental split on age requires just that variable, whereas a decision tree approach requires many more attributes in order for those most explanatory to be identified. Note, however, that as the size of a dataset increases so too do computational requirements. This is less of an issue for the judgemental splits and prescribed binning criteria classes of method (which simply involve binning and can be parallelised if necessary), but can pose a problem for derived methods which often involve computationally intensive single-threaded processes. For the 51,072 observation dataset of this study, desktop computation times ranged from a matter of seconds (for a judgemental split) to one hour (*k*-medoids).

All of these considerations contribute to gauging method appropriateness for use within particular settings. For general practitioners keen to segment their patient lists, the simplicity and accessibility offered by a judgemental split or prescribed binning criteria may outweigh inferior discrimination when compared to derived methods. Use of such solutions would also facilitate consistent cohort-level spend comparisons or benchmarking between practices or sub-systems within a wider system, due to the fixed descriptions for segment membership. If the onus is more on system-level strategy, then uncovering greater differences between population segments associated with a more discriminative decision tree approach may be valued greater than the costs in terms of sourcing data, computation, and resourcing. For instance, is the odd day or two of computational time really of consequence when developing the long-term strategic direction of population health over many years?

4.3 Variable selection

It is possible that more discrimination could be achieved if social and community care utilisation were included alongside the primary and secondary care data considered both here and in previous segmentation efforts [30,31]. This is because there is a more pronounced demarcation in the types of people care is provided to in these settings (i.e. typically to those that are more frail). Further work to investigate this in the UK will need to overcome the challenge of burdensome information governance policy, which is restricting linkage of healthcare datasets.

As previously mentioned (Section 3.3) this study has excluded activity as an explanatory variable due to concerns around circularity. Further caution should be exercised around variables which may proxy activity or spend, e.g. whether a person has had a fall or not in the last 6 months (this recorded social status variable typically would result in A+E attendance and/or admission). Ultimately, for utilisation-based population segmentation, explanatory variables should be relatively temporally-stable, such as the chronic condition data used here (noting the possibility of false negatives in these, e.g. through persons not regularly attending GP check-ups).

It should be noted that while this paper has considered spend as the targeted measure for which discrimination is sought, other variables could also have been used. For instance, activity may have been an alternative measure of utilisation to spend (noting that some discrimination may be lost since each non-elective admission, for example, would be considered equally regardless of length of stay – an important marker for resource utilisation). Alternatively, clinical outcome measures such as 30-day re-admission risk or performance measures such as waiting time could also have been targeted.

4.4 Selection of segmentation methods used in this study

While this study has sought to present and compare a variety of methods for utilisation-based population segmentation, it is not exhaustive and, depending on the setting and question posited, there may be more appropriate approaches than those considered here. Healthcare practitioners with an interest in population segmentation would be wise to maintain awareness of how the statistical literature in this field evolves, both in the healthcare setting and in others such as market segmentation (Section 2.2). In particular, machine learning is a discipline for which there is currently much interest, and where objective and non-objective classification techniques are developing at a substantial rate.

A method not assessed here but that may be familiar to some healthcare organisations is the Johns Hopkins ACG system. This approach is, in the UK, often associated with segmentation. However, while it does provide groupings for individuals, these are too high in number to be meaningful and useful for strategic population segmentation. The main focus of the ACG system is in patient-level cost prediction (which it does via linear and logistic regression-based approaches), and therefore, to this end, it should be pitted against random forests and other derived predictive approaches which have the important advantage of being calibrated specifically to the region in question [23,24].

Finally, it is worth commenting that such cost predictive methods can be used in conjunction with the segmentation methods explored here. In this study, segmentation has been performed descriptively on the basis of actual spend, but this can readily be replaced by predicted spend, i.e. the output from a predictive model as above. The notion behind this would be that observed spend used here is merely a sample of what could possibly occur in a window of time, but that by first predicting cost one would be controlling for the effects of extreme or unusual values that could occur for some individuals within the period considered.

4.5. Conclusions

There is not necessarily a *right* or *wrong* method for descriptive population segmentation: there are many factors to weigh up relating to discrimination and practicality. That being said, this study finds that clustering approaches are on-the-whole unsuitable; being computationally expensive, lacking discrimination, and requiring laborious pre-processing. Prescribed binning criteria, such as Bridges to Health, are unlikely to achieve high levels of discrimination between cohorts but may be useful on the grounds of segment interpretability or benchmarking with other health systems. Decision trees, particularly Breiman's CART, should be the preferred option since they offer a sound conceptual fit to the problem and promote good discrimination through calibration to local data. If these cannot be readily implemented, whether due to insufficient data or expertise, then judgemental splits focusing on the number of chronic conditions should be favoured.

Competing interests

The authors declare that they have no competing interests.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author contributions

RW designed the study and obtained the data. RB and RW conducted the literature review. RB, BM and RW performed the statistical investigations. RW wrote the paper.

Acknowledgements

The authors would like to acknowledge the help and support of business intelligence colleagues with a special thanks to Emma Gara and Chris Davies. The authors are also grateful to the general practice surgeries who have contributed the primary care data that has enabled this study.

References

[1] Kindig D, Stoddart G. What Is Population Health? *Am J Public Health*. 2003;93(3):380-383. doi:10.2105/AJPH.93.3.380

[2] Bartley M. *Health Inequality: An Introduction to Concepts, Theories and Methods*. Cambridge, UK: Polity; 2003.

[3] Charlton J, Rudisill C, Bhattarai N, Gulliford M. Impact of deprivation on occurrence, outcomes and health care costs of people with multiple morbidity. *J Health Serv Res Policy*. 2013;18(4):215-223. doi:10.1177/1355819613493772

[4] Lillrank P, Groop PJ, Malmström TJ. Demand and supply-based operating modes--a framework for analyzing health care service production. *Milbank Q*. 2010;88(4):595-615. doi:10.1111/j.1468-0009.2010.00613.x

[5] Baxter S, Johnson M, Chambers D, Sutton A, Goyder E, Booth A. The effects of integrated care: a systematic review of UK and international evidence. *BMC Health Serv Res*. 2018;18(1):350. doi:10.1186/s12913-018-3161-3

[6] NHS. *NHS Long Term Plan » Online Version of the NHS Long Term Plan*.; 2019. https://www.longtermplan.nhs.uk/online-version/. Accessed January 18, 2019.

[7] Bertsimas D, Bjarnadóttir MV, Kane MA, et al. Algorithmic Prediction of Health-Care Costs. *Operations Research*. 2008;56(6):1382-1392. doi:10.1287/opre.1080.0619

[8] Oh H, Parks SC, Demicco FJ. Age- and Gender-Based Market Segmentation. *International Journal of Hospitality & Tourism Administration*. 2002;3(1):1-20. doi:10.1300/J149v03n01_01

[9] Chang M-H, Molla MT, Truman BI, Athar H, Moonesinghe R, Yoon PW. Differences in healthy life expectancy for the US population by sex, race/ethnicity and geographic region: 2008. *J Public Health (Oxf)*. 2015;37(3):470-479. doi:10.1093/pubmed/fdu059

[10] NHS England. *"How to" Guide: The BCF Technical Toolkit; Population Segmentation, Risk Stratification, and Information Governance*.; 2014. https://www.england.nhs.uk/wp-content/uploads/2014/09/1-seg-strat.pdf.

[11] Lynn J, Straube BM, Bell KM, Jencks SF, Kambic RT. Using Population Segmentation to Provide Better Health Care for All: The "Bridges to Health" Model. *The Milbank Quarterly*. 2007;85(2):185-208. doi:10.1111/j.1468-0009.2007.00483.x

[12] Low LL, Kwan YH, Liu N, Jing X, Low ECT, Thumboo J. Evaluation of a practical expert defined approach to patient population segmentation: a case study in Singapore. *BMC Health Services Research*. 2017;17(1):771. doi:10.1186/s12913-017-2736-8

[13] Joynt KE, Figueroa JF, Beaulieu N, Wild RC, Orav EJ, Jha AK. Segmenting high-cost Medicare patients into potentially actionable cohorts. *Healthcare*. 2017;5(1):62-67. doi:10.1016/j.hjdsi.2016.11.002

[14] Clegg A, Bates C, Young J, et al. Development and validation of an electronic frailty index using routine primary care electronic health record data. *Age Ageing*. 2016;45(3):353-360. doi:10.1093/ageing/afw039

[15] Breiman L. *Classification and Regression Trees*. Routledge; 2017. doi:10.1201/9781315139470

[16] Bardsley M. *Understanding Analytical Capability in Health Care: Do We Have More Data Than Insight?* The Health Foundation; 2016. https://www.health.org.uk/publications/understanding-analytical-capability-in-health-care.

[17] Griffiths JD, Williams JE, Wood RM. Modelling activities at a neurological rehabilitation unit. *European Journal of Operational Research*. 2013;226(2):301-312. doi:10.1016/j.ejor.2012.10.037

[18] Lemon SC, Roy J, Clark MA, Friedmann PD, Rakowski W. Classification and regression tree analysis in public health: Methodological review and comparison with logistic regression. *ann behav med*. 2003;26(3):172-181. doi:10.1207/S15324796ABM2603_02

[19] Wu BU, Johannes RS, Sun X, Tabak Y, Conwell DL, Banks PA. The early prediction of mortality in acute pancreatitis: a large population-based study. *Gut*. 2008;57(12):1698-1703. doi:10.1136/gut.2008.152702

[20] Patel RB, Mathur MB, Gould M, et al. Demographic and Clinical Predictors of Mortality from Highly Pathogenic Avian Influenza A (H5N1) Virus Infection: CART Analysis of International Cases. *PLOS ONE*. 2014;9(3):e91630. doi:10.1371/journal.pone.0091630

[21] Cairney J, Veldhuizen S, Vigod S, Streiner DL, Wade TJ, Kurdyak P. Exploring the social determinants of mental health service use using intersectionality theory and CART analysis. *J Epidemiol Community Health*. 2014;68(2):145-150. doi:10.1136/jech-2013-203120

[22] Peltz A, Hall M, Rubin DM, et al. Hospital Utilization Among Children With the Highest Annual Inpatient Cost. *Pediatrics*. 2016;137(2):e20151829. doi:10.1542/peds.2015-1829

[23] Sushmita S, Newman S, Marquardt J, et al. Population Cost Prediction on Public Healthcare Datasets. In: *Proceedings of the 5th International Conference on Digital Health 2015*. DH '15. New York, NY, USA: ACM; 2015:87–94. doi:10.1145/2750511.2750521

[24] Morid MA, Kawamoto K, Ault T, Dorius J, Abdelrahman S. Supervised Learning Methods for Predicting Healthcare Costs: Systematic Literature Review and Empirical Evaluation. *AMIA Annu Symp Proc*. 2018;2017:1312-1321.

[25] Hothorn T, Hornik K, Zeileis A. Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*. 2006;15(3):651-674. doi:10.1198/106186006X133933

[26] Quinlan JR. *C4.5: Programs for Machine Learning*. 1 edition. San Mateo, Calif: Morgan Kaufmann; 1992.

[27] Kass GV. An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Journal of the Royal Statistical Society Series C (Applied Statistics)*. 1980;29(2):119-127. doi:10.2307/2986296

[28] 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):2198-2205. doi:10.1016/j.apmr.2012.05.026

[29] Liao M, Li Y, Kianifard F, Obi E, Arcona S. Cluster analysis and its application to healthcare claims data: a study of end-stage renal disease patients who initiated hemodialysis. *BMC Nephrology*. 2016;17(1):25. doi:10.1186/s12882-016-0238-2

[30] Vuik SI, Mayer E, Darzi A. A quantitative evidence base for population health: applying utilization-based cluster analysis to segment a patient population. *Population Health Metrics*. 2016;14(1):44. doi:10.1186/s12963-016-0115-z

[31] Low LL, Yan S, Kwan YH, Tan CS, Thumboo J. Assessing the validity of a data driven segmentation approach: A 4 year longitudinal study of healthcare utilization and mortality. *PLoS ONE*. 2018;13(4):e0195243. doi:10.1371/journal.pone.0195243

[32] Hartigan JA, Wong MA. Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society Series C (Applied Statistics)*. 1979;28(1):100-108. doi:10.2307/2346830

[33] Huang Z. A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining. In: *In Research Issues on Data Mining and Knowledge Discovery*. ; 1997:1–8.

[34] 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

[35] Kaufman L, Rousseeuw PJ. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons; 2009.

[36] Wagner S, Wagner D. *Comparing Clusterings - An Overview*. Karlsruhe: Universität Karlsruhe, Fakultät für Informatik; 2007:19.

[37] Hubert L, Arabie P. Comparing partitions. *Journal of Classification*. 1985;2(1):193-218. doi:10.1007/BF01908075