Appendix E: Data Analysis

Data Analysis Team: Peter Martin, Amy Macdougall, Andy Whale

1. Introduction

This appendix describes the analysis carried out on data collected as part of the Payment Systems Pilot Project, insofar as it informed the development of a casemix classification for CAMHS. It will be useful to give a short overview of the structure of this appendix.

- The next section, Section A.2, describes the sample selection, presents a description of the Analysis Sample, and comments on data quality.
- Section A.3 lays out our analytical strategy, which involved carrying out three different approaches to classification, and comparing the results in terms of model fit and clinical meaningfulness. Two of these approaches were "statistical", in the sense that in each case the classification was devised via a data analytic procedure (with current view ratings as input), rather than from clinical concepts. These two approaches were unsupervised cluster analysis, and supervised cluster analysis. The third approach to classification was to define a grouping scheme from clinical considerations, drawing on NICE guidance and the THRIVE model for CAMHS (Wolpert et al. 2014), and devise an algorithm that assigns CAMHS patient to a group based on their characteristics (as rated by clinicians on the current view form).
- Section A.4 describes the findings from each of the three approaches.
- The three approaches are then compared with one another in Section A.5. The conceptually-derived classification performed better than either of the two statistical approaches.
- Section A.6 describes the resulting model of classification in more detail, and investigates the evidence for an additional influence of context problems, EET issues, and complexity factors.
- Finally, Section A.7 reports sensitivity analyses that investigate the robustness of our results to assumptions we have made, and to certain limitations in the operationalizations used.

A note on terminology: For the purpose of this appendix, the terms "cluster", "grouping" and group are used interchangeably. We will tend to use "grouping" when referring to the categories of CAMHS patients proposed in the main report. The term "cluster" will refer to groups defined on the basis of statistical procedures, or groups used in statistical analysis. We speak of "cluster development" to refer to the statistical process of devising and testing different ways to group CAMHS users.

2. Data and Data Quality

Data were collected from 20 CAMH services. Data were submitted according to the Children and Young People's Improving Access to Psychological Therapies data set specification (Version 3). A rigorous process was conducted to select services for inclusion in the analysis sample for the purpose of cluster development. This included inspection of the data received, as well as inquiries with data managers and other staff at the participating services to ascertain the quality of the data collected. In particular, we took care to ascertain how information regarding treatment activity was recorded, and drew conclusions regarding the completeness (or otherwise) of activity records from each participating service. Time constraints meant that unfortunately we were unable to investigate data quality at the level of the individual child; for example, we were unable to validate data received for individual children by comparing them with clinical case notes.

The unit of analysis was a "period of contact": a period from the point at which referral is received or accepted, to the point of case closure. We applied strict criteria to decide whether to include data from a particular period of contact for cluster development. The inclusion criteria for periods of contact were as follows:

- Must be closed or "dormant" (without activity for at least six months);
- Must have Current View Form completed at assessment;
- Must have information on activity, and at least one direct contact ("appointment") must be recorded as having been attended by the child or young person;
- Must come from a service whose data quality overall was sufficiently strong.

A decision on the last criterion was made via a combination of data inspection and communication with the services. In many cases, service representatives told us about problems in the collection of activity data that suggested that information from their service was incomplete or otherwise not valid. If this was the case, we excluded data from the service from the Analysis Sample for cluster development.

The Analysis Data Set comprised clinical records from 4573 periods of contact in 11 CAMH services. All periods of contact had a completed Current View Form at assessment, and at least one direct contact had taken place. Table E.1 displays the numbers of periods of contact included from each of the 11 services. A single service ("A") provided almost a third of useable cases. The service that made the smallest contribution ("K") to the Analysis Data Set provided just under 1 % of useable cases.

Table E.2 gives the age and gender distribution of the Analysis Data Set. A full account of the demographic characteristics of the Analysis Sample, and a comparison with large samples of children seen in CAMHS from other data sources is provided in Appendix F.

Table E.1 Analysis Sample

Service	Number of POCs	Proportion in Analysis Sample
A	1451	31.7%
B	537	11.7%
C	499	10.9%
D	470	10.3%
E	445	9.7%
F	314	6.9%
G	299	6.5%
H	293	6.4%
I	116	2.5%
J	114	2.5%
K	35	0.8%
Total	4573	100.0%

Table E2. Payment Systems age and gender breakdown

		Age Group				
		0-4 years (%)	**5-9 years (%)**	**10-14 years (%)**	**15-19 years (%)**	**Totals (%)**
Gender	**Male (%)**	87 (64.4%)	605 (66.5%)	836 (48.0%)	547 (32.7%)	**2114 (47.3%)**
	Female (%)	48 (35.6%)	305 (33.5%)	916 (52.0%)	1125 (67.3%)	**2438 (52.7%)**
	Totals (%)	**135 (3.0%)**	**910 (20.3%)**	**1742 (39.0%)**	**1672 (37.4%)**	**4469 (100%)**

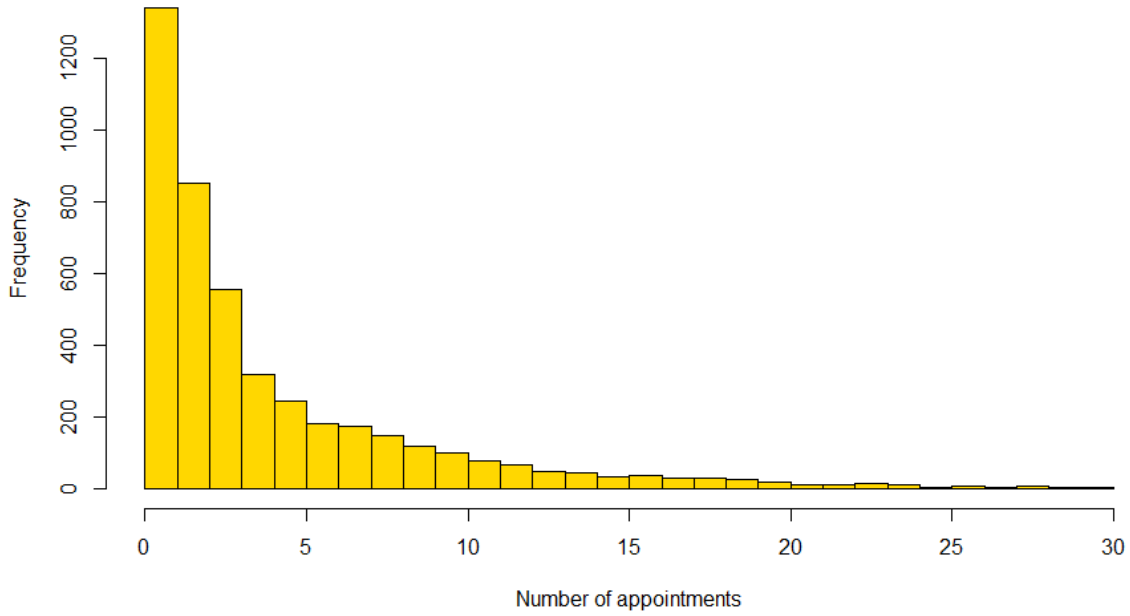
Note: 83 periods of contact had no information on the child's age; 21 periods of contact had no information on the child's gender. These are not included in this table, but are included in the analysis.

Measures

Resource Use. Our primary indicator of resource use was "Number of Appointments", i.e. the number of direct contacts (face-to-face or telephone) that the service user (child or young person) had with the CAMH service. We counted only appointments that had actually been attended, and did not count missed or cancelled appointments. The distribution of "Number of Appointments" is illustrated in Figure E.1, and shown in Table E.2. We show this distribution in the Analysis Sample to inform the reader of the data used for the development and testing of classifications. Note that statistics (such as the mean, the median, etc.) derived from this Sample do not represent good estimates of the population ("real") distribution of number of appointments, mainly because the sample is biased towards

shorter periods of contact (because of the relatively short observation period; see Appendix F). A detailed account of this bias, along with better estimates of the population distribution of "Number of Appointments" in participating services, are given in Appendix F.

Figure E.1: Bar chart of "Number of Appointments" in the Analysis Sample



Note: n=4573. Forty children (0.87 % of the sample) attended more than 30 appointments. These are not shown in this graph, but are included in the analysis.

Table E.2: Frequency distribution of "Number of Appointments" in the Analysis Sample

Number of appointments	Frequency	Total %
1	1340	29.3
2	851	18.6
3	556	12.2
4	319	7.0
5	244	5.3
6	182	4.0
7	174	3.8
8	150	3.3
9	119	2.6
10	101	2.2
11-20	413	9.0
21-30	84	1.8
31 or more	40	0.9
Total	4573	100.0

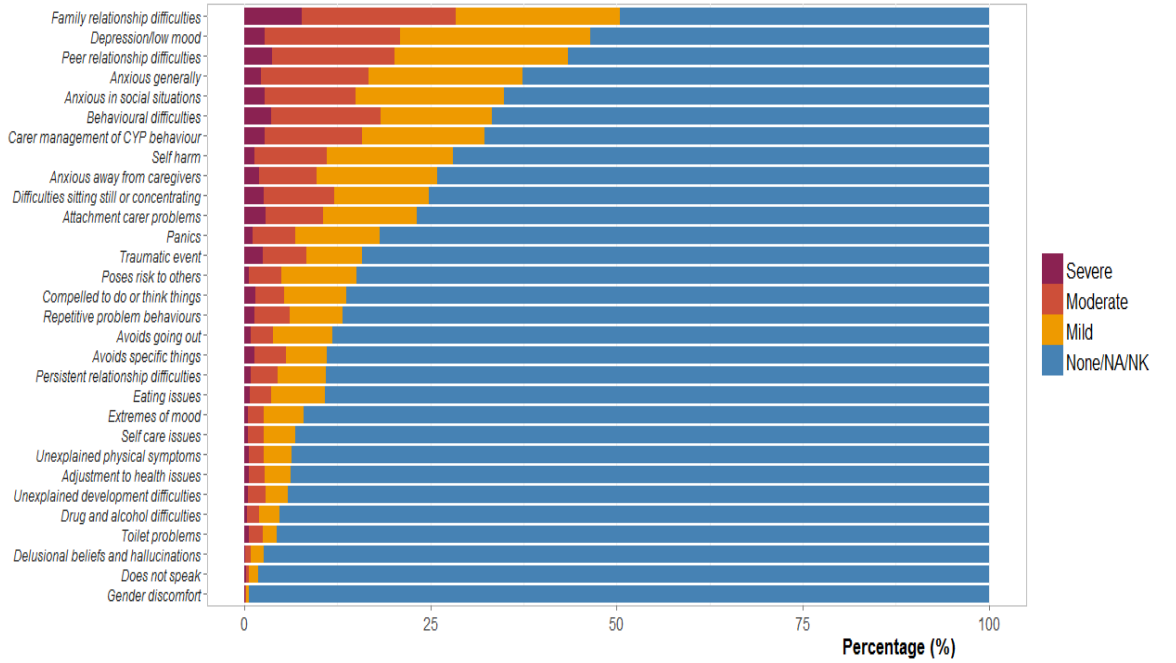
Notes: Summary statistics: Mean = 4.96, SD = 6.53,
1st quartile = 1, Median = 3, 3rd quartile = 6, Maximum = 101.

Duration and staff presence. The data specification also included variables recording the duration of appointments, and the number and type of staff present. These variables contained a high proportion of missing values, and were therefore initially not used in cluster development or testing. However, we did conduct a sensitivity analysis, employing multiple imputation of missing values, in order to gauge the extent to which our findings may have been influenced by 'ignoring' duration of appointments and staff involvement. This sensitivity analysis is reported in section E.7.

Presenting Information. Our indicators of presenting information were the ratings the clinician gave their client on the Current View Form at assessment. For the purpose of our analysis, ratings of "None", "Not Known" were combined with missing ratings to form a common category that records if there is no evidence for the presence of a problem. The distributions of current view problem descriptions, complexity factors, contextual problems, and EET (Education, Employment, Training) issues are shown in Figures E.2 to E.6. A table of the distribution of presenting problems by age and gender is given in Appendix F.

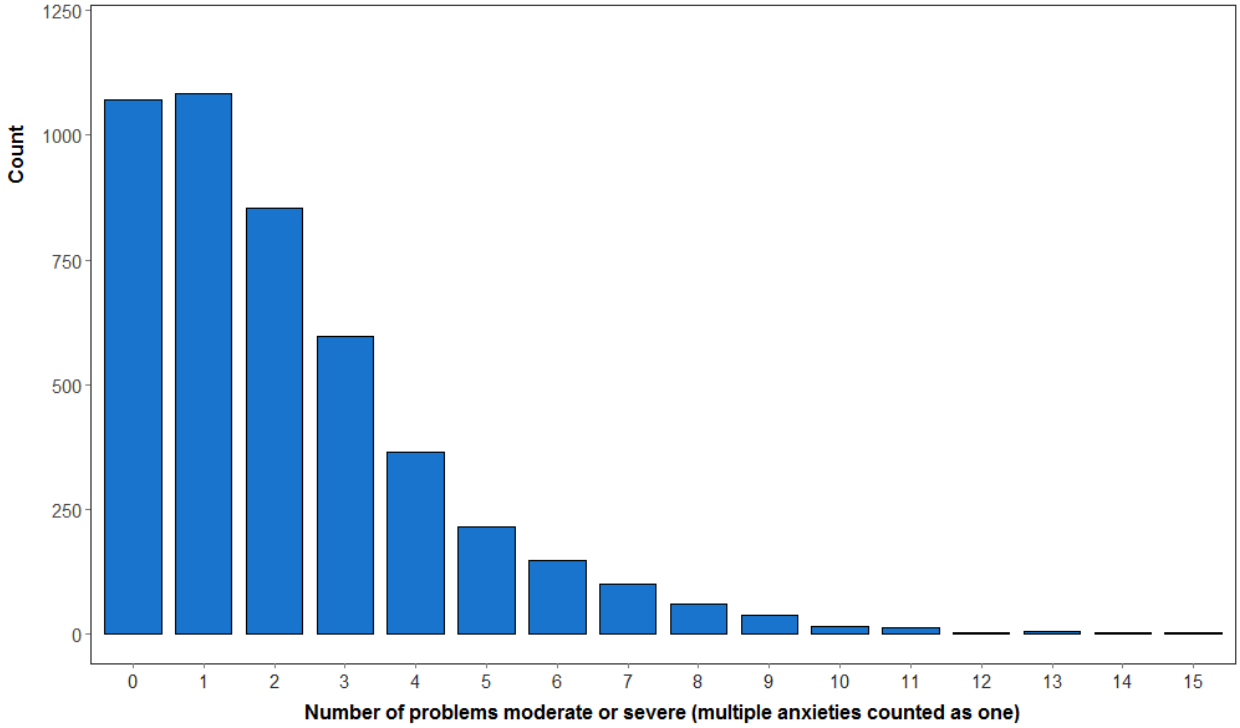
The data illustrate the diversity of children seen in CAMHS, and the complexity of many presentations. It is noteworthy that two of the three problem descriptions rated as present most frequently are "Family Relationship Difficulties" and "Peer Relationship Difficulties" – types of difficulties that are not associated with a particular diagnosis (see Figure E.2). As Figure E.3 shows, about a quarter of periods of contacts began with the child not having any other than mild problems, according to current view ratings. About another quarter of children came with a single moderate or severe problem. And over half of all children presented with more than one problem rated moderate or severe. Figures E.5 and E.6 demonstrate the frequent presence of complexity factors and contextual problems that children present with at CAMHS.

Figure E.2: Frequency Distributions of Current View Presenting Problems



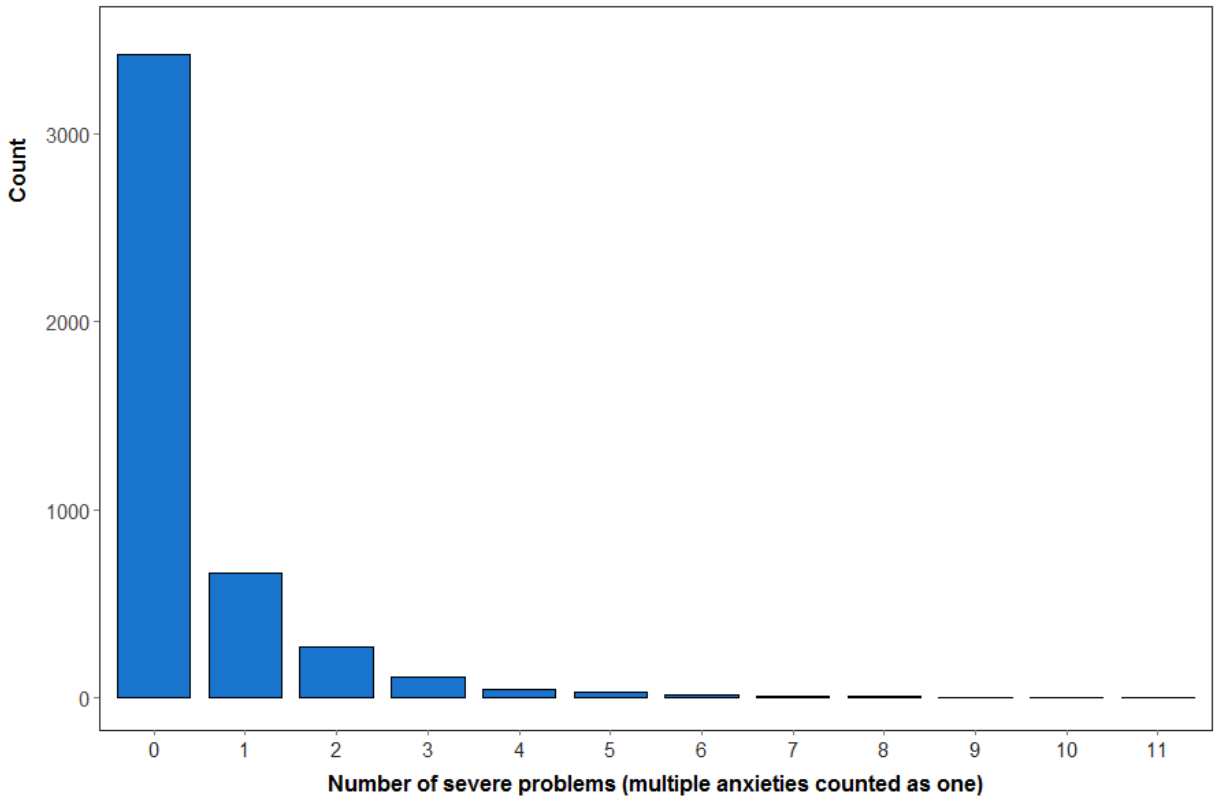
Note: Analysis Sample, n = 4573.

Figure E.3: Number of Presenting Problems Rated Moderate or Severe



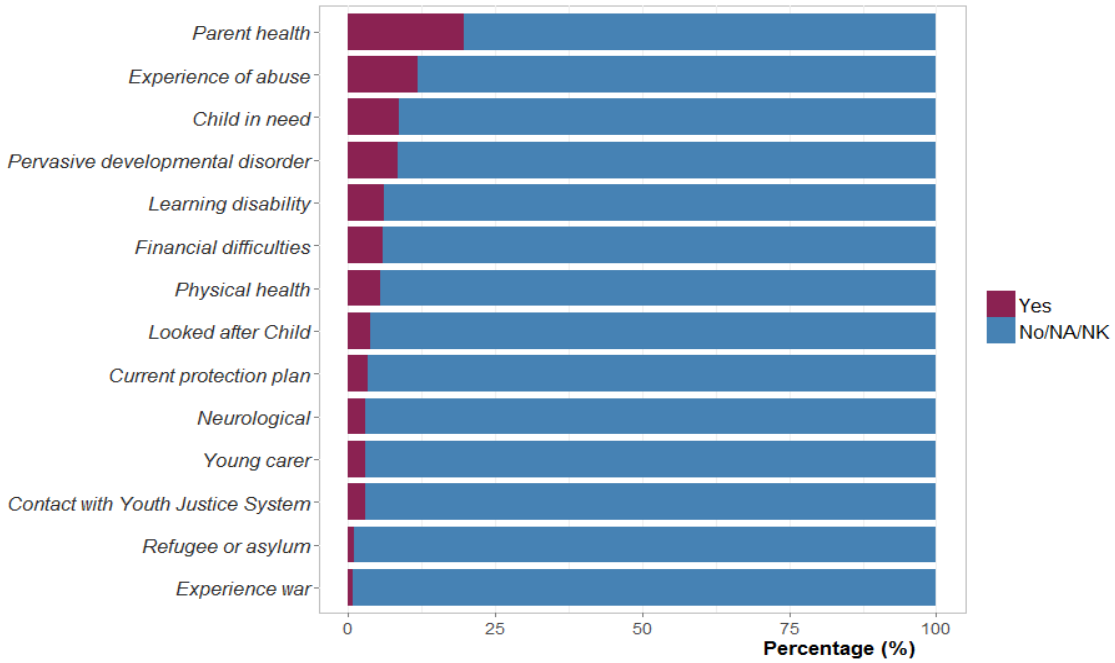
Note: Analysis Sample, n = 4573. If several types of anxiety were rated as moderate or severe, this was counted as a single problem ("anxiety") only, for the purpose of this graph.

Figure E.4: Number of Presenting Problems rated Severe



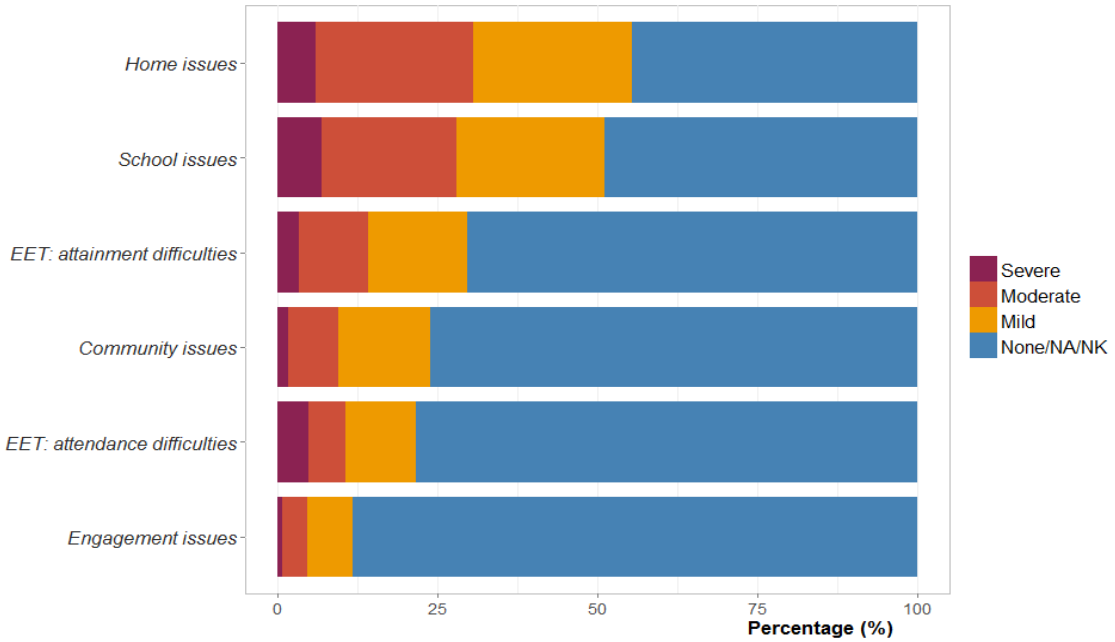
Note: Analysis Sample, n = 4573. If several types of anxiety were rated as moderate or severe, this was counted as a single problem ("anxiety") only, for the purpose of this graph.

Figure E.5: Distributions of Current View Complexity Factors



Note: Analysis Sample, n = 4573.

Figure E.6: Current View Contextual Problems and EET Issues



Note: Analysis Sample, n = 4573. EET: Employment, Education, Training.

3. Research Strategy

The main aim of data analysis was to produce and test a classification ("groups", "clusters") of children seen in CAMHS. The quality of this classification was to be assessed by three criteria: clinical meaningfulness, reliability of cluster allocation, and relationship of clusters to resource use. We considered three approaches to classification: unsupervised cluster analysis, supervised cluster analysis, and conceptually informed grouping.

Unsupervised Cluster Analysis

Unsupervised cluster analysis (UCA) is a summary term for methods of classification that cluster cases (here: individual periods of contact) on the basis of a given set of characteristics, without reference to a dependent variable (outcome, response). For our case, this approach implies that clusters are derived from presenting information, collected via the Current View Form, without taking into account any indicators of resource use.

Supervised Cluster Analysis

Supervised cluster analysis (SCA) refers to methods of classification that cluster cases on the basis of the relationship between a set of characteristics (in our case, the Current View Information) and a dependent variable (in our case, indicators of resource use). The aim is to find subsets of a sample that can be identified by their presenting information, so that the members of each subset are similar with respect to resource use, while tending to be different from members of other subsets with respect to resource use.

Clinically informed grouping

We defined a conceptual classification of children coming to CAMHS based on two principles:

(1) A broad distinction between patients in terms of the type and intensity of their needs into three supergroupings: "Getting Advice", "Getting Help", and "Getting More Help";
(2) our review of the NICE guidance (described in section 6.3 of the main report; see also Vostanis et al., in press).

We posited that children for whom the same NICE guidance was appropriate would have similar needs, both qualitatively – in terms of the type of treatment – and quantitatively, in terms of the cost of resource provision to the service. In order to identify which NICE guidance may be appropriate for which child, we used information from the current view form. Two practicing CAMHS clinicians prepared a decision table that specifies precise rules for allocating a CAMHS patient to a group based on a clinician's current view ratings of the patient.

4. Classification

4.1 Unsupervised CA

The primary aim of the project is to create groupings of patients which are clinically meaningful, predictive of resource use, and able to be reliably identified. When investigating groups within data an obvious statistical method to consider is Cluster Analysis. The aim of Cluster Analysis is to detect groups of patients who are more similar to each other, than to other patients in the dataset. In this section we assess the contribution that Cluster Analysis can make to the overall aim of clustering patients.

Note that in this appendix the term 'cluster' will refer to clusters found using statistical methods, **not** the groupings defined in the main report (which will be referred to here as 'conceptually derived '). The clustering will be based on information from the Current View form only, and not include any dependent variables such as number of sessions. This type of clustering is called 'unsupervised'. Clusters will be assessed in terms of:

- how well they fit the data,
- how clinically meaningful they are,
- how well they predict resource usage (compared to the conceptually driven groups).

The first criterion relates to the objective that clusters are able to be reliably identified. That is, given a patient, it is clear which cluster they fall into. Well-fitting clusters divide patients into distinct groups. Poorly fitting clusters contain patients who are no more similar to others within their own cluster, than in other clusters. If we fail to find well-fitting clusters it would be very difficult to create a system of clustering in which cluster membership could be reliably identified.

4.1.1 Methods

The main clustering method selected, *k*-medoids, was one which finds mutually exclusive clusters, based on information from the Current View form only.

Taking all items on the Current View form provides in total 50 variables for each patient, all categorical, some ordinal. Clustering was carried out using all 50 variables, and then using only the first 30 from the 'Provisional Problem Description' (PPD) section.

An appropriate metric for this type of data is the Gower distance, as defined in Gower (1970). A distance matrix was created using this metric, with the 'vegdist' function from the 'vegan' package (method 'altGower' which excludes double zeros[1]), within the statistical program R.

Before finding the distance matrix using the 30 variables from the PPD section, 156 patients were excluded who did not have any problems rated. These patients form their own cluster

[1] See appendix (*reliability page …)* for an explanation of why double zeros, that is when both practitioners have recorded the problem as not present, are not counted as 'agreement'.

which is added in to any clustering solution found using the reduced set of 30 variables. There were no patients with nothing selected over all 50 variables.

The main clustering method used was *k*-medoids, using the Partitioning Around Medoids algorithm (PAM, as described in Kaufman and Rousseeuw (2005)) implemented in the 'pam' function from the 'cluster' package in R. This is similar to *k*-means, but is suitable for categorical data.

To assess fit the silhouette was used (Kaufman and Rousseeuw (2009)), a measure of how close patients are to others in their assigned cluster compared to the next nearest cluster.

Following Rousseeuw (1986), for any individual *i* who has been placed in cluster *A,* define:

a(i) = average dissimilarity of *i* to all other individuals **in *A***

d(i,C) = average dissimilarity of *i* to all individuals in *C*

$$b(i) = \underset{C \neq A}{\text{minimum}}\ d(i,C)$$

Then the silhouette for *i* is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

The average silhouette is found by taking the average *s(i)* over all individuals in the data set. The range for a silhouette is between -1 and 1. A high silhouette (approximately between 0.7 and 1) indicates that an individual is close to individuals within its own cluster, and far from individuals in other clusters. That is, there is a greater distance between clusters than within clusters. A silhouette close to 0 (and as an approximate rule less than 0.25) suggests that an individual is no closer to others within its own cluster, than individuals in other clusters.

To inform decisions about the number of clusters k, the average silhouette was inspected. See Figures E.7 and E.8 for hypothetical examples (not based on our own data) of one data set with a a low silhouette (0.35), and one with a high one (0.88). The clusters found using cluster analysis are indicated by the differently shaped and coloured points. In Figure E.8, the distances between points within the clusters are smaller than between the clusters – so there is a high average silhouette. These clusters appear to fit well. In Figure E.7, there is no such clear separation, so the average silhouette is low. Although the data we will be clustering here is categorical, not continuous as in these examples, they indicate the meaning of the silhouette in general.

Solutions with locally high average silhouettes were assessed in terms of how well they could predict resource usage, and how clinically meaningful the clusters were.

To assess how well clusters predicted resource usage, generalised linear mixed models were fitted as in the main analysis and the AIC and BIC inspected (see section 5 of this appendix).

The provisional problem descriptors for each cluster were plotted in order to assess clinical meaningfulness. A few illustrative examples will be shown in the results section.

Figure E7: A poorly fitting 2 cluster solution

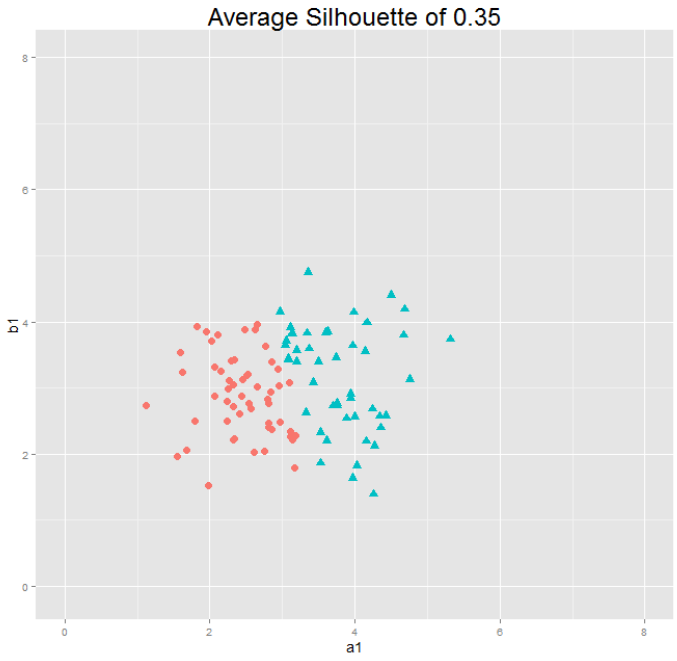
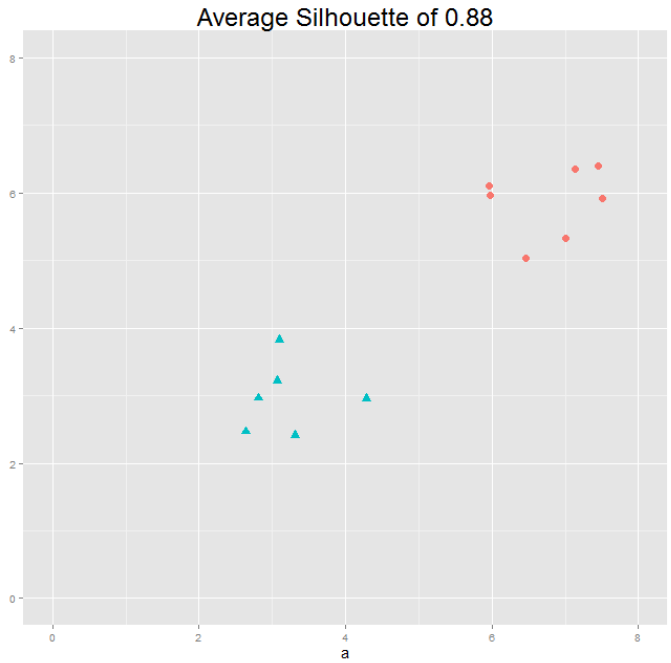


Figure E8: A well-fitting 2 cluster solution

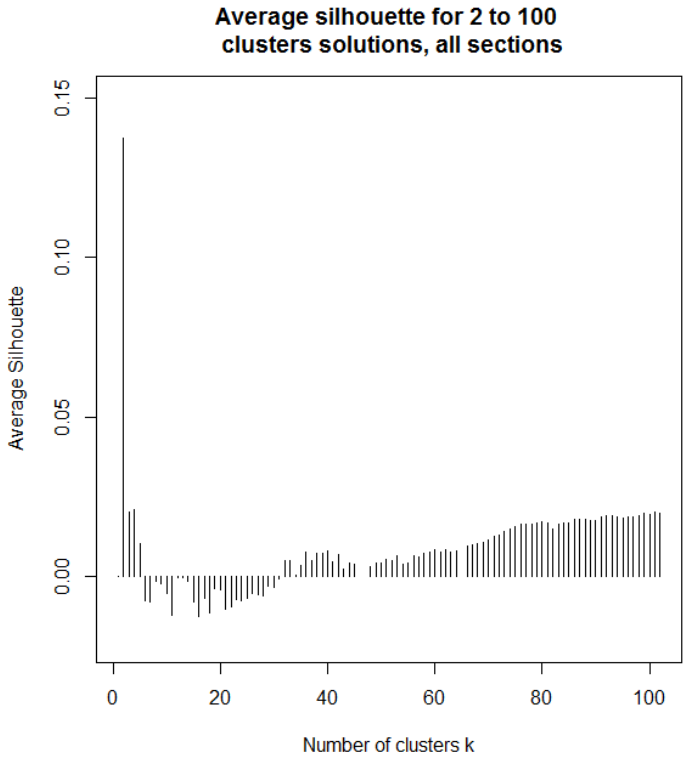


Other methods were explored alongside PAM, including hierarchical clustering with average linkage and fuzzy clustering (using the same distance matrix as described above, both also implemented in the 'cluster' package in R). The least poor results were found using PAM.

4.1.2 Results A: Fit of clusters

The first approach clustered patients using all 50 variables from the Current View form, using the PAM clustering technique. For an initial impression of an appropriate number of clusters *k* the average silhouette was calculated for 2 to 100 cluster solutions. The results are plotted in figure E.8.

Figure E.8: The average silhouette for *k* from 2 to 100



The average silhouette is very close to zero, if not negative, for *k* up to 100. This indicates that the clusters fit the data poorly. The least poor solution is for *k*=2, which has an average silhouette of 0.14.

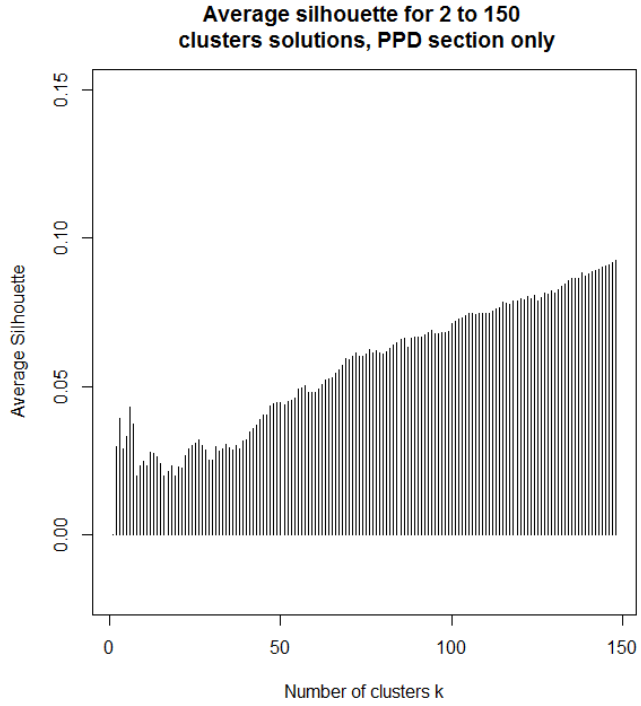
Another approach was considered: using only the 30 variables from the PPD part of the form. This yielded higher average silhouettes, however still significantly below what might be a reasonable value (taken to be at least 0.25, as a very low minimum). The average silhouette is shown for *k* from 2 to 150 in figure E.9.

The average silhouette is clearly rising with *k.* However, solutions with hundreds of clusters may provide a better fit but be of little use practically.

Failure to find a solution with a reasonably high average silhouette means that no solution which separates the patients into distinct groups was found. That is, many patients were no more similar to others within their own cluster, than to many patients in other clusters. It would therefore be difficult to create a system of clustering patients in which cluster membership could be reliably identified.

As mentioned above, poor results were also found using other methods. This suggests that it may not be possible to form distinct, well separated clusters on the basis of the Current View form information alone, using an unsupervised cluster analysis approach.

Figure E.9: Average silhouette for *k* from 2 to 150



4.1.3 Results B: Clinical meaningfulness

The cluster analysis failed to find well-fitting clusters. In this section we look at the meaningfulness of the clusters from a clinical perspective (using an illustrative sample of clusters). That is: do the patients within each cluster form coherent groups in terms of their presenting problems?

As the number of conceptually driven groups was under 20, only solutions with relatively small numbers of clusters were considered. Solutions with locally high average silhouettes were selected:

- 2 clusters, all variables (all items on CV form) **average silhouette=0.137**,
- 6 clusters, PPD section only **average silhouette=0.043**,
- 26 clusters, PPD section only **average silhouette=0.0319**.

The proportions of patients with each problem from the PPD section were inspected for each cluster. As the 26 cluster solution (using variables from the PPD section only) provided the best prediction of resource use (according to the AIC), the first six clusters of this solution are shown in Figure E.10.

Given the poor fit of the clusters to the data, we might expect to see an almost random mix of problems within each cluster. This is not the case, as some clusters correspond to recognisable groups of patients. For example: clusters 1 (mild depression), 2 (multiple anxieties), and 6 (ADHD). Clusters 4 and 5 are less well defined, and a particular feature is that there is no common problem or set of problems across all patients in that cluster. In cluster 5, less than half of the patients have the most common problem.

The rest of the clusters were also a mix of recognisable groups, and random mixtures. Overall the clusters were dominated by the most common problems (for example ‘depression/low mood’ and ‘family relationship difficulties’). Patients with less common problems were not captured: for example those with Psychosis or Eating Disorders. This is perhaps because of low numbers. Patients with these problems are spread across multiple clusters.

4.1.4 Results C: Prediction of resource use

So far we have seen that clusters found using PAM have failed to create clearly distinct groups, but that some groups were somewhat meaningful in terms of patients’ problems. Lastly we see how they compare with the conceptually driven clusters in terms of prediction of resource use (number of sessions attended).

Number of sessions attended was predicted using cluster membership and service ID, as in the main analysis (see section 5 of this appendix for details). Inspection of the AIC suggests that, out of the classifications found by unsupervised cluster analysis, the best fitting model included the 26 cluster solution (based on information from the PPD section only), plus the patients who were excluded, making 27 clusters in total. However, this model fit worse than the conceptually-driven model.

4.1.5 Summary

The aim of cluster analysis is to find groups (if they exist) of individuals who are more similar to each other, than to other patients. The best fitting cluster solution (with more than two clusters) used only information from the PPD section of the Current View form, and did not fit the data well. Patients were no more similar to those within their own cluster, than to patients in other clusters (this conclusion is based on distances between patients). This would make creating a clustering system in which membership could be reliably identified difficult.

There was some degree of clinical meaningfulness in the least poor cluster solution, however no advantage was found over the conceptually driven clusters in this respect. There was also no improvement in prediction of resource use, compared to the conceptually driven clusters.

Furthermore, since the clusters fit this relatively small dataset poorly, there would be no reason to believe that these clusters would generalise well to larger numbers of patients.

It was therefore decided that there was no compelling evidence for the use of clusters found using unsupervised cluster analysis.

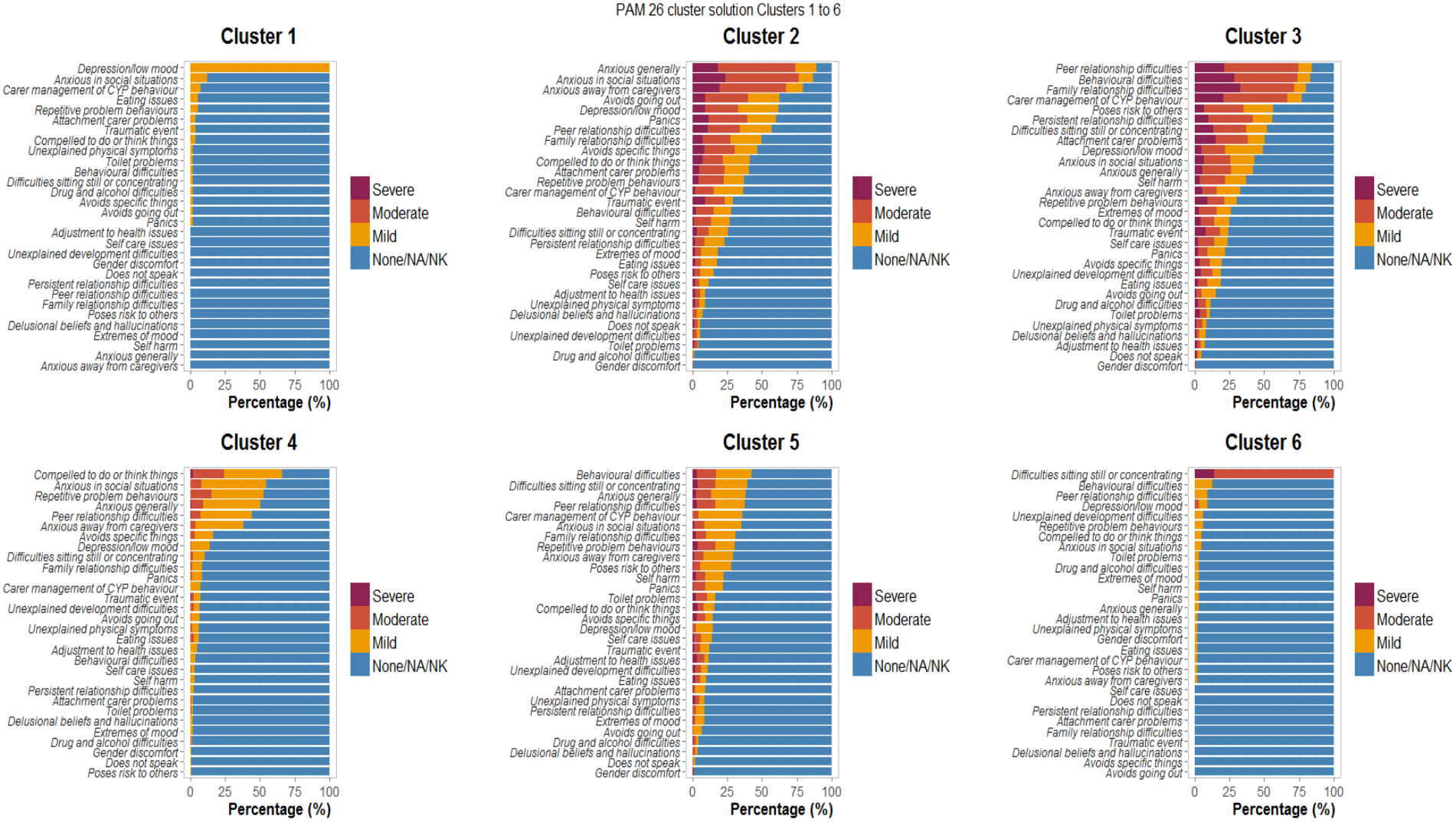


Figure E.10: Proportion of patients with each problem descriptor (plus severity) for clusters 1 to 6 of the 26 cluster solution.

4.2 Supervised CA

The aim of supervised cluster analysis was to attain a classification of children seen in CAMHS that would directly take into account resource use, operationalized as the number of appointments attended until case closure. We used recursive partitioning (Hothorn, Hornik & Zeileis 2006), which is a type of regression tree analysis. Regression trees consider the relationship between a dependent variable (in our case: "Number of Appointments") and a set of independent variables. We used 53 independent variables: the 50 ratings from the assessment current view form, and three variables that summarize the information from the 30 presenting problem variables: the maximum problem rating, the number of problems rated moderate or higher, and the number of problems rated severe. Recursive partitioning works by first considering the whole sample and looking for that division of the sample that maximizes the gain in prediction of the dependent variable. A division is defined via a particular dichotomization of one of the independent variables. For example, a possible division may split the sample into two groups: those with a "Low Mood" rating of mild or lower, and those with a "Low Mood" rating of moderate or higher. This results in the creation of two subgroups. Subsequently, the procedure considers all possible divisions within each resulting subgroup, and again selects the division that offers the best improvement in the prediction of the dependent variable. This procedure is repeated until some stopping criterion is satisfied. Different types of regression trees differ by the type of stopping rule employed. In recursive partitioning, a significance level is set, such that a division is not implemented if the resulting improvement in the prediction is not statistically significant (i.e. has a p-value larger than the set significance level).

Regression trees have the advantage that they are able to discover complex interactions between the independent variables. A disadvantage of regression trees is that the divisions are orthogonal with respect to the independent variables, and thus are not always very good at taking into account the correlations between independent variables. Another disadvantage of regression trees is their liability to overfitting, i.e. to grow a tree that provides an optimum division for a given sample, but that may not necessarily be reproducible in a new sample. Put differently, a regression tree grown from a single sample may not represent a reliable partition of the data.

We chose an analytic strategy that was designed to put the strengths of regression trees to best use, while minimizing the risk of unreliable results. We proceeded as follows: We produced ten random split-half samples (i.e. we randomly divided the Analysis Data Set into two halves in ten different ways), stratified by gender and age group so that the proportions of gender and age were nearly identical in each of the resulting twenty sub-samples. In each pair of samples, one was randomly assigned to be the "development data set", the other to be the "test data set". We then grew a regression tree on each development data set, using a range of significance levels to vary the point at which the procedure would stop.[2] (This procedure was suggested by Kuhn & Johnsson [2013: 183].) We then tested the trees resulting from each development sample on its associated test sample. We used the statistical model implied by each regression tree to predict the expected number of appointments for each member of the test sample, and assessed model fit via two statistics: "correlation R^2" (where R^2 is defined as the squared correlation between the

[2] The significance levels used were: 0.001, 0.005, 0.010, 0.015, 0.020, 0.025, 0.030, 0.040, 0.050, 0.060, 0.070, 0.080, 0.090, and 0.100.

observed and the predicted number of appointments; this is R_6^2 in Kvalseth [1985]), and "outlier-resistant R^2", called R_9^2 in Kvalseth (1985: 283), which is defined as $R_9^2 = 1 - \left(\frac{Median\{|y_i - \hat{y}_i|\}}{Median\{|y_i - \bar{y}_i|\}}\right)^2$.

For each of the ten test samples, we then selected two regression trees: the tree with the highest correlation R^2, and the tree with the highest outlier-resistant R^2. (Usually, the outlier-resistant R^2 selected a larger tree than the correlation R^2.) In order to compare the model fit of the regression trees to the conceptually-derived model, we then estimated a mixed effects negative binomial regression (adding a random effect for the service attended by the child). We also inspected the ten regression trees for consistency: if trees grown on different random subsets of our data were similar to one another, this would indicate that a classification based on these trees would be reliable (i.e. likely to be reproducible in new data). If trees differed considerably, this would indicate that such a classification would be unreliable. Results are reported in Section 5.2 of this appendix.

4.3 Conceptually-derived classification

The rules that decide which cluster a given child was allocated to for the purpose of analysis are displayed in Table E. 3. The clinical meaning of the clusters is explained in Section 7 of the full report. Note that there is no algorithmic identification for one of the proposed groups: 'Getting More Help: Presentation Suggestive of Potential BPD (Guided by NICE Guideline 78)'. This is because clinical experience suggested that young people will rarely be identified as belonging to this group at assessment (when the Current View items our algorithm relies upon were rated). This group is therefore not represented in figures and graphs, and was not part of statistical modelling.

Table E.4 shows the proportion of group membership in the Analysis Sample. The three largest groups are "Getting Advice" (30 %), "Getting Help: Difficulties Not Covered by Other Groupings" (16 %), and "Getting More Help: Difficulties of Severe Impact" (8 %). Note that these three groups are not defined via one particular NICE guidance. These proportions do not present our best estimates of the cluster proportion, since the Analysis Sample is likely to be biased towards periods of contact of shorter duration (see Appendix F). We are reporting these proportions, because these data form the basis of the statistical modelling that follows (which is based on closed and dormant cases only).

Figure E.11 displays proportional cluster membership separately for four different age groups. The age differences in the cluster proportions make clinical sense. Note, for example, the relatively high proportions of children classified as "ADHD" and "Conduct Disorder" in the younger age groups, and the relatively high proportions of children classified within "Depression", "Self-Harm", and "Co-occurring Emotional Difficulties" in the older groups.

Table E3: Algorithm for Grouping Allocation based on Current View Ratings

E3 a: Groups defined by a single "index" presenting problem

Current View Presenting Problem	SHA Self-Harm	PTS PTSD	DEP Depression	OCD OCD/BDD	BIP Bipolar Disorder	ADH ADHD	GAP Generalised Anxiety / Panic Disorder		BEH Conduct Disorder	SOC Social Anxiety	AUT Autism management	EAT Eating Disorders	PSY Psychosis	NEU Neurodev. Assessment
ANXIOUS AWAY CAREGIVERS	≤Self-harm	<Trau-ma	≤mild	<Comp do think	Any	≤mild	≤Anx Gen	<Panic	≤mild	<Anx Soc	≤mild	≤mild	Any	Any
ANXIOUS SOCIAL	≤Self-harm	<Trau-ma	≤mild	<Comp do think	Any	≤mild	≤Anx Gen	<Panic	≤mild	≥mod-erate	≤mild	≤mild	Any	Any
ANXIOUS GENERALLY	≤Self-harm	<Trau-ma	≤mild	<Comp do think	Any	≤mild	≥mod-erate	Any	≤mild	<Anx Soc	≤mild	≤mild	Any	Any
COMPELLED DO THINK	≤Self-harm	<Trau-ma	≤mild	≥mod-erate	Any	≤mild	<Anx Gen	<Panic	≤mild	<Anx Soc	≤mild	<Eat	Any	Any
PANICS	≤Self-harm	≤ Trau-ma	≤mild	<Comp do think	Any	≤mild	Any	≥mod-erate	≤mild	<Anx Soc	≤mild	≤mild	Any	Any
AVOIDS GOING OUT	≤Self-harm	≤ Trau-ma	≤mild	<Comp do think	Any	≤mild	≤Anx Gen	<Panic	≤mild	<Anx Soc	≤mild	≤mild	Any	Any
AVOIDS SPECIFIC THINGS	≤Self-harm	<Trau-ma	≤mild	≤Comp do think	Any	≤mild	≤Anx Gen	<Panic	≤mild	<Anx Soc	≤mild	<Eat	Any	Any
REPETITIVE PROBLEM BEHARS	≤Self-harm	<Trau-ma	≤mild	<Comp do think	Any	≤mild	≤Anx Gen	<Panic	≤mild	<Anx Soc	≤mild	≤mild	Any	Any
LOW MOOD	≤Self-harm	<Trau-ma	≥mod-erate	≤mild	Any	≤mild	≤mild	≤mild	≤mild	≤mild	≤mild	≤Eat	Any	Any
SELF HARM	≥mod-erate	≤mild	<Low Mood	Any	<Ex-tremes Mood	≤mild	<Anx. Gen.	<Panic	≤mild	≤mild	≤mild	Any	Any	Any
EXTREMES OF MOOD	<Self-harm	≤mild	≤mild	≤mild	mod-erate	≤mild	≤mild	≤mild	≤mild	≤mild	≤mild	≤mild	severe OR	Any
DELUSIONAL BELIEF HALLUCINATIONS	≤mild	≤mild	≤mild	≤mild	<Ex-tremes Mood	≤mild	≤mild	≤mild	≤mild	≤mild	≤mild	≤mild	≥mod-erate	Any
DRUG ALCOHOL DIFFICULTIES	Any	Any	Any	Any	Any	Any	Any	Any	Any	Any	≤mild	Any	Any	Any
DIFFICULTIES SITTING STILL CONCENTRATE	≤mild	<Trau-ma	<Low Mood	<Comp do think	Any	≥mod-erate	≤mild	≤mild	≤mild	≤mild	≤mild	≤mild	Any	Any
BEHAVIOURAL DIFFICULTIES	≤mild	<Trau-ma	<Low Mood	<Comp do think	Any	Any	≤mild	≤mild	≥mod-erate	≤mild	Any	≤mild	Any	Any
POSES RISK OTHERS	≤mild	≤mild	≤mild	<Comp do think	Any	Any	≤mild	≤mild	Any	≤mild	Any	≤mild	Any	Any
CARER MANAGENT	Any	Any	Any	Any	Any	Any	Any	Any	Any	Any	Any	Any	Any	Any
TOILET PROBLEMS	≤mild	≤mild	≤mild	<Comp do think	Any	≤mild	≤mild	≤mild	≤mild	≤mild	≤mild	≤mild	Any	Any
TRAUMATIC EVENT	≤mild	≥mod-erate	≤mild	<Comp do think	Any	≤mild	≤mild	≤mild	≤mild	≤mild	≤mild	≤mild	Any	Any
EATING ISSUES	≤mild	≤mild	<Low Mood	<Comp do think	Any	≤mild	≤mild	≤mild	≤mild	≤mild	≤mild	≥mod-erate	Any	Any
FAMILY REL'SHIP DIFFICULTIES	Any	Any	Any	Any	Any	Any	Any	Any	Any	Any	Any	Any	Any	Any
ATTACHMENT CARER PROBLEMS	Any	Any	Any	Any	Any	Any	Any	Any	Any	Any	Any	Any	Any	Any
PEER RELATIONSHIP DIFFICULTIES	Any	Any	≤Low Mood	≤mild	Any	Any	≤mild	<Panic	Any	Any	Any	Any	Any	Any
PERSIST. REL'SHIP DIFFICULT.	≤Self Harm	Any	<Low Mood	≤mild	Any	Any	≤mild	<Panic	Any	Any	Any	Any	Any	Any
DOES NOT SPEAK	≤mild	Any	<Low Mood	<Comp do think	Any	≤mild	≤Anx Gen	<Panic	≤mild	Any	Any	Any	Any	Any
GENDER DISCOMFORT	<Self Harm	Any	Any	<Comp do think	Any	Any	<Anx Gen	Any	<Beha v diffs	Any	Any	Any	Any	Any
UNEXPLAINED PHYSICAL SYMPTOMS	≤mild	Any	≤Low Mood	<Comp do think	Any	≤mild	≤Anx Gen	<Panic	≤mild	Any	Any	≤mild	Any	Any
UNEXPLAINED DEVELOPM. DIF.	≤mild	≤mild	≤mild	≤mild	Any	Any	≤mild	≤mild	≤mild	≤mild	≤mild	Any	Any	≥mod-erate
SELF CARE ISSUES	<Self Harm	Any	≤Low Mood	<comp do think	Any	Any	<Anx Gen	<Panic	≤mild	Any	Any	Any	Any	Any
ADJUSTMENT HEALTH ISSUES	≤mild	Any	Any	Any	≤mild	Any	Any	Any	≤mild	Any	Any	Any	≤mild	Any
Complexity: Pervasive Develop. Disorder	Any	Any	Any	Any	Any	Any	Any	Any	NO	Any	YES	Any	Any	Any
Age	Any	Any	Any	Any	≥ 10 years	Any	Any	Any	Any	Any	Any	≥ 10 years	≥ 10 years	Any

Notes:

Colour key:

Colour	Meaning
Green	"Index problem", or required condition. Example from Self-Harm: Self-Harm must be rated moderate or severe.
Yellow	Exclusion criterion compared to the index problem. Example from Self-Harm: "Anxious away from caregivers" must be rated as less severe or of equal severity as the index problem, Self-Harm.
Red	Absolute Exclusion Criterion. Problem must be absent or mild (where appropriate).
Blue	No restrictions on ratings apply.
Light Green	Additional required condition (age restriction for BIP, EAT, and PSY).

Symbols

≤ ***Less severe or equal severity****. Example: "≤mild" means "Must be rated 'none' or 'mild'".*

< ***Less severe than****. Example: "<Self-Harm" means "must be rated as being less severe than self-harm"*

≥ ***More severe or equal severity****. Example: "≥moderate" means "must be rated 'moderate' or 'severe'"*

NO ***Must be absent****.*

YES ***Must be present****.*

Any **No conditions on ratings apply.**

Table E3: b: Remaining Groups

	ADV: Getting Advice: Signposting and Self-management Advice	DNC: Getting Help: Difficulties Not Covered by Other Groupings	DSI: Getting More Help: Difficulties of Severe Impact	EMO: Getting Help: Co-occurring Emotional Difficulties	BEM: Getting Help: Co-occurring Behavioural and Emotional Difficulties
Does not fit the criteria of any of the groups in Table E3 a (except Neurodevelopmental Assessment)	**YES**	**YES**	**YES**	**YES**	**YES**
Number of presenting problems rated moderate or higher ≤1	**YES**	**Any**	**NO**	**NO**	**NO**
Number of presenting problems rated moderate or higher ≥2 OR Number of presenting problems rated severe =1 AND number of presenting problems rated mod-erate = 0	**NO**	**YES**	**Any**	**Any**	**Any**
Number of presenting problems rated severe ≥2 OR [Number of presenting problems rated moderate or higher ≥ 2 if one of these is from list A AND the child is aged≥10]	**NO**	**NO**	**YES**	**Any**	**Any**
Number of "emotional" problems rated moderate or higher ≥2	**NO**	**Any**	**Any**	**YES**	**Any**
Any "emotional" problem rated moderate or higher AND Behavioural Difficulties rated moderate or higher	**NO**	**Any**	**Any**	**NO**	**YES**
Number of problems from list B rated moderate or higher = 0	**Any**	**Any**	**Any**	**YES**	**YES**

Notes: For the purpose of this table, the complexity factor "Pervasive Developmental Disorder" is counted as a "moderate presenting problem" if present.

List A: Delusional Beliefs/Hallucinations; Eating Issues; Extremes of Mood (severe rating only)

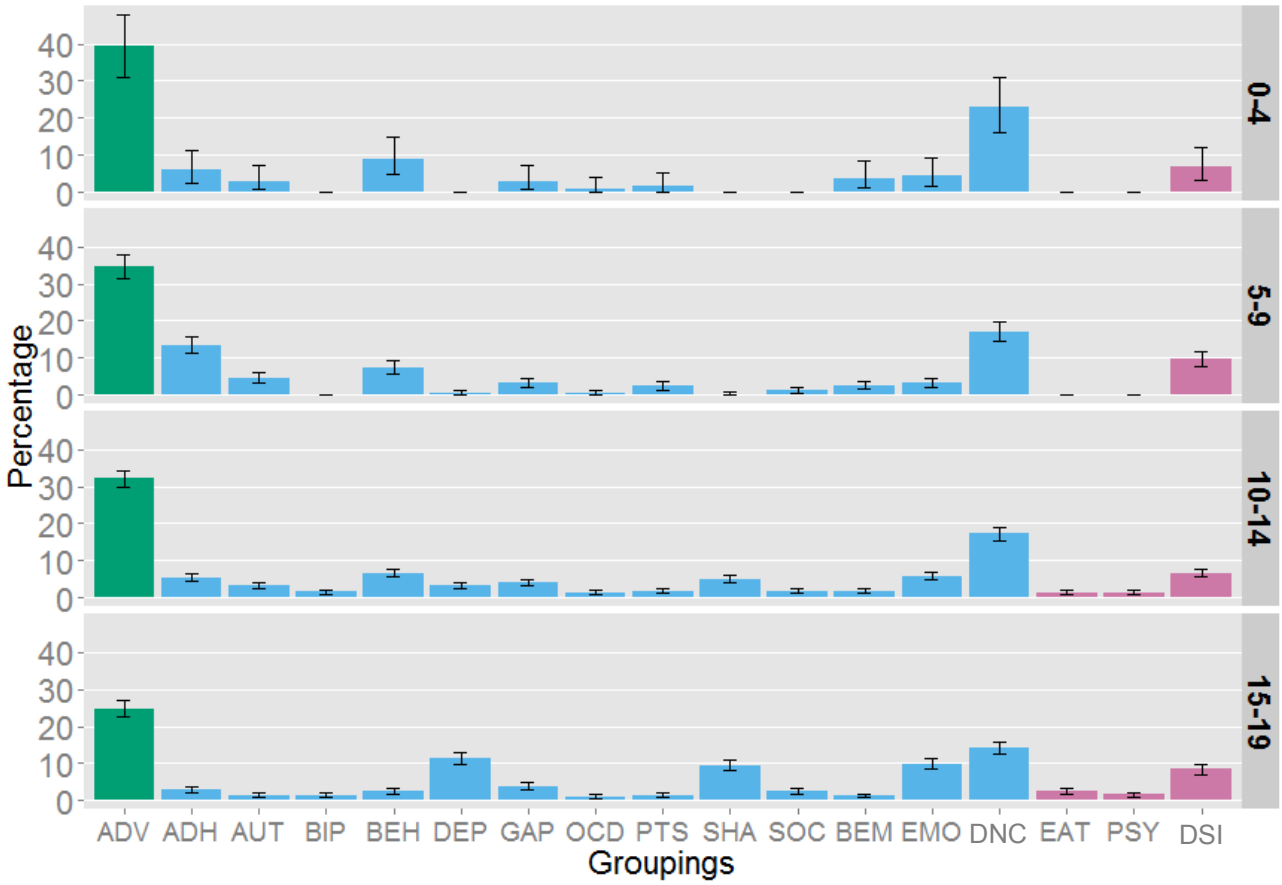
List B: Extremes of mood (Bipolar disorder); Pervasive Developmental Disorders (Autism/Asperger's); Delusional beliefs and hallucinations (Psychosis); Eating issues (Anorexia/Bulimia); Disturbed by traumatic event (PTSD); Self-Harm (Self injury or self-harm); Difficulties sitting still or concentrating (ADHD/Hyperactivity)
"Emotional" presenting problems: Depression/low mood (Depression); Panics (Panic Disorder); Anxious generally (Generalized anxiety); Compelled to do or think things (OCD); Anxious in social situations (Social anxiety/phobia); Anxious away from caregivers (Separation anxiety); Avoids going out (Agoraphobia); Avoids specific things (Specific phobia).

Table E.4: Frequency counts and percentages of Groupings in the Analysis Sample

Grouping	Short Label	Count	Total %
Getting Advice: Signposting and Self-management Advice	ADV	1378	30.1
Getting Help: ADHD (Guided by NICE Guideline 72)	ADH	282	6.2
Getting Help: Autism Spectrum (Guided by NICE Guideline 170)	AUT	127	2.8
Getting Help: Bipolar Disorder (Guided by NICE Guideline 185)	BIP	50	1.1
Getting Help: Behavioural and/or Conduct Disorders (Guided by NICE Guideline 158)	BEH	246	5.4
Getting Help: Depression (Guided by NICE Guideline 28)	DEP	255	5.6
Getting Help: GAD and/or Panic Disorder (Guided by NICE Guideline 113)	GAP	174	3.8
Getting Help: OCD (Guided by NICE Guideline 31)	OCD	46	1.0
Getting Help: PTSD (Guided by NICE Guideline 26)	PTS	77	1.7
Getting Help: Self-harm (Guided by NICE Guidelines 16 and/or 133)	SHA	251	5.5
Getting Help: Social Anxiety Disorder (Guided by NICE Guideline 159)	SOC	83	1.8
Getting Help: Co-occurring Behavioural and Emotional Difficulties	BEM	80	1.7
Getting Help: Co-occurring Emotional Difficulties	EMO	308	6.7
Getting Help: Difficulties Not Covered by Other Groupings	DNC	744	16.3
Getting More Help: Eating Disorders (Guided by NICE Guideline 9)	EAT	63	1.4
Getting More Help: Psychosis (Guided by NICE Guidelines 155 and/or 185)	PSY	49	1.1
Getting More Help: Difficulties of Severe Impact	DSI	360	7.9
	Total	4573	100.0

Note: In addition to the classification shown in the table, 130 children (2.84 % of the sample) were classified as in need of "Neurodevelopmental Assessment"; this group is not mutually exclusive with any others. The group 'Getting More Help: Presentation Suggestive of Potential BPD (Guided by NICE Guideline 78)' is not represented, since there is currently no allocation algorithm for this group.

Figure E.11: Grouping membership in the Analysis Sample, by Age Group



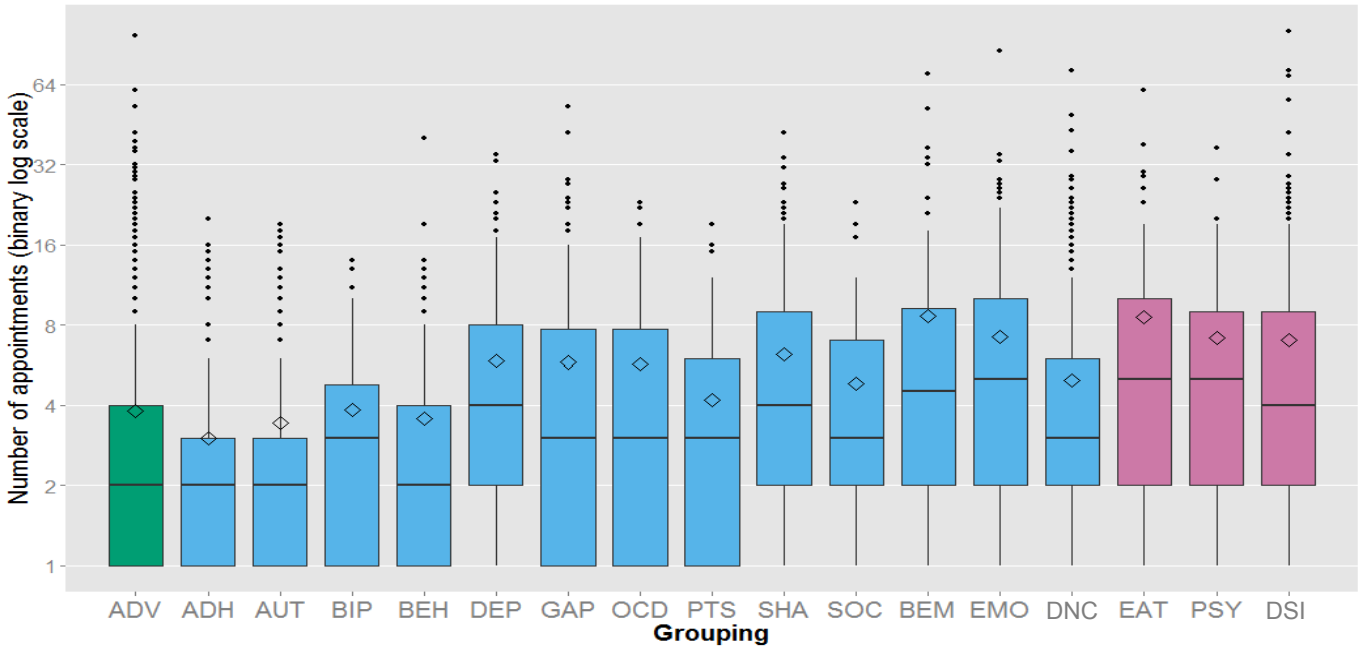
Notes: For definition of grouping labels see Table E.4. The grouping 'Getting Advice: Neurodevelopmental Assessment' is not represented, as it is not mutually exclusive with the remaining groupings. The grouping 'Getting More Help: Presentation Suggestive of Potential BPD (Guided by NICE Guideline 78)' is not represented, since there is currently no allocation algorithm for this group. N = 4573.

Table E.5: Distribution of Number of Appointments by Grouping

Grouping	Mean	Trimmed	SD	Min	Max	Median	P25	P75	COV
ADV	3.82	2.93	5.77	1	97	2	1	4	1.51
ADH	3.02	2.57	3.11	1	20	2	1	3	1.03
AUT	3.43	2.88	3.88	1	19	2	1	3	1.13
BIP	3.84	3.52	3.58	1	14	3	1	4.75	0.93
BEH	3.57	3.05	3.86	1	40	2	1	4	1.08
DEP	5.89	5.25	5.72	1	35	4	2	8	0.97
GAP	5.79	4.73	7.21	1	53	3	1	7.75	1.25
OCD	5.72	5.14	6.36	1	23	3	1	7.75	1.11
PTS	4.17	3.77	3.91	1	19	3	1	6	0.94
SHA	6.22	5.45	6.37	1	42	4	2	9	1.02
SOC	4.83	4.25	4.59	1	23	3	2	7	0.95
BEM	8.64	6.86	11.68	1	70	4.5	2	9.25	1.35
EMO	7.22	6.30	7.96	1	85	5	2	10	1.10
DNC	4.93	4.17	5.77	1	72	3	2	6	1.17
EAT	8.59	7.18	10.69	1	61	5	2	10	1.24
PSY	7.14	6.29	7.45	1	37	5	2	9	1.04
DSI	7.02	5.66	9.68	1	101	4	2	9	1.38
Total	4.96	4.03	6.53	1	101	3	1	6	1.32

Notes: Trimmed: 90 % trimmed mean; SD: Standard Variation; P25: 25th percentile; P75: 75th percentile; COV: Coefficient of Variation [COV = SD/Mean]. For definition of grouping labels see Table E.4. The grouping 'Getting Advice: Neurodevelopmental Assessment' is not represented, as it is not mutually exclusive with the remaining groupings. The grouping 'Getting More Help: Presentation Suggestive of Potential BPD (Guided by NICE Guideline 78)' is not represented, since there is currently no allocation algorithm for this group. N = 4573.

Figure E.12: Distributions of "Number of Appointments" (on binary log scale) by Grouping



Notes: For definition of grouping labels see Table E.4. The graph shows boxplots. The lower end of the box denotes the 25th percentile, the line in the middle of the box denotes the median, and the upper end of the box denotes the 75th percentile. The vertical lines and dots above and below the boxes represent the range.The arithmetic mean is represented by a rhombus. Data are shown on a binary log scale. The grouping 'Getting Advice: Neurodevelopmental Assessment' is not represented, as it is not mutually exclusive with the remaining groupings. The grouping 'Getting More Help: Presentation Suggestive of Potential BPD (Guided by NICE Guideline 78)' is not represented, since there is currently no allocation algorithm for this group. N = 4573.

We now turn to the relationship of the groupings to measured resource use. The data are displayed in Table E5, and illustrated in Figure E12. In general, children who were allocated to the three groups within the "Getting More Help" supergrouping tended to attend a higher number of appointments than children in most other groups, on average. The averages of these three groups are 7.0, 7.1 and 8.6, respectively. Some of the groups within the "Getting Help" supergrouping have means at similar levels, notably the two groups defined by specific co-occurring difficulties (Co-occurring Emotional Difficulties and Co-occuring Behavioural and Emotional Difficulties), with average number of appointments of 7.2 and 8.6, respectively. Children allocated to the "Getting Advice" group tended to attend few appointments on average (with a median of 2.0 appointments and a mean of 3.8 appointments). Four of the groups within the "Getting Help" supergrouping have similarly low averages; these are the groups defined by ADHD, Autism Management, Bipolar Disorder (Extremes of Mood), and Conduct Disorder (with means ranging from 3.0 to 3.8). Seven groups within the Getting Help supergrouping have average numbers of sessions somewhere between the Getting Advice and Getting More Help groups.

There is considerable variation within the groups, in terms of the numbers of appointments attended. This is maybe best illustrated by Figure E12: the boxes indicating the interquartile ranges all overlap. Also, the standard deviations and the ranges of numbers of appointments attended within the groups are high relative to the differences in means and medians between the groups. It is fair to conclude that the variation within groups exceeds the variation between groups. The groups, insofar as membership is decided by the algorithm on the basis of current view ratings alone, are thus internally heterogeneous with respect to resource use. Nonetheless, as

section 5 of this appendix will show, this conceptually-driven grouping provides a better and more reliable prediction of the number of appointments attended than the other two classification methods that we have employed.

One question of interest regarding the relationship between our classification and resource use is whether we need as many as 18 clusters to achieve a given level of resource use prediction, or whether a smaller number of clusters might suffice. We considered four models of classification overall: a three-cluster model, a five-cluster model, a sixteen cluster model, and the eighteen-cluster model described above. These models are summarized in Table E.6. The three cluster model simply distinguishes between “Getting Advice”, “Getting Help”, and “Getting More Help”. This model is nested within each of the three other models. The five-cluster model is the same as the three-cluster model in terms of the categories “Getting Advice” and “Getting Help”, but introduces a distinction within “Getting More Help” into “Eating Disorder”, “Psychosis”, and “Other co-occurring difficulties”. The sixteen cluster model retains this structure for Getting More Help, but introduces additional distinctions within Getting Help (introducing all categories defined in Table E3.a). Finally, the 18-cluster model adds two further categories to the sixteen cluster model: namely the ‘comorbid’ categories “Co-occurring emotional difficulties” and “Co-occurring emotional and behavioural difficulties”.

In the next section, we will investigate how well each of these different models of classification predicts the number of appointments attended, and also compare the conceptual approach to classification to each of the two data-driven approaches.

Table E.6: Description of conceptually derived clustering models tested

Number of clusters	Getting Advice	Getting Help	Getting More Help
3	One cluster: • "Getting Advice"	One cluster: • "Getting Help"	One cluster: "Getting More Help"
5	One cluster: • "Getting Advice"	One cluster: • "Getting Help"	Three clusters: • "Eating Disorder" • "Psychosis" • "GMH with other Co-occurring Difficulties".
16	Two clusters: • "Getting Advice" • "Neurodevelopmental Assessment"	Eleven clusters: • ADHD • Autism • Bipolar • Conduct • Depression • GAD/Panic • OCD • PTSD • Self Harm • Social Anxiety • Other Co-occurring Difficulties	Three clusters: • "Eating Disorder" • "Psychosis" • "GMH with other Co-occurring Difficulties".
18	As described	As described	As described

Note: In the sixteen-cluster model, the clusters ""GH with other co-occurring difficulties" and "GMH with other co-occurring difficulties" contain some members who, in the 18-cluster model would be allocated to either "Co-occurring emotional difficulties" or "Co-occurring emotional and behavioural difficulties".

5 Model Comparison

A mixed-effects negative binomial regression approach was used to compare the classifications with respect to how well they predict the number of appointments, and to explore the effect of contextual problems and complexity factors. Since we included only children who had attended at least one appointment, the raw "number of appointments" variable contained no zeroes. We created a new outcome variable "Number of appointments after initial session" by subtracting one (1) from the raw number of appointments. This is a mere practical convenience, in order to avoid having to fit a zero-truncated model.

The model includes a random effect for "CAMH service" in order to take into account the nested data structure.

$$\log(\mu_{ij}) = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ijk} + u_i \ ,$$

where:

- $Y_{ij} \sim NB(\mu_{ij}, V(\alpha))$ is the number of appointments (after initial session) for the j^{th} child treated by the i^{th} service, with mean μ_{ij} and variance $V(\alpha)$, which depends on the dispersion parameter α;
- β_0 is an intercept term;
- β_k, $k = 1, \ldots, p$, is a vector of slope coefficients corresponding to the p predictor variables $x_1, \ldots, x_p$;
- $u_i \sim N(0, \sigma_u^2)$ is the random intercept term for the i^{th} service, $i = 1, \ldots, 11$;
- the variance function is defined as: $V(\alpha) = \mu + \alpha\mu^2$. (This is called the "NB2 parameterization".)

5.1 Unsupervised Cluster Analysis and Conceptually-derived classification

Table E.7 shows fit indices for classification models derived conceptually and from unsupervised cluster analysis. Models are compared using the fit indices AIC ("Akaike Information Criterion") and BIC ("Bayesian Information Criterion"). AIC and BIC are model quality criteria. Each balances model fit ("log-likelihood") with model parsimony ("number of parameters") in a different way. For each AIC and BIC, a smaller number indicates a better model. The lower value for both AIC and BIC indicates that the best classification model is the theory-derived model with 18 clusters. It is shown to be superior to smaller models posited from theory, as well as to the data driven classifications suggested by unsupervised cluster analysis.

Having accepted the 18-cluster model, we were further interested in investigating the potential effect of complexity factors, contextual problems, and EET issues. We thus added nineteen variables representing current view ratings of these factors to the 18-cluster model. As Table E.7 shows, the lower value for AIC suggests that the model including additional factors is superior to the simple 18-cluster model, while the lower value for BIC prefers the simpler model. The following section describes a detailed analysis of the larger model and draws conclusions regarding the size and importance of effects of complexity factors, context problems, and EET issues.

Table E7: Model comparison: Mixed negative binomial regression

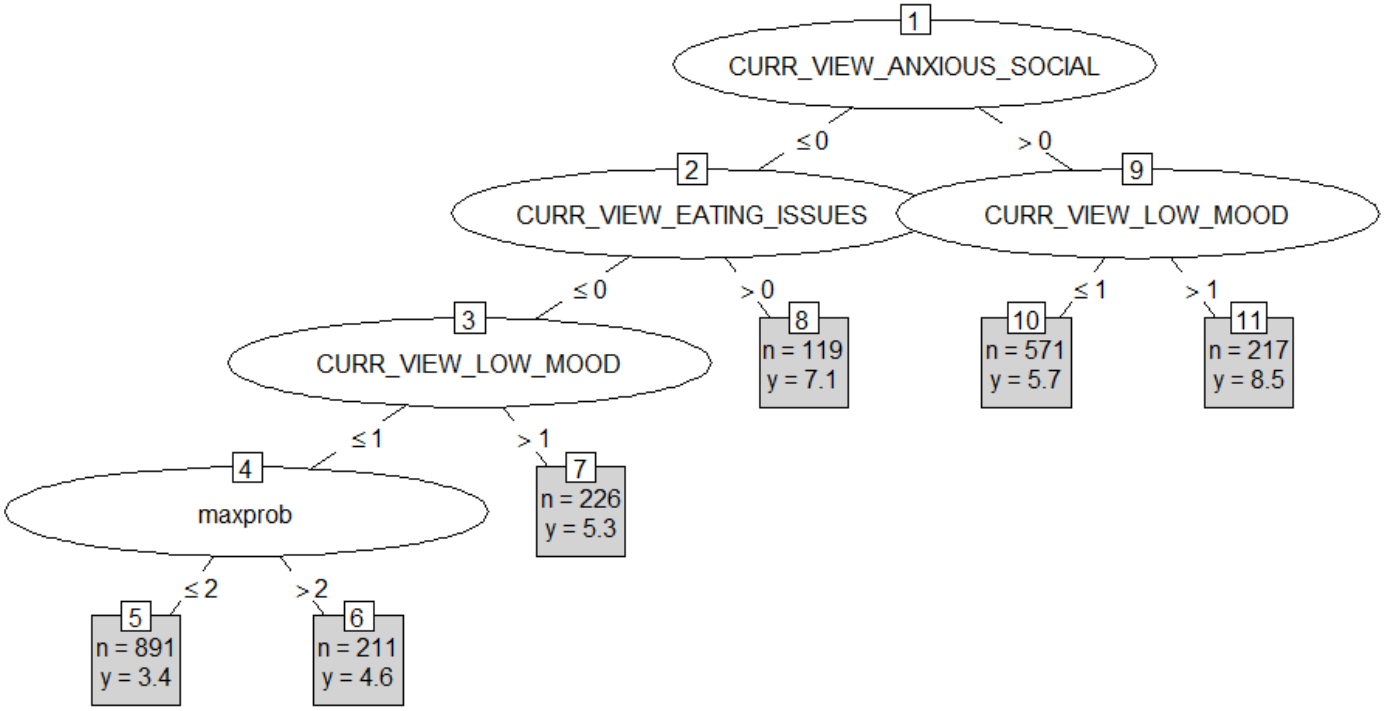
	Model	Log-likelihood	Parameters	AIC	BIC
Null Model	Intercept & Random Effect for Service only	-10968.4	3	21942.8	21962.1
Conceptual	Three Clusters	-10905.9	5	21821.8	21853.9
	Five Clusters	-10904.3	7	21822.6	21867.6
	Sixteen Clusters	-10844.6	18	21725.2	21840.9
	Eighteen Clusters	-10832.3	20	21704.6	**21833.2**
	Eighteen Clusters + Complexity Factors, Contextual Problems & EET Issues	-10806.5	39	**21691.0**	21941.7
Unsupervised Cluster Analysis	UCA: 2 clusters	-10960.9	5	21931.8	21963.9
	UCA: 6 clusters	-10886.5	9	21791.0	21848.9
	UCA: 26 clusters	-10839.3	29	21736.6	21923.0

Note: All models include a random effect for service. AIC: Akaike Information Criterion. BIC: Bayesian Information Criterion. Dependent Variable: Number of appointments.
AIC = -2 × LL + k × 2; BIC = -2 × LL + k × ln(n), where LL is the log-likelihood, k is the number of parameters, and n is the sample size.
In Unsupervised CA (UCA), a small group of cases had no problems rated as present and was excluded from CA, but is included in the models as a separate group.

5.2 Supervised Cluster Analysis and the 18-cluster conceptually-derived model

Regression trees derived from Supervised Cluster Analysis (SCA) were compared with theory-derived classification using mixed-effects negative binomial regression, with the same model structure as for the comparison of conceptually-derived classification with Unsupervised Cluster Analysis. Each regression tree, grown on a random development sample, was tested on its associated test sample (the other half of the data not used in growing the tree). For each random split of the data, the models with the highest correlation R^2 and outlier-resistant R^2 were then tested to determine how well they fit the other half of the sample. This means that a total of 20 models were tested (ten optimal correlation R^2 models, ten optimal outlier-resistant R^2 models). An example of one of the regression trees obtained is presented in figure E.12.

Figure E.12: Example SCA Regression Tree



Note: Current View Ratings: 0 = none, 1 = mild, 2 = moderate, 3 = severe. the variable "maxprob" indicates the maximum rating among the 30 presenting problems; e.g. "maxprob >2" indicates that at least one problem was rated 3 (i.e. "severe").

It will be useful to briefly explain in detail the classification implied by the tree in Figure A.8. This model suggest that there are 6 clusters, from right to left:

- 217 patients, attending 8.5 sessions on average, presenting with social anxiety rated at mild or above, and low mood rated as moderate or severe;
- 571 patients attending 5.7 sessions on average, presenting with social anxiety at mild and above, but low mood at mild or below;
- 119 patients attending 7.1 sessions on average, without social anxiety but with an eating issue rated at mild or above;

- 226 patients attending 5.3 sessions on average, presenting without social anxiety or eating issues, but with low mood rated at moderate or severe;
- 211 patients attending 4.6 sessions on average, presenting without social anxiety, eating issues, with low mood rated at mild or below, but at least one other problem rated as severe;
- and a group of 891 patients attending 3.4 sessions on average, presenting without social anxiety, eating issues or low mood above mild, and have no problem rated severe.

The fit for each model, as well as the fit for the conceptually derived model, was examined using the test samples. The fit indices for these models are presented in table E.8. The first thing to note about the models presented here is the tendency towards a lower number of clusters in the SCA derived models compared to the conceptually derived model. This is also reflected once again in the relative level of fit; there is a tendency for the AIC criteria to select the conceptually derived model as the best fitting, while the BIC tends to favour the smaller, SCA derived models, and almost exclusively the correlation R^2 determined models (since these tend to be smaller).

Table E8: Model comparison: Mixed negative binomial regression

	Log-likelihood	Parameters	AIC	BIC
Test sample 1				
R^2 model	-5329.78	10	10679.56	**10736.68**
Kval model	-5324.95	19	10687.42	10796.42
Conceptually derived model	-5301.99	20	**10643.98**	10758.21
Test sample 2				
R^2 model	-5382.21	11	10786.42	**10849.25**
Kval model	-5375.53	17	10785.06	10882.16
Conceptually derived model	-5371.55	20	**10783.10**	10897.33
Test sample 3				
R^2 model	NA	NA	NA	NA
Kval model	-5378.45	14	10784.90	**10864.86**
Conceptually derived model	-5363.37	20	**10766.74**	10880.97
Test sample 4				
R^2 model	-5327.07	9	10672.14	**10723.54**
Kval model	-5311.88	22	10667.76	10793.41
Conceptually derived model	-5292.57	20	**10625.14**	10739.37
Test sample 5				
R^2 model	-5314.40	10	10648.80	**10705.92**
Kval model	-5304.63	16	10641.26	10732.64
Conceptually derived model	5294.28	20	**10628.56**	10742.72
Test sample 6				
R^2 model	-5373.93	12	10771.86	**10840.40**
Kval model	-5371.72	20	10783.44	10897.67
Conceptually derived model	-5365.02	20	**10770.04**	10884.27
Test sample 7				
R^2 model	-5312.25	20	10664.50	10778.73
Kval model	-5311.78	22	10667.56	10793.21
Conceptually derived model	-5298.92	20	**10637.84**	**10752.07**
Test sample 8				
R^2 model	-5334.34	13	10768.93	**10768.93**
Kval model	-5331.00	20	10816.23	10816.23
Conceptually derived model	-5315.32	20	**10670.64**	10784.87
Test sample 9				
R^2 model	-5419.21	9	10856.42	**10907.82**
Kval model	-5407.37	20	10854.74	10968.97
Conceptually derived model	-5390.77	20	**10821.54**	10935.77
Test sample 10				
R^2 model	-5325.80	8	**10667.60**	**10713.29**
Kval model	-5313.22	24	10674.44	10811.52
Conceptually derived model	5315.93	20	10671.86	10786.09

Note: All models include a random effect for service. AIC: Akaike Information Criterion. BIC: Bayesian Information Criterion. Dependent Variable: Number of appointments.
AIC = -2 × LL + k × 2; BIC = -2 × LL + k × ln(n), where LL is the log-likelihood, k is the number of parameters, and n is the sample size.
The model fit indices for the conceptually derived model are not identical, as each model is fitted to the test sample that the SCA model is also being fitted to.
The r^2 model for the 3rd random split did not converge for the negative binomial regression model.
For each data split, the best model according to each criterion is highlighted.

Another key observation to make is the degree of inconsistency in the models derived using the SCA method. Three of the better fitting examples are presented below in figures E.13a, E.13b and E.13c. While there is a tendency for the SCA method to result in smaller (and potentially more parsimonious) models, only a few current view items are selected in each model; low mood, social anxiety, eating issues and panics are the four most common parameters selected by the SCA method, but beyond these, each model has a tendency to select different parameters from a relatively wide range of criteria to separate out the smaller groups 'lower down' in the selection. The practical consequence of this is that, while models derived using the SCA method do show a better fit to the data than the conceptually derived model (at least according to the BIC index), there is no real way to be certain which of the derived models would be the most reliable and best fitting to population level data. This is likely to a result of the tendency for the SCA method to 'overfit' the data; each time the sample was randomly split in half (stratified for age and gender), different groups of relatively small but very specific groups of patients will have been selected, which will then influence the iterative 'splitting' process which separates the sample. This is evident even in the smaller models below; even the first split is not caused by the same predictor in each model, and as the model progresses down the tree, there is variability in which predictors are chosen to make further splits depending upon the exact make-up of the random sample upon which they were developed. This is even more pronounced in the larger SCA derived models; while there is a common set of predictors that frequently appear near the tops of the trees (low mood, eating issues, social anxiety), the smaller and smaller disparate groups begin to vary much more as the number of small sub-groups is identified. If the SCA method provided a reliable way of determining disparate groups of patients, there should be a greater level of consistency in the criteria that are used to determine the groups. Consequently, it would seem that the conceptually derived model of patient classification would offer a more reliable and meaningful method of patient grouping than the SCA method.

Figure E.13a. R^2 derived regression tree with 8 clusters, from the 5th sample random test sample.

Figure E13b. R^2 derived regression tree with 6 clusters, from the 10th random test sample.

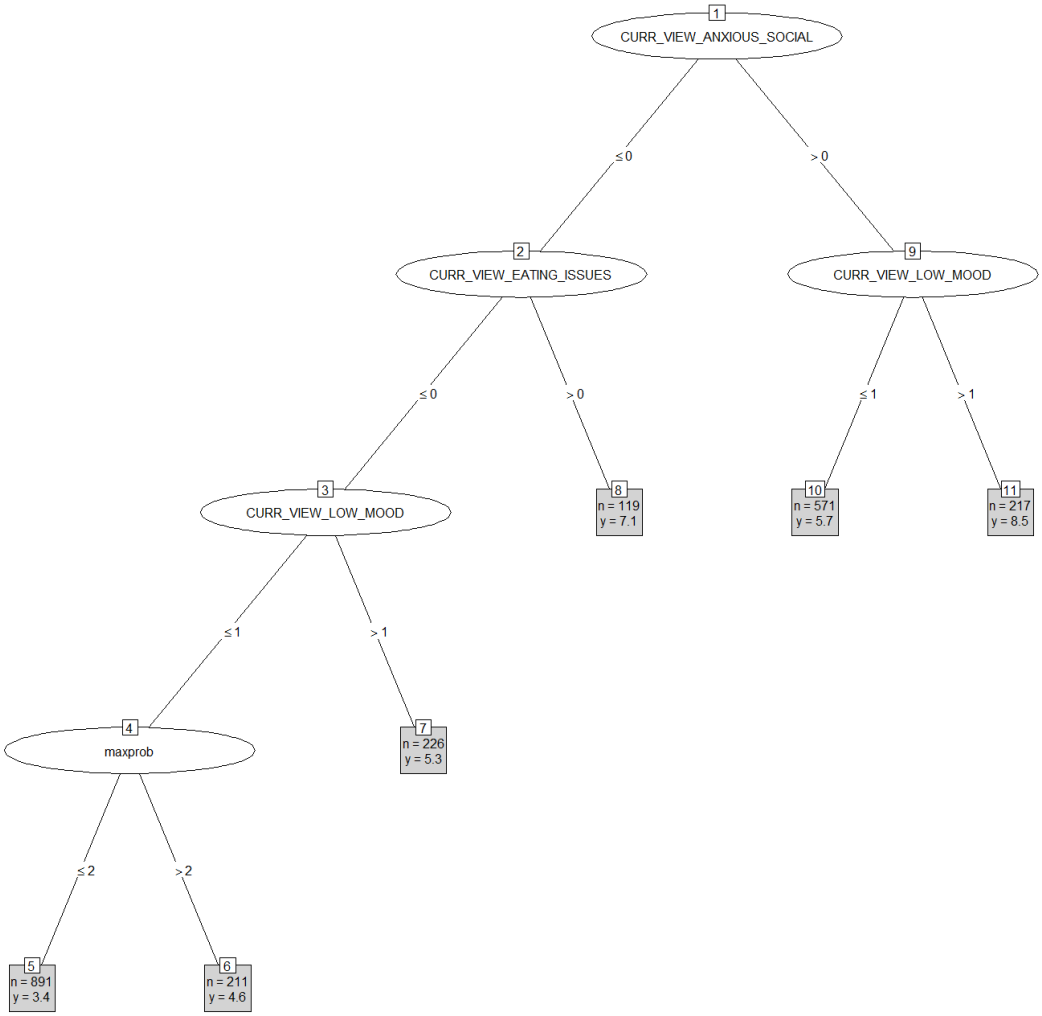
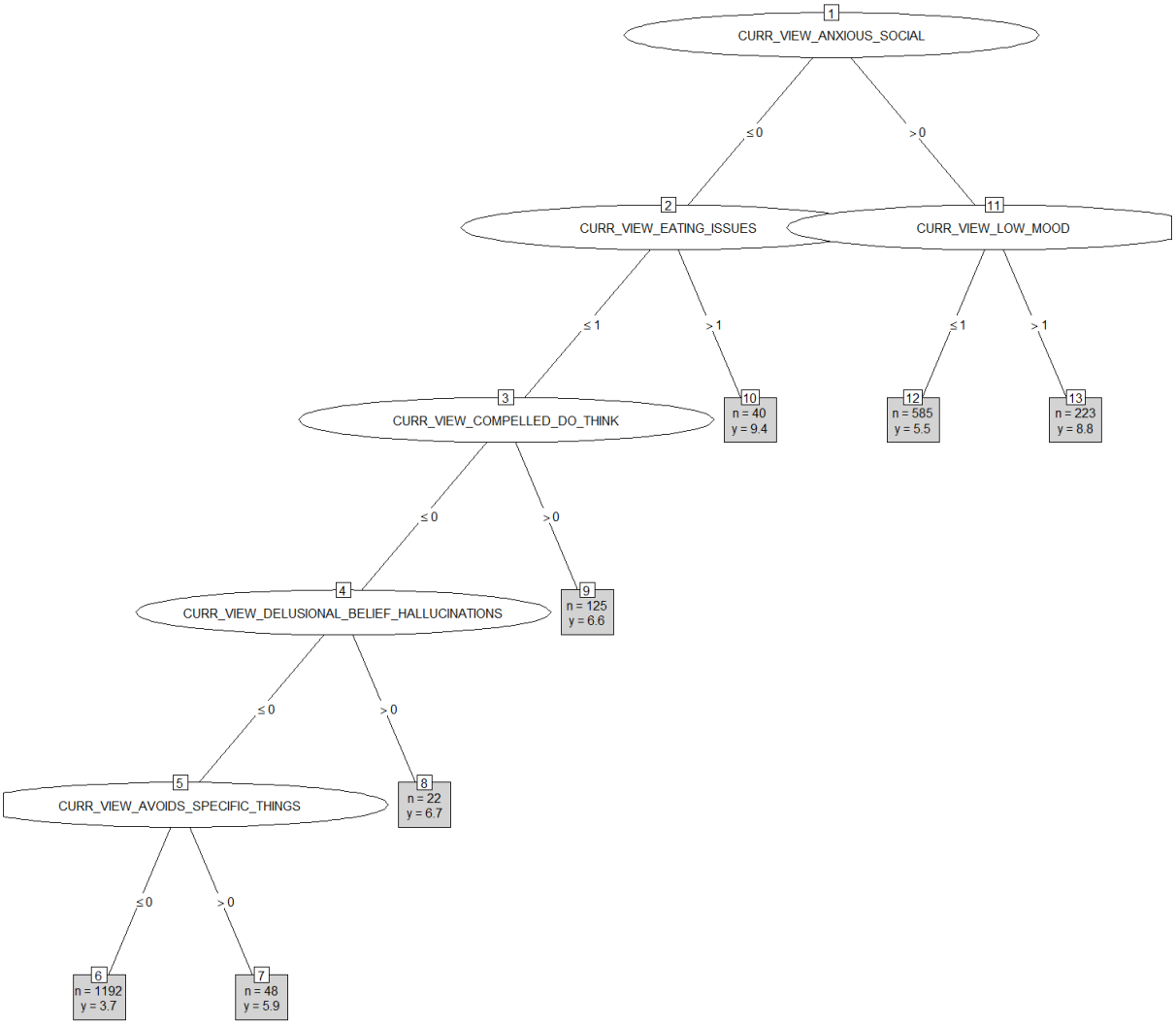


Figure E13c. R^2 derived regression tree with 7 clusters, from the 4th random test sample.



6 Final Model

This section presents details of the 18-cluster model with complexity, context, and EET factors. Model estimates are presented in Table E9. Regression coefficients and their confidence intervals are illustrated in Figure E.14. Coefficients represent estimates of the effect of being in a given cluster, or having a particular additional factor, on the log number of appointments until case closure, relative to being in the "Getting Advice" cluster with no additional factors. The coefficient estimates for the 17 clusters in the 18-cluster model *without* additional factors (judged to be the better model according to BIC – see previous section) are very similar to those in the model presented, suggesting that the additional factors have little, if any, effect on the predicted average number of appointments within each cluster. We present the model with additional factors, because this helps us illustrate one of our main points: that we did not find evidence for any strong effect of contextual, complexity, or EET factors on resource use.

Table E.9 and Figure E.14 show that the largest estimated effects on resource use are associated with the three clusters that make up the supergrouping "Getting More Help" and seven of the thirteen clusters that form the "Getting Help" supergrouping. Specifically, the largest predicted effects are for the clusters "Eating Disorders", "Co-occurring Behavioural and Emotional Difficulties", "Self Harm", "Depression", "Psychosis", "Co-occurring Emotional Difficulties", "OCD", "Getting More Help: Difficulties of Severe Impact", "General Anxiety and Panic Disorder", and "Getting Help: Difficulties Not Covered by Other Groupings". In general, this result confirms our expectation that the children classified within one of clusters from the "Getting More Help" supergrouping would tend to have the highest average resource use, while those in groups from the "Getting Help" supergrouping would tend to have the next highest, with children in the "Getting Advice: Signposting and Self-Management" grouping (the reference category in this model) predicted to have the lowest average resource use. Note, however, that most of the confidence intervals for these ten strongest estimates have substantial overlap with one another, which illustrates that there is little basis, from our data, to derive a definite rank order of "average number of appointments" for these ten groups.

Six groups from the "Getting Help" supergrouping – "Social Anxiety", "Conduct Disorder", "ADHD", "Autism Management", "Bipolar Disorder", and "PTSD" – are not found to have an average number of sessions that differs significantly from the "Getting Advice" cluster. Neither do we have evidence, from these data, that membership in the cluster "Neurodevelopmental Assessment" (which is not mutually exclusive with other clusters) adds to resource use, once the assigned cluster membership among the 17 mutually exclusive clusters has been taken into account.

6.1 Contextual Problems, Complexity Factors, and EET Issues

We simultaneously entered nineteen additional factors – representing contextual problems, complexity factors, and EET issues – to the 18-cluster model. We made no specific predictions about which additional factors should have an influence on resource use. In the absence of such specific predictions, then under the null hypothesis that none of the nineteen additional factors has an effect on resource use after controlling for cluster membership, we would expect on average about one coefficient to be statistically significant (at the conventional 5 % level of significance) due to random error. In our analysis, we found five such 'statistically significant' effects. The estimates are presented in Table A8 and illustrated in Figure A9.

The complexity factor "Living in financial difficulties" and EET Attendance Difficulties are predicted by the model to contribute to higher average resource use (after controlling for all other variables in the model). Conversely, contextual problems at "School, Work or Training" or with "Service Engagement", as well as the complexity factor "Serious physical health issues", predict lower average resource use (after controlling for all other variables in the model). Although "statistically significant" in the sense that their 95 % confidence interval did not include the value 0, the estimated size of the effect was relatively small for each of these five factors, compared to the effects of the ten clusters that are predicted to be most resource-intensive. Overall, then, our inspection of estimates effect sizes confirms the conclusion that we drew from the model comparison presented in the previous section, namely that there is little evidence, from our data, that contextual, complexity or EET factors have an important influence on resource use, once cluster membership has been taken into account.

As can be seen from the model specification presented in Section 5, we estimated not only fixed effects pertaining to clusters and additional factors, but also the variation in resource use that is due to differences between services, i.e. the random effects variance. This was estimated to be 0.237, so that the random effects standard deviation (the square root of the variance) is estimated to be 0.487. This indicates that there is a large variation between services relative to the effects of different clusters. This finding will be illustrated in the following section.

Table E.9a: Estimates from a mixed negative binomial regression.
Predicated Variable: "Number of Appointments"

	Coefficient Estimate	Standard Error	95 % Confidence Interval for the Coefficient	
			Lower Bound	Upper Bound
Intercept	1.036	0.154	0.734	1.338
Groups (Ref: ADV)				
ADH	-0.154	0.096	-0.342	0.034
AUT	0.046	0.132	-0.213	0.304
BIP	0.011	0.198	-0.377	0.399
BEH	-0.019	0.099	-0.212	0.174
DEP	0.697	0.091	0.518	0.876
GAP	0.386	0.107	0.176	0.595
OCD	0.645	0.197	0.260	1.031
PTS	0.018	0.162	-0.299	0.335
SHA	0.738	0.091	0.560	0.915
SOC	0.244	0.151	-0.052	0.541
BEM	0.881	0.151	0.586	1.177
EMO	0.675	0.085	0.509	0.840
DNC	0.319	0.065	0.192	0.445
EAT	1.007	0.165	0.684	1.331
PSY	0.674	0.193	0.296	1.052
DSI	0.639	0.085	0.473	0.805
NEU and Other Factors				
NEU	-0.193	0.126	-0.441	0.055
LAC	0.035	0.117	-0.194	0.263
YOU	-0.242	0.129	-0.495	0.012
LD	0.138	0.094	-0.045	0.322
PHY	-0.210	0.092	-0.391	-0.029
NEI	-0.174	0.129	-0.427	0.078
PRO	-0.074	0.133	-0.334	0.186
CIN	-0.102	0.086	-0.270	0.067
REF	0.069	0.308	-0.536	0.673
WAR	-0.258	0.352	-0.949	0.432
ABU	0.089	0.072	-0.053	0.230
PAR	0.064	0.054	-0.042	0.170
JUS	-0.019	0.130	-0.274	0.235
FIN	0.221	0.093	0.040	0.402
HOM	0.005	0.025	-0.044	0.054
SCL	-0.057	0.026	-0.108	-0.006
COM	0.031	0.033	-0.033	0.095
ENG	-0.102	0.041	-0.183	-0.022
ATE	0.142	0.029	0.085	0.198
ATA	-0.041	0.03	-0.100	0.018

Notes:
N = 4573; Loglikelihood = - 10806.5, AIC = 21691.0, BIC = 21941.7,
Dispersion parameter estimate: $\hat{\alpha} = 0.687$ *(s.e. = 0.019),*
Random effect variance estimate: $\widehat{\sigma_u^2} = 0.237$ *.*
See Table E9b for legend.

Table E9.b: Legend to Table E9.a and Figure E14

Complexity Factors
ABU: Experience of Abuse or Neglect
CIN: Child in Need
FIN: Living in financial difficulty
JUS: Contact with Youth Justice System
LAC: Looked after Child
LD: Learning Disability
NEI: Neurological Issues
PAR: Parental Health Issues
PHY: Physical Health Problems
PRO: Current Protection Plan
REF: Refugee or asylum seeker
WAR: Experience of War, Torture or Trafficking
YC: Young Carer

Contextual Problems
ENG: Service Engagement
COM: Community Issues
HOM: Home
SCL: School, Work or Training

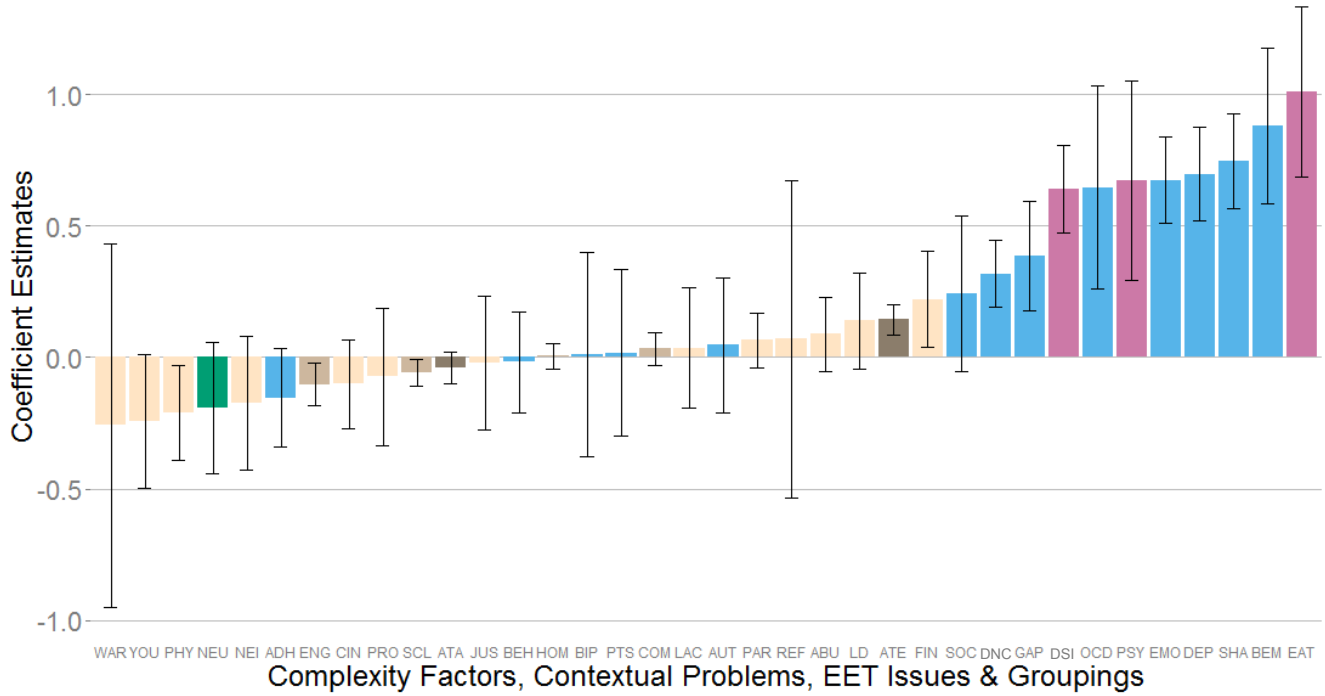
Education/Employment/Training
ATA: Attainment Difficulties
ATE: Attendance Difficulties

Groupings: Getting Advice
ADV: Getting Advice: signposting/self-management
NEU: Neurodevelopmental Assessment

Groupings: Getting Help
ADH: ADHD
AUT: Autism
BIP: Bipolar Disorder (moderate)
BEH: Antisocial Behaviour / Conduct Problems
DEP: Depression
GAP: Generalized Anxiety or Panic Disorder
OCD: Obsessive Compulsive Disorder
PAN: Panics
PTS: PTSD
SHA: Self Harm
SOC: Social Anxiety
BEM: Behavioural and Emotional Difficulties
EMO: Co-occuring Emotional Difficulties
DNC: Difficulties Not Covered by Other Groupings

Groupings: Getting More Help
EAT: Eating Disorder
PSY: Psychosis
DSI: Difficulties of Severe Impact

Figure E.14: Estimates from a mixed negative binomial regression.
Predicated Variable: “Number of Appointments”



Notes: Reference category: “Getting Advice: Signposting and Self-management”. Coefficients represent estimates the effect of being in a given cluster, or having a particular additional factor, relative to being in the “Getting Advice: Signposting and Self-management” cluster with no additional factors. Error bars show 95 % confidence intervals. See Table E9b for legend.

6.2 Model Interpretation

We visualize the model results by taking three "example cases", i.e. idealized children with the following characteristics:

Child A is a member of the Eating Disorder cluster, Child B is a member of the GAD/Panic cluster, and Child C is a member of the Getting Advice cluster. So Child A would be expected to have relatively high resource use, Child C would be expected to have relatively low resource use, and Child B would be expected to be somewhere in the middle between the two others, with respect to resource use.

We now imagine that each of these three idealized children presents at each of three services: one service with relatively low average service provision, one service with approximately average service provision, and one service with relatively high service provision. The number of sessions a child is predicted to attend before case closure depends on (a) the child's characteristics (here: their cluster membership), and (b) the kind of service the child presents at.

For the purpose of illustration, we selected three services according to their estimated effects on the average number of sessions, taking those whose best linear unbiased estimates (BLUP) of the random intercepts (u_j) represented the 10^{th}, 50^{th} (approximately[3]), and 90^{th} percentile. Since we had eleven services overall, this meant that we selected the services with the second lowest, approximate median, and the second highest intercept. We suggest that these three services provide an indication of the diversity in service provision in CAMHS, and can respectively represent services with low, average, and high levels of service provision, respectively.

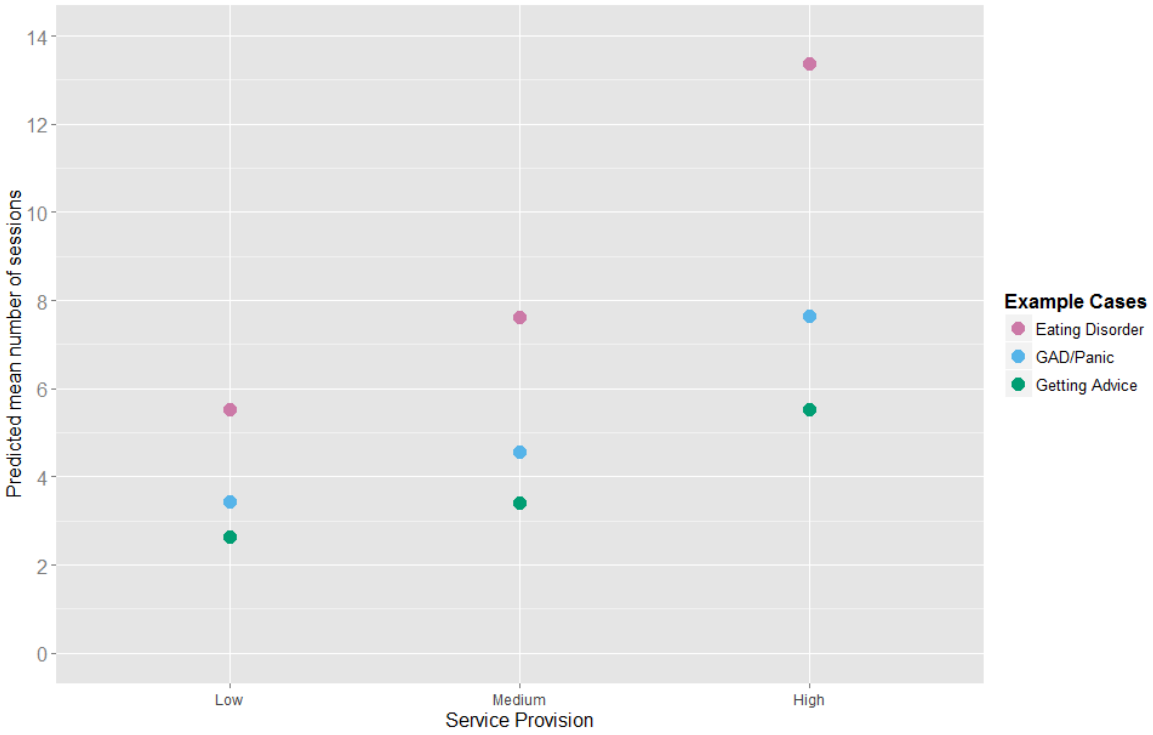
Figure E15 presents the predicted average number of sessions for the three example clusters, contingent on the level of treatment provision in the service attended. The figure illustrates that the number of appointments a child is predicted to attend is influenced at least as much by the type of service the child presents at, as it is by the child's characteristics. Thus, the predicted average number of sessions for children classified as "Getting Advice" varies between about 2.6 and 5.5 between the three services, while the predicted average number of sessions for children classified as "Eating Disorder" varies between about 5.5 and 13.4. That is, a high provision service is predicted to provide about as much resources to children with relatively small estimated need as a low provision services is predicted to provide for children with a very high estimated need.

Note that the model assumes that services don't differ in the way they distribute resources to different types of children, i.e. the model does not allow for service-level random coefficients of the cluster indicators. This represents of course a limitation of the model. For example, services with relatively small funding may well decide to concentrate resources on children with the highest need, which may not be a necessity for more generously funded services.

[3] The median service happened to be small in terms of the number of cases in our analysis sample (n = 114), and did not have any children in the Eating Disorder cluster. We therefore instead selected the service whose random effect estimate was next closest to the median (and that contributed a larger number of cases, n = 1451).

Figure E16 provides a check on this assumption. It presents the observed average number of sessions attended by children in the three clusters, within the three selected services. We see that the pattern of both cluster and service differences is similar to the one predicted by the model. In the low provision service, children in the “Generalized Anxiety/Panic” cluster appear to attend approximately the same average number of sessions as those in the “Getting Advice” cluster – possibly an indication of reserving a relatively high proportion of scarce resources for children with higher need. Once again, the number of appointments that a child can expect to attend is influenced at least as much by the service that the child presents at, as it is influenced by the characteristics of the child (as far as we have been able to measure them).

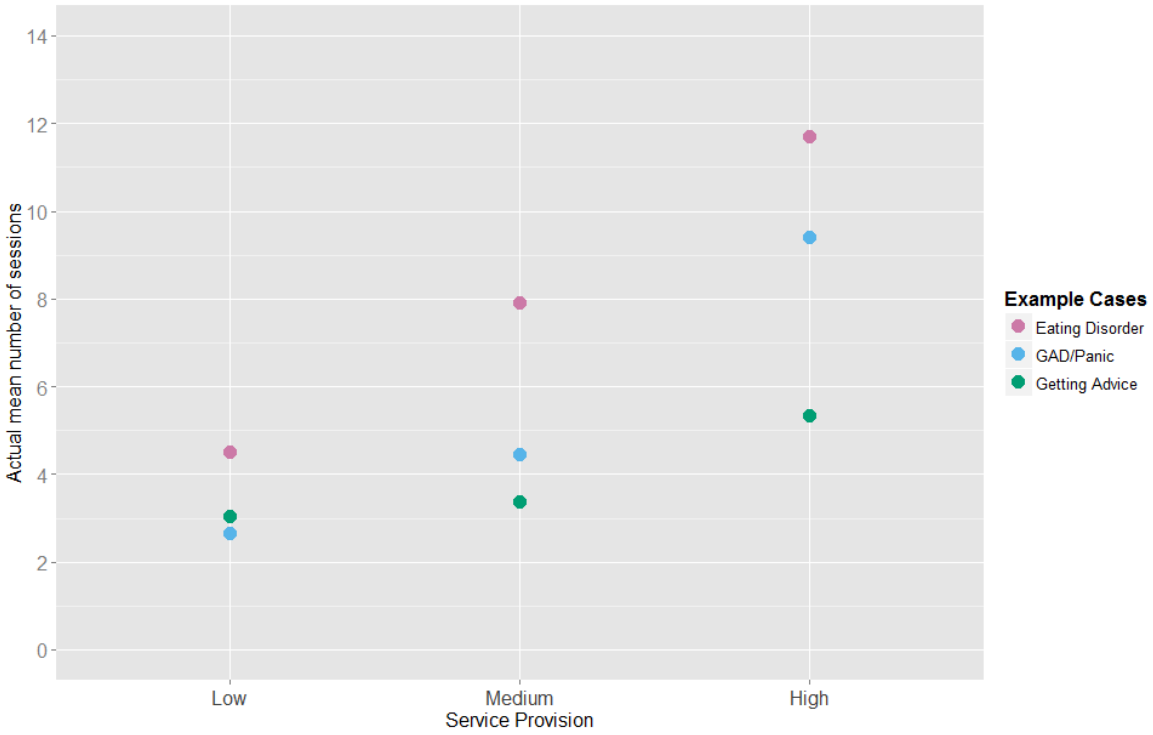
Figure E15 Predicted average number of sessions for three clusters, for three different CAMH services



A note on interpretation: the model predicted averages, as well as the actually observed averages, presented in Figures E15 and E16 are likely to be underestimates of true averages, because of the short observation period, which means that shorter periods of contact are overrepresented in the sample. Thus the reported (predicted and observed) averages should not straightforwardly be taken as estimates of population averages. Similarly, the short observation period may also have somewhat distorted the comparison between the three groupings, because we would expect cases with Eating Disorder to feature a higher proportion of ‘long duration’ compared to cases classified as GAD/Panic or Getting Advice. That means, the differences in means between the three groups may be larger in the population than either predicted or observed in our study. We don’t have a reason to think that the size of service-level differences would be strongly affected by the observation period. However, some of the differences in average resource use between services may be due to data quality (i.e. the degree to which treatment activity was recorded fully and accurately), rather than real differences in treatment provision. We recommend further efforts

to ensure accurate and complete data collection in CAMHS, which would allow more precise estimates of a service's treatment provisions for different types of patients.

Figure E16 Observed average number of sessions for three clusters, for three different CAMH services



Note: Sample sizes for the three services are displayed in Table E10.

Table E10: Sample sizes for Figure A11

	ADV	GAP	EAT	Total N in service (all clusters)
Low	67	9	6	293
Medium	411	46	26	1451
High	135	36	7	445

7 Sensitivity Analysis

7.1 Treatment Cost Weights

We considered that the number of appointments attended was an imperfect indicator of resource use, insofar as the cost of direct treatment activity depends not only on the number of appointments, but also on their duration, and on the types and numbers of staff present. Using information about salary and other costs associated with different professions working in CAMHS, we should in principle be able to estimate the staff costs required for each appointment. Unfortunately, the variables recording appointment duration and staff presence featured missing values, so that relative staff costs could be derived only for about half of overall appointments. This was the reason that we based our main analysis on the number of appointments only.

In this section, we present a sensitivity analysis, investigating the relationship between our casemix classification and estimated total staff costs for each period of contact. We use conditional mean replacement and multiple imputation of missing values to estimate staff costs in the absence of information on duration or staff presence, or both. The purpose of the sensitivity analysis is to see whether our conclusions regarding cluster differences in resource use change when we use relative staff costs as our indicator for resource use, rather than the number of appointments.

7.1.1 Measures

Duration
The duration of appointments was recorded in minutes.

Staff Presence
For each of fifteen categories of staff, the number of individuals present at the appointment was recorded. A list of these categories is provided in the first column of Table E11.

Staff Weights

Estimates of staff salary costs were made from a survey of participating CAMHS services, who were asked to provide salaries for staff that have direct contact with CAMH service users. Estimated salaries are displayed in the second column of Table E11. These estimates represent averages for each grade of staff, where first an average was computed for each responding service, and then the (unweighted) average was taken over all responding services. The numbers represent salaries only, and do not take into account any of the additional employer costs (such as national insurance, pension contributions, and so forth). We assumed that these additional employer costs are roughly proportional to basic salaries. Since our aim is to

estimate *relative* costs of different staff categories only, rather than derive precise estimates of treatment costs, we felt we were justified in working with basic salary estimates. The figures were also compared with a national pay scales file.

Cost weights for different professional categories were then computed by setting the lowest estimated salary equal to 1, and expressing salaries of other staff categories as multiples of this lowest average salary.

Table E11: Reported salaries and assigned weights by profession

Professional Category	Average Reported Salary	Weight
Medical Professional	73,686	3.75
Child and Adolescent Psychotherapist	50,559	2.57
Psychologist	41,045	2.09
Family Therapist	40,043	2.04
Counselling	35,542	1.81
Educational Psychologist	35,542	1.81
Other Qualified Staff	34,136	1.74
Other Therapy Qualified Professional	32,755	1.67
Creative Therapist	32,184	1.64
Social Worker	31,667	1.61
Nurse	31,367	1.60
Primary Mental Health Professional	29,887	1.52
Occupational Therapist	29,211	1.49
Other Educational Professional	20,480	1.04
Other Unqualified Staff	19,640	1.00

Note: Weights were computed by dividing all average salaries by the lowest average, i.e. by dividing the average salaries by 19,640.

We then computed a staff weight for each appointment, by multiplying the number of professionals of each category that were recorded as present at the appointment with that category's weight, and summing over all professional categories. For example, if one psychologist was present at an appointment, this appointment was designed a staff cost weight of 2.09 (from Table …). If an appointment was attended by one psychiatrist and two nurses, the resulting staff cost weight was 1 × 3.75 + 2 × 1.60 = 6.95, and so forth.

Finally, to take into account both the duration of appointments and the staff presence, the total relative cost for a POC is computed as the sum of the products of appointment duration by appointment weight, i.e.

$$c_j = \sum_{i=1}^{n_j} (d_{ij}\, w_{ij}),$$

where:
c_j is the total relative cost for the j^{th} POC;
d_{ij} is the duration of the i^{th} appointment of the j^{th} POC, where duration was measured in hours;
w_{ij} is the staff weight computed from the number and types of staff present at the i^{th} appointment of the j^{th} POC; and
n_j is the number of appointments within the j^{th} POC.

7.1.2 Missing Data

The analysis sample consisted of 4573 POCs with altogether 22676 appointments.
We dealt with missing values in two stages:

In Stage 1, "Conditional Mean Replacement", we considered POCs that had information on duration and/or staff costs on some, but not all of their appointments. For those POCs, we filled in the missing values as the average of all observed values within that POC. For example, if a POC had six recorded appointments, but had duration information on only four of them, the two missing durations were each imputed to be equal to the average of the four observed durations.

There were 22676 appointments overall. Of these, 22119 had valid information on duration.[4] Conditional mean replacement, as described above, allowed us to impute durations for 538 appointments, so that the total number of appointments with (observed or imputed) duration information was 22657 after Stage 1 (99.9 % of all appointments). Analogously, 11,970 appointments had valid information on staff weights. Conditional mean replacement allowed us to impute staff weights for 733 additional appointments, so that the total number of appointments with (observed or imputed) information on staff weights was 12703 after Stage 1 (56.0 % of all appointments).

In Stage 2, "Multiple Imputation", we used multiple imputation of missing values (Rubin 1987; Carpenter & Kenward 2013) to impute missing treatment costs for POCs based on a statistical model. Sixteen POCs (0.3 %) had no information about the duration of any of their appointments, while 2145 POCs (46.9 %) had no information about staff present at appointments. Total weights could be computed for 2428 POCs (53.1 %), who had both duration and staff cost information.

For the purpose of the imputation model, we aggregated the appointment data over all POCs. For each POC, we computed the "total duration" as the sum of the durations of all appointments. We

[4] We set durations that were recorded to be 0 (but where the appointment was recorded as attended) to missing. We also regarded durations as missing if the recorded duration exceeded three hours.

also computed the "total staff weight" as the sum of all staff weights. We then log-transformed the observed treatment costs (so that their distribution approximated the normal) and specified a linear regression with the following predictors: the log-transformed total durations, the log-transformed total staff weights, the number of sessions, the cluster membership, and the Current View ratings of all complexity, contextual and EET factors. We used the "multiple imputation by chained equations" (mice) approach in order to account for the fact that two of the predictors of the total weights (total duration and total staff costs) were themselves missing in some instances. We used the "mice" package in R (van Buuren & Groothuis-Oudshoorn 2011) to carry out the imputations.

Multiple imputation assumes that data are Missing At Random (i.e. that the true values of missing observations only depend on data that were observed). We were unable to submit this assumption to a strict test. However, the observed correlations between the variables "number of sessions", "total duration", "total staff costs" and "cost weights" were all very high (varying between 0.89 for the correlation between number of sessions and cost weights, and 0.95 between number of sessions and total staff costs). These high correlations suggest that, knowing the number of sessions (as we do for all POCs in this sample), we can predict the total relative treatment costs with high accuracy.

We imputed 200 data sets, and performed the substantive analysis (i.e. the prediction of treatment cost by cluster membership) on each of these 200 data sets. We then combined estimates and adjusted standard errors according to Rubin's rules (Rubin 1987).

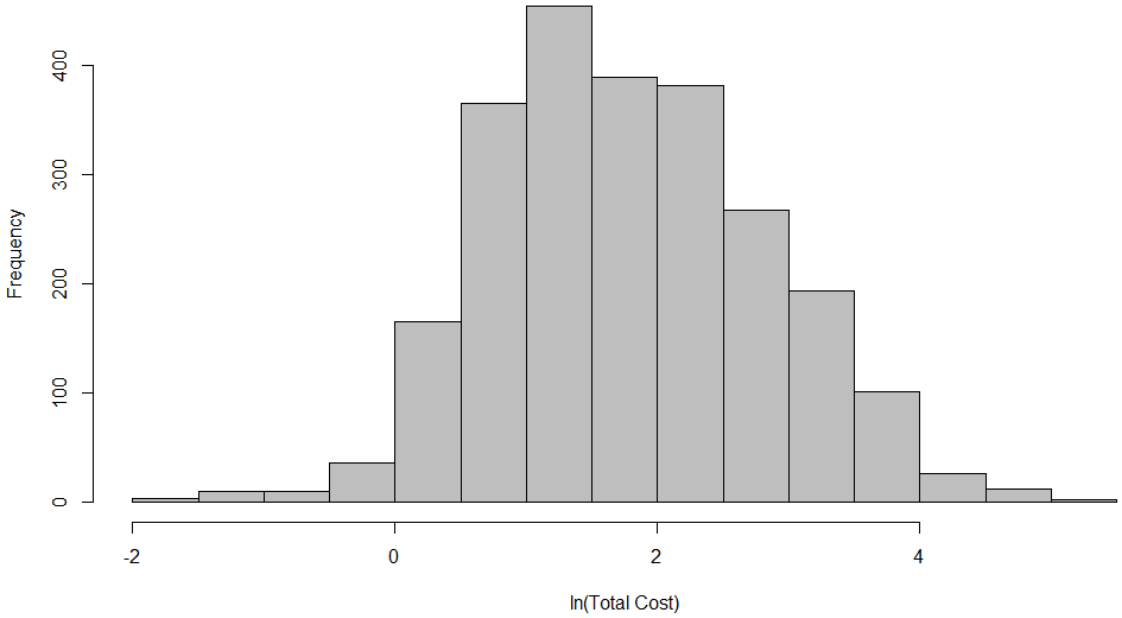
7.2 Substantive analysis

The substantive model of interest was specified so as to ascertain whether substantive conclusions about the relationships between clusters and resource use, as well as between complexity, contextual or EET issues and resource use, would change if we operationalized resource use as the total treatment cost for a POC, rather than simply as the number of sessions.

The available treatment costs (n = 2414) had a mean of 10.55 (SD: 14.4), with a minimum of 0.14, a maximum of 90.91, and a median of 5.61.[5] Their distribution was highly positively skewed (skew statistic: 4.25). Before analysis, we log-transformed the total weights, using the natural logarithm. The resulting distribution is shown in Figure E17. The log-transformation resulted in a distribution that approximated the normal quite well (skew = 0.23, kurtosis = -.0.06). We felt therefore justified in estimating a linear regression, predicting the log-transformed weights.

[5] Recall that these numbers represent relative treatment costs. We make no attempt to arrive at a precise estimate for a given child coming to CAMHS, but only aim to distinguish between children according to the (estimated) relative costs of their treatment.

Figure E17 Histogram of log-transformed total costs (n=2414)



Note: 2159 POCs with missing total costs are not represented in this graph.

7.3 Results

The multiple imputation appeared to have worked well. The maximum fraction of missing information (fmi) in the substantive model of interest was fmi = 0.00067 (so that fewer than 200 imputations would probably have been sufficient). The distribution of imputed relative costs was very similar in shape to the distribution of observed relative costs. With respect to the substantive model of interest, Table E12 and Figure E18 display the coefficients from the linear model predicting the logarithm of relative costs.

The findings provide no reason to change the conclusions we drew from the analysis predicting the number of appointments. The order and relative magnitude of the effects of the clusters is almost the same, regardless whether we are predicting the number of appointments or treatment cost. Children in the Eating Disorder, Self Harm, or Co-occuring Conduct and Emotional Problem clusters are predicted to have the highest resource use compared to other clusters. In general, clusters within the “Getting More Help” supercluster tend to have higher predicted resource use than those in most of the clusters within the “Getting Help” supercluster.

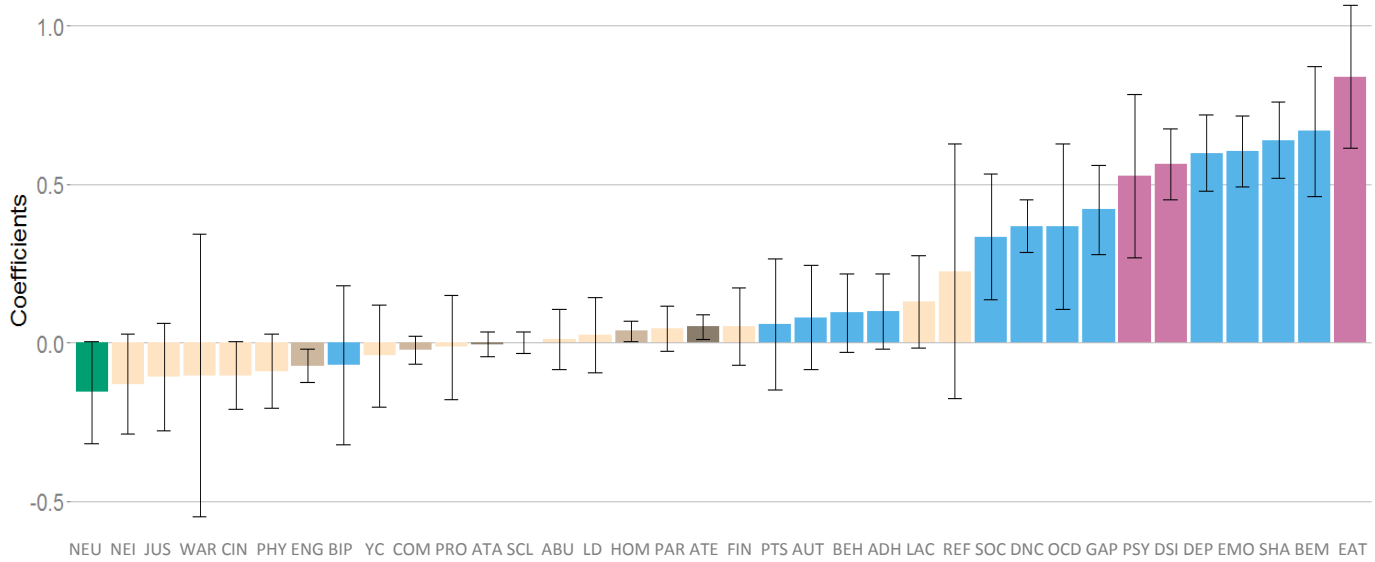
Like the main analysis, the sensitivity analysis provides little evidence for strong effects of any complexity factors, contextual problems, or EET issues. Point estimates of the effects sizes of these variables are very much smaller than those of the clusters with the highest predicted resource use. Moreover, all but three of the 95 % confidence intervals of the effects relating to complexity, contextual or EET issues include the value zero, suggesting that most of the small observed effects may be due to random variation, rather than constituting evidence for a real effect of the variables tested on resource use.

Table E12: Combined estimates from 200 imputed data sets. Linear regression predicting “total relative costs”

	Coefficient	Standard Error	95 % confidence interval	
Intercept	1.525	0.134	1.262	1.788
Clusters, Ref: ADV				
ADH	0.099	0.060	-0.019	0.217
AUT	0.080	0.084	-0.084	0.244
BIP	-0.070	0.129	-0.322	0.182
BEH	0.096	0.063	-0.028	0.219
DEP	0.599	0.062	0.478	0.721
GAP	0.420	0.072	0.279	0.561
OCD	0.368	0.134	0.106	0.630
PTS	0.060	0.106	-0.147	0.267
SHA	0.640	0.061	0.519	0.760
SOC	0.335	0.101	0.137	0.533
BEM	0.667	0.104	0.463	0.871
EMO	0.605	0.058	0.492	0.718
DNC	0.368	0.043	0.285	0.451
EAT	0.840	0.115	0.615	1.065
PSY	0.527	0.131	0.270	0.783
DSI	0.565	0.057	0.453	0.677
NEU and other factors				
NEU	-0.155	0.082	-0.317	0.006
LAC	0.130	0.074	-0.016	0.275
YC	-0.040	0.082	-0.200	0.120
LD	0.025	0.060	-0.092	0.143
PHY	-0.089	0.060	-0.207	0.029
NEI	-0.130	0.080	-0.288	0.027
PRO	-0.014	0.084	-0.179	0.152
CIN	-0.103	0.054	-0.209	0.004
REF	0.226	0.205	-0.175	0.628
WAR	-0.103	0.227	-0.549	0.343
ABU	0.012	0.048	-0.082	0.107
PAR	0.046	0.036	-0.025	0.116
JUS	-0.107	0.086	-0.276	0.061
FIN	0.053	0.062	-0.069	0.175
HOM	0.038	0.017	0.005	0.071
SCL	0.002	0.017	-0.033	0.036
COM	-0.022	0.022	-0.066	0.021
ENG	-0.072	0.027	-0.125	-0.019
ATE	0.051	0.020	0.013	0.089
ATA	-0.005	0.020	-0.044	0.034

Notes: See Table E9.b for key to abbreviations used.

Figure E18: Graphical representation of combined estimates from 200 imputed data sets. Linear regression predicting “total relative costs”



Notes: See Table E9.b for key to abbreviations used.

7.2 Closure Reason and Complexity Factors

Exploration of Closure Reasons

Amy Macdougall

As part of the data collection process, information on case closure reason was recorded for those cases which closed within the 22 month time frame. In this section we look at the quality of information for insight into whether the number of sessions has been underestimated for certain groups of patients. We primarily consider patients who stopped attending their appointments or finished their period of contact against the advice of the clinician.

Description of the Data

The main sample is comprised of 4573 cases which were either closed (87%) or assumed closed after six months of inactivity (13%). Of the closed cases, a quarter were missing information on the reason for case closure. The remaining 3009 cases with reasons for closure provide the data for this section.

Table E13 shows the possible codes for case closure reason[6] and the distribution of cases across these categories.

Table E13: Percentage of patients with each case closure reason.

Case closure reason	Percentage of patients (*n*)
1. Discharged on professional advice	82.18% (*2491)*
2. Discharged against professional advice	3.46% (*105)*
3. Patient non-attendance	9.40% (*285)*
4. Transferred to other health care provider - medium secure unit	0.36% (*11)*
5. Transferred to other health care provider - high secure unit	0%
6. Transferred to other health care provider - not medium/high secure	1.58% (*48)*
7. Transferred to adult mental health service	1.29% (*39)*
8. Patient moved out of area	1.62% (*49)*
9. Patient died	0.10% (*3)*

Most cases were recorded as being closed on professional advice (82%). A further 13% were closed either against professional advice or due to patient non-attendance; the remaining categories contain very small numbers of patients.

These categories fail to discriminate between the majority of the closed cases in our sample and we are left with fairly limited information on case closure.

We will focus on the second (discharged against advice) and third (patient non-attendance) categories, which explicitly indicate that treatment may have finished prematurely. The remaining categories with very small numbers of patients (4-9) are combined into one, labelled “transfer/move”.

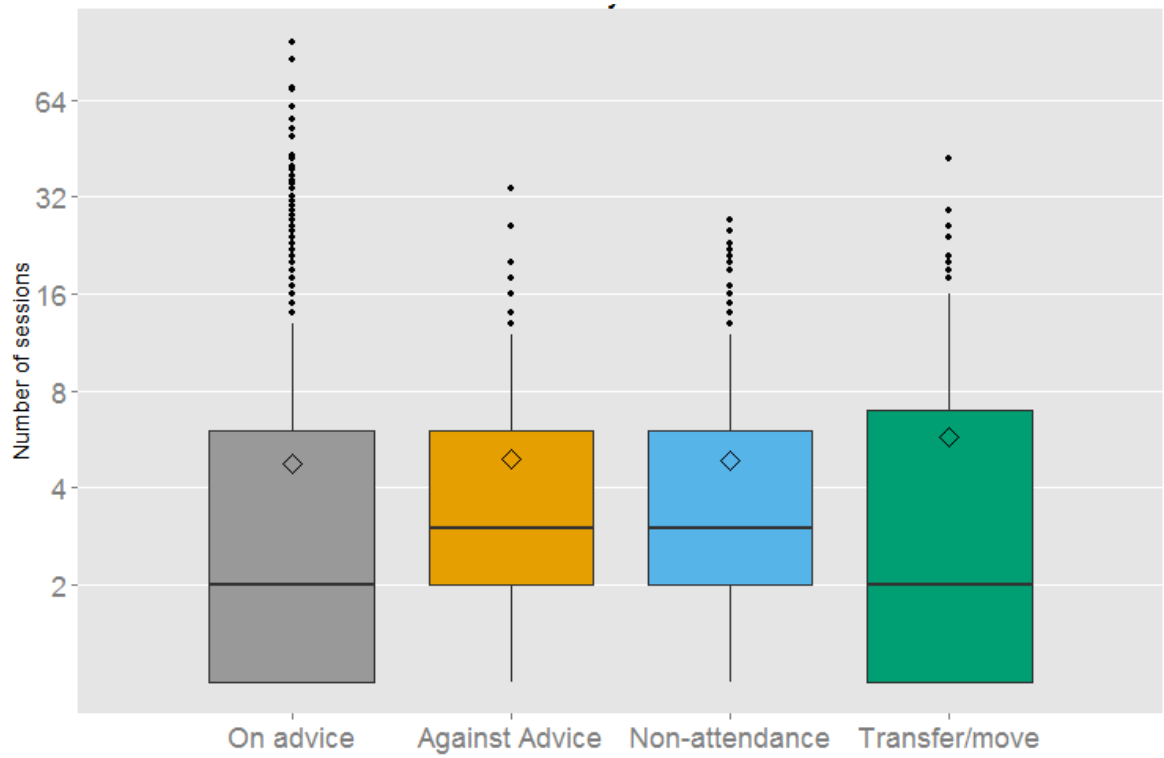
[6] These are National Codes for discharge from mental health service reason.

Table E14 gives the updated proportions for each combined category. Figure E.19 displays boxplots of number of sessions for each category, with the mean plotted as a diamond. Inspection of these reveals no clear differences in the distribution of number of sessions between subgroups: the means and medians are very similar, if not identical. Therefore, there is no suggestion of any particular group having lower resource use and biasing the sample.

Table E14: Percentage of patients

Closure Type	Percentage of patients
1. Discharged on professional advice	82.22%
2. Discharged against professional advice	3.49%
3. Patient non-attendance	9.37%
4. Transfer/move	3.24%

Figure E19: Number of Appointments by Case Closure Reason



Notes: Boxplots of number of sessions on the binary log scale. Means are plotted as diamonds.

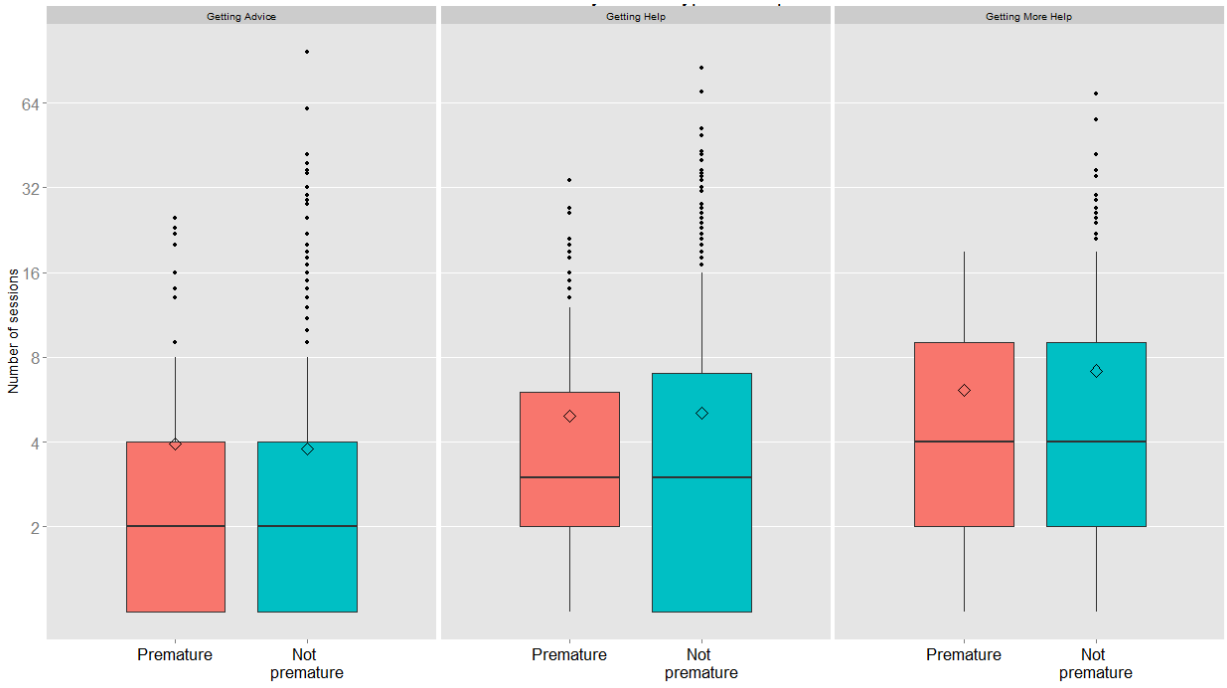
Given the small numbers discussed above it was not possible to break patients down by cluster. To increase sample sizes, patients in the groups discharged **against professional advice** and **patient non-attendance** are amalgamated into one group, labelled 'Premature'. Remaining patients who left on advice, or were transferred or moved are in one group labelled 'Not premature'.

The boxplots in Figure E20 are of number of sessions, broken down by:

I. super cluster
II. closure type:
- 'Premature' (those who left treatment against advice, or stopped attending)
- 'Not premature' (all other patients)

Again, the distributions look very similar across closure types. Medians are identical and means differ by very little within super clusters. There is no indication that resource use is different for those in the 'Premature' group, whose case closure reasons suggested that they may have finished treatment prematurely.

Figure E20: Number of Appointments by Closure Type and Super Grouping



Notes: Number of sessions is plotted on the binary log scale.

Summary

Services provided some information on the reason for case closures within the data collection period. It was noted that these categories were fairly limited in that they did not discriminate between most patients.

Resource use (as measured by number of sessions attended within each period of contact) was inspected for patients by broken down by their closure type. There was no evidence that resource use was lower for certain patients; namely those that had stopped attending or whose cases had been closed against professional advice. The same was true when we distinguished between patients in different superclusters. From these results, there is little reason to suspect that closure reason played the role of a confounding variable in our main analysis.

It was not possible to investigate the interaction between complexity factors and closure reason due to small numbers. We recommend a study designed specifically for this purpose or simply a larger sample size.

References (add references from Amy's & Andy's sections)

Carpenter, J. R., & Kenward, M. G. (2013). *Multiple Imputation and its Application*. Chichester: Wiley.

Gower, J.C. (1971), *A General Coefficient of Similarity and Some of its Properties.* Biometrics, 27(4), pp 857-871.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651–674.

Kaufman and Rousseeuw (2009), *Finding Groups in Data: an Introduction to Cluster Analysis.* Wiley

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. New York: Springer.

Kvalseth, T. (2014). Cautionary Note About R2, *39*(4), 279–285.

Rousseeuew (1986), *Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis.* Journal of Computational and Applied Mathematics, 20, pp. 53-65

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice : Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3), 1–67.

Wolpert, M., Harris, R., Jones, M., Hodges, S., Fuggle, P., James, R., Wiener, A., Mckenna, C.,Law, D., Fonagy, P. (2014) THRIVE. The AFC-Tavistock Model for CAMHS. London: CAMHS Press. Available: http://www.tavistockandportman.nhs.uk/sites/default/files/files/Thrive%20model%20for%20CAMHS.pdf .