Health Sciences

UNIVERSITY OF Southampton

Learning-based text mining in analysis of free-text responses to the (2013) Wales Cancer Patient Experience Survey (WCPES) –

Findings and future applications

09/10/2014

Dr. Mike Bracher

Dr. Richard Wagland

Prof. Dame Jessica Corner

Contents

Key points:

- Learning-based text mining has the potential to save time and resources in analysing free-text data from patients.

- The possibility of using this approach, and the quality of the results that it produces, are dependent upon the size and quality of the training data sets available for sorting the free-text material.

- Care must be taken when verifying the data sorted by text mining, ensuring full coverage of the free-text material and where necessary undertaking manual coding of the remainder of unsorted data. Researchers must also remain alert to the potential presence of novel data that does to map to the existing taxonomy of thematic categories used to classify responses.

- In future, a rules-based approach to text mining may be preferable to a learning-based approach. The former allows for direct control over data sorting through manual construction of rules, and offers the potential for integrating expert knowledge into the sorting of data (this is not practical with the latter).

- Exploration of rules-based text mining in analysis of free-text comments from patients is currently under way at the University of Southampton, Faculty of Health Sciences, in partnership with Nominet UK[1].

1 – Background and aims

Introduction

Researchers from the University of Southampton Faculty of Health Sciences were commissioned by Macmillan Cancer Support to analyse results from the free text portion of the 2013 cancer patient experience survey for Wales (WCPES). This analysis was carried out through a thematic content approach, in which the data were organised into themes by the research team (Bracher et al., 2014). As part of this work, it was agreed that the resulting coded data would be used as a test set in order to explore the potential of using text mining (TM) techniques using machine learning algorithms in future analysis of free text data. The purpose of this report is to explore the potential for using machine learning algorithms in processing patients comments, to evaluate their effectiveness vs. the 'gold standard' of manual coding, and to discuss the implications of these findings for future analysis of free-text data from patients.

[1] The authors acknowledge the contribution of David Simpson (Senior Researcher, Nominet UK), who gave feedback on an earlier version of this report.

What is text mining?

TM refers to the process of deriving high quality information (i.e. some particular aspect of the data that is of interest) from a given set of textual data, typically through identification of patterns and trends using statistical pattern learning (Hearst, 2003). In our application, the process is facilitated by supervised machine learning algorithms. The term 'supervised' here refers to the process of presenting an algorithm with a set of data that has already been coded (or 'labelled') as belonging to different categories (in this case, different aspects of cancer patient experience), so that it can learn to build a model of the patterns and associations within the data such that it will be able to classify future data in the same way (Collingwood and Wilkerson, 2012). In this application, our data are coded according to themes assigned to each comment left by WCPES respondents, and the algorithms are then given test sets of this coded data from which to derive patterns (e.g. a portion of the comments coded as relating to 'Nursing'). Each algorithm is then given a further test set on which to test their accuracy. This is then judged on a weighted average of their ability to correctly identify comments as belonging to a theme (referred to as *precision*) and how many comments corresponding to a particular theme they are able to identify from the total population of comments given (referred to as *recall*). Comments that have been coded as belonging to a particular theme are referred to as *positive* results, while comments not coded to the theme of interest are referred to as *negative* results. In testing the algorithms, we are looking to explore how well they are able to identify positive results and differentiate them from negative results.

Different algorithms build models and solve problems using different approaches, and therefore may differ in accuracy. This can also vary between themes, as the algorithm will need to build a different model for each theme it is given. For this reason, a combination of algorithms can sometimes be used (referred to as an *ensemble* approach) (Jurka et al., 2013). The accuracy of a particular ensemble is a function of the *coverage* (i.e. what percentage of the positive results in a given data set are agreed by a given set of algorithms) and *precision* (i.e. the number of true positive identified minus the number of false positives identified).

What is the potential for applying text mining techniques to analysis of free-text comments from patients?

Qualitative analysis of free-text comments has been used in analysis of free-text data from previous patient experience surveys (PES) and patient reported outcome measures (PROMs) instruments. This process involves a manual approach to classifying the data for analysis. This typically involves identification of *semantic* and *latent* themes (Boyatzis, 1998). The former refer to themes that are identified by their semantic content (e.g. terms corresponding to areas of treatment or care, such as 'nurse/nurses/nursing'), while the latter refer to meaningful connections between disparate themes that may not correspond directly to their semantic content (e.g. material referring to themes such as 'more information during treatment', 'better communication from staff in pre-operative stage' and 'lack of contact post-treatment' may be identified with a wider latent theme relating to 'importance of being prepared'). The latter type of theme often cuts across different aspects of patient journeys and experiences, and can be seen as meaningful or 'narrative' themes that arise from analysis of semantic themes. The process of identifying these themes involves a team of trained researchers, and agreement on the themes between researchers can be quantified using approaches such as the Cohen's kappa statistic (Carletta, 1996).

The algorithms classify data by building models based on observable aspects of the information with which they are presented (i.e. words, word classes, associations between words or phrases/constructions). Therefore, their use relates to the organisation of data into semantic themes, rather than 'reading for meaning'. Their use can therefore be described as a form of data organisation, rather than true analysis which requires the judgement of trained human researchers. Qualitative analysis by teams of trained researchers involving manual-only sorting and coding of the data can be seen as the 'gold standard' against which other ways of organising and classifying data can be judged. The potential advantage of the proposed use of a learning-based text mining approach is to cut down on the amount of manual coding necessary, and in so doing make the process faster and more resource-efficient. This presents two potential benefits:

- *Cutting down on the time needed to produce analyses of free-text data, with the potential for the process to become more responsive.*

- *Reducing costs and resource use associated with this type of work.*

Both of these potential advantages could also be seen to enable analysis of free-text comments in situations where time and/or resource constraints may make this impractical using a 'gold-standard' approach. However, there are also some potential limitations to this approach in terms of both the effectiveness of algorithms and how their use may affect the quality of analysis that results.

2 – Methods

Free-text data from the WCPES were coded manually by researchers, and the data collated into a spreadsheet in which each row represented the free-text response of a single patient (in the second row of each column). Subsequent columns were used to assign comments to corresponding themes. This provided the 'gold standard' data set against which the algorithms would be tested, our objective being to see how well they could replicate this coding. Training and testing of the algorithms was conducted using the *RTextTools* package for R Statistical Computing software (Jurka et al., 2012). This package contains nine machine learning algorithms:

- *Support Vector Machines (SVM)*

- *Self-adaption Link-quality Detection Algorithm (SLDA)*

- *Boosting*

- *Bagging*

- *Random Forests (RF)*

- *Generalised Linear Models Network (GLMNET)*

- *Decision Trees (TREE)*

- *Neural Networks (NNET)*

- *Maximum Entropy (MAXENT)*

Based upon initial sorting with a smaller test set of coded data, the four most successful algorithms were chosen for use in this exploratory investigation:

- SVM
- SLDA
- RF
- TREE

These four algorithms were trained and tested using the coded WCPES data set (both separately and as ensemble).

Full testing of the algorithms involved the construction of test sets of data taken from the WCPES data set. The taxonomy into which the WCPES data set were sorted comprised 258 categories across five levels of specificity[2], and in this phase only the most general categories (n=29) with numbers of positive results >=50 were used (see table 1). The ability of algorithms to make successful predictions is a function of the size of the training set from which they are able to derive their rules and models, amongst many other things including the quality of the data and the approach used by the algorithm itself. While the numbers of respondents in the WCPES data set represent relatively large numbers compared with those involved in most forms of qualitative research, they are very small in relation to the number of data points found in most text mining applications. The limit of 50 is an arbitrary limit, but one that was set below that which could be expected to be a cut off for effectiveness in most applications of learning-based text-mining.

Test sets for each category were constructed using equally weighted numbers of positive and negative results (the latter were selected at random from the pool of available results)[3]. The test sets were then randomised again to ensure a random distribution of negative and positive results across the set. The next step was to define the training and testing portions of the data sets, where the testing set was defined as being from the 1st to 90th percentile, and the testing set from the 91st to 100th percentile in each set. This is a standard approach in supervised machine learning, allowing for the maximum possible range of data for training while retaining a suitable pool for testing (Collingwood and Wilkerson, 2012). The results of this training and testing procedure produced recall, precision and f-scores for each algorithm, as well as ensemble agreement data for each category. In addition, for each category the algorithms were subjected (individually) to k-fold[4] cross-validation, where the algorithm performs a given number of tests (in our case, 10). In each step of

[2] E.g. Nursing / Nursing Positive / Nursing communication positive / Nursing Communication Information Positive. The full taxonomy with associated numbers of positive results is given in the Appendix to this report.

[3] In most cases, the number of positive results was greater than the number of negatives, and so it was possible to match negatives at random to the positive results. However, for the two largest categories (see table 1) where positives were greater than negatives, the number of positive results was determined as being equal to the maximum number of available negatives. In these cases the positive results were also selected at random.

[4] k = 10.

this process 90% of the data are used for training and 10% for testing, meaning that across all of the steps each comment will be trained on and tested. This process ensures that the algorithm is capable of processing the entire set (rather than just the user-defined 10%), and provides accuracy scores that can give indications of consistency across the data.

Category	***True positive results (n)***	***Training set (n)***
Positive comments	3818	1708
Negative comments	2313	4626
Nursing	1074	2147
Communication between patients and healthcare staff	1013	2026
Waiting for appointments	670	1340
Surgery	541	1081
Hospital Doctors NOS[5]	476	952
Diagnostic and investigative processes and procedures	475	950
Consultant and Specialist Doctors NOS	466	932
General Practitioners	401	626
Chemotherapy	306	612
Follow up and aftercare in the post-treatment phase of the cancer journey.	290	552
Radiotherapy	251	502
Hospital environments	240	480
Communication between healthcare staff and/or institutions	238	476
Waiting times on the day of appointments	188	377
Travel relating to cancer treatment	161	322
Hospital food and catering	153	306
Emotional, social and psychological support	136	272
Staffing levels NOS	130	260
Oncology	117	234
Pain Management	82	164
Treatment and care at night, weekend and in the evening	69	138

Table 1 - Number of true positive results and size of training sets for tested categories.

[5] Not otherwise specified (in this case, not identified with any other area of medical specialty).

3 – Findings

3.1 - F-score results

The accuracy of algorithms in terms of their ability to correctly identify patient comments as belonging to a particular category is measured in terms of the f-scores. These scores are a function of separate scores for *precision* and *recall*. Precision (P) is calculated using by dividing the number of true positives (TP) identified by an algorithm, by the number of true and false (FP) positives (P = TP / (TP + FP)), while recall (R) is calculated by dividing the number of true positives by the number of true positives plus false negatives (FN) (R = TP / (TP + FN)) (Fawcett, 2006). The harmonic mean of these two scores gives us the f-score. Put simply, *precision* scores tell us what percentage of positive results identified by an algorithm are true positives. *Recall* indicates the percentage of true positives that have been identified by a given algorithm. F-scores thus represent a weighted average of these results (full results for the categories tested can be viewed in the appendix). The highest f-score for any algorithm in any category was 1 (i.e. 100% for the TREE algorithm in the '3.8.Oncology' category), while the lowest was 0.43 (i.e. 43% for the SLDA algorithm in both the '3.16.Emotional.Social.MH.support' and '2.5.Out.of.hours.Weekend.NOS' categories). The mean f-score for all algorithms in all categories was 0.79 (with a standard deviation of 0.086).

3.2 – Ensemble results

The accuracy of the process may be improved using an ensemble approach, where two or more algorithms 'agree' on a particular label. Typically, as more algorithms are added to the ensemble, we would expect to see an increase in precision and a decrease in coverage, and this was observed (Jurka et al., 2013). The full results in the appendix compare the best performing single algorithm in each category with the ensemble giving the highest recall score where coverage was equal to or greater than the recall score for the single algorithm. Coverage is similar to recall, except that in this case the criteria are the number of true positives that are agreed upon by the specified number of algorithms (as opposed to simply being identified by at least one in the ensemble) (Jurka et al., 2013). Across all categories, mean coverage score for the best performing ensembles was on average 0.10 higher than for the best performing single algorithm. Recall scores for the best performing ensemble were on average 0.03 lower than the best performing single algorithm.

3.3 – K-fold results

All algorithms were subjected to k-fold cross validation across all categories to ensure that they were able to process the entirety of the data set, and this process also generated accuracy scores. Accuracy (ACC) is calculated by dividing the number of true positives plus true negatives by the number of identified positives plus identified negatives (i.e. those identified by the algorithm – [ACC

= (TP + TN / (P + N)]) (Fawcett, 2006). The mean accuracy score for all k-fold procedures for all algorithms in all categories was 0.822 (with a standard deviation of 0.058).

3.4 – Trends in the results

The findings presented in the appendix, as well as the average scores indicated above give a broad picture of the performance of the individual algorithms and ensembles. However, for the purposes of evaluating the potential applications of learning-based text mining to future work in processing patient feedback from free-text, several other results must be considered. The quality of the results are a function of the quality of the data provided, the approach taken by the different algorithms, and the size (n) of the training sets that the algorithms have for developing their approach to classifying comments.

Data quality

How well an individual algorithm or ensemble of algorithms performs will be determined in part by the quality of the data (i.e. how 'difficult' or 'easy' it is for a given algorithm to generate rules from data sets). For example, comments belonging to a category which has a clear marker (e.g. a word or partial word, such as 'Nurse/Nurses/Nursing' etc.) may be easier to identify than those belonging to a category which has a broad set of terms, or is expressed in fuzzy terminology, or involves implicit meaning (e.g. language relating to emotional, social and/or psychological issues). This is reflected as a broad trend in the distribution of algorithm-average f-scores for categories with the highest and lowest values. Table 2 presents the four categories with highest overall f-scores and those four with the lowest values. Those categories where f-scores were highest overall tended to be those for which clear and consistent markers exist in the comments, while the categories with comparatively poorer scores tended to be those with broader or 'fuzzier' markers or terminology (this trend is observable across the data presented in the appendix).

Highest average f-score categories	n	Score	Lowest average f-score categories	n	Score
X3.18.Nursing	2147	0.915	X2.1.Com.Inter.Intra.Agency	476	0.71375
X3.6.Hospital.Doctors.NOS	952	0.915	X3.3.Consultants.SpecialistsNOS	932	0.68875
X3.11.Radiotherapy	502	0.91375	X3.16.Emotional.Social.MH.support	272	0.65875
X3.8.Oncology	234	0.91125	X2.5.Out.of.hours.Weekend.NOS	138	0.65875

Table 2 - Highest and lowest algorithm-average f-score categories.

Approach of the algorithms

It is also important to take into account variations in individual algorithm scores within the different categories, and what this indicates about variations in their suitability for particular kinds of free-text data. A detailed examination of each category in relation to the theoretical approach of each

algorithm is beyond the scope of this paper; however, for illustrative purposes we can take an example such as that given in Table 3.

Category	Test n	TREE F-score
X3.8.Oncology	234	1
X3.18.Nursing	2147	0.94
X3.11.Radiotherapy	502	0.935
X3.9.Pain.Management	164	0.935
X3.6.Hospital.Doctors.NOS	952	0.925
X3.2.Chemotherapy	612	0.915
X3.19.Surgery	1081	0.915
X2.4.Wait.On.Day	377	0.88
X3.3.Consultants.SpecialistsNOS	932	0.865
X3.7.Investigations	950	0.84
X4.4.Travel	322	0.81
X4.2.Food.Catering	306	0.805
X3.16.Emotional.Social.MH.support	272	0.785
X2.5.Out.of.hours.Weekend.NOS	138	0.785
X1.1.Positive.Clean	1708	0.775
X2.2.Com.Pat.Prov	2026	0.765
X2.1.Com.Inter.Intra.Agency	476	0.765
X3.15.After.care	398	0.74
X2.3.Wait.App	1340	0.735
X1.2.Improve.Clean	4626	0.715
X4.1.Environment	480	0.69
X3.4.GP	626	0.68
X2.6.Staff.Levels.NOS	260	0.61

Table 3 - TREE f-scores for all categories

The table above presents the f-scores for the TREE algorithm for all categories tested, ordered from highest to lowest score. In this table, categories towards the lower end of the table tend to be those with broader or 'fuzzier' sets of markers than those at the top. Furthermore, the 'test n' value does not follow the distribution of the f-scores. Both of these indicate that for TREE, the quality of the data was the more significant of the two factors in terms of their effect on algorithm performance, and we may wonder how might this relate to the approach used by the algorithm? TREE (decisions trees) is a type of algorithm that solves problems by creating multi-level branched decision maps (or 'trees'), creating a series of binary classifications that the algorithm will use to sort and code the data. This type of approach appears in theoretical terms to be well suited to classifying comments where there are clear and/or narrow sets of markers that identify comments with a particular category (e.g. in the category 'Oncology', where every comment identified with this category will contain at least the partial word 'Oncolog…', for which this algorithm was 100% accurate). It appears less well suited to making associations between sets with broader or fuzzier sets of markers, and this is borne out by

the distribution of scores. What this example indicates is that appreciation of the general approach taken by different algorithms is necessary in order to make informed judgements about the type of information sorting to which they may best be suited (the implications of this for future work in analysis of patient free-text comments is discussed in the next section).

4 – Discussion

4.1 - What does this tell us about the potential use of learning-based text mining in analysis of free-text comments from patients?

The intended outcome of using learning-based text mining in analysis of free-text data from patients is that the process cuts down on the amount of manual sorting required, making coding and analysis of this data quicker and more resource efficient. The gold-standard for this type of work involves a team of trained researchers, who sort and code the entirety of the data set manually, and perform appropriate checks on agreement between researchers on how codes are applied to the comments (for more details on this type of approach, see Bracher et al., 2014). This is the quality standard against which any advantages in terms of time and resource efficiency from using learning-based text mining are assessed.

Speed and accuracy of coding

While the exact amount of time taken on manual sorting and cleaning of the data was not measured formally, the potential for sorting the data into a general framework prior to initial coding has obvious advantages. Firstly, it means that 'cleaning' the data becomes a more focused process, requiring verification of membership of one category (i.e. does this comment belong in the nursing category?), rather than reading and sorting each comment into multiple categories (as in Bracher et al., 2014). Secondly, in the stage two of coding, when more detailed codes are applied within the most general categories (e.g. 'Nursing Positive' / 'Nursing Negative' / Nursing Communication with Patients'), the researcher has only to work with a small taxonomy of categories, with associated benefits for speed and accuracy (i.e. that they are likely to miss fewer terms if they are working on a smaller set of codes at any one time).

The manual coding of the WCPES data, at the level of detail present in the full taxonomy, was extremely labour-intensive. Further, while this method can be seen as the gold-standard, and was essential for developing from the bottom-up a taxonomy of terms for sorting patient experience data, the existence of this taxonomy presents new opportunities for future work. The fact that we now have a system of categories for coding that is derived from a national survey population of free-text respondents who are cancer patients (one that appears broadly representative of respondents to the full survey) means that this may be applied in future work of this type (Bracher et al., 2014).

Accuracy is lower for the algorithms in almost all cases when measured against the number of true positives (i.e. those coded by human coders), and it is impossible for them to exceed this standard. However, given that the results from the algorithms would not be taken 'as is' but rather verified or 'cleaned' by researchers in future applications, it is likely most errors relating to false positives or

false negatives would be detected, for example, a comment that does not belong in 'Nursing' could excluded, and if necessary recoded into the appropriate category. Should the algorithms leave a remainder of comments unprocessed (i.e. they are not labelled to a specific category), these can be coded manually by the researcher. This dual approach would reduce greatly the errors present in algorithm-only sorting.

Does learning-based text mining using an existing taxonomy involve risk of loss of novel data in future applications?

One of the major strengths of using free-text data is that it is largely unstructured (save for the questions that prompt responses, e.g. 'what are positive/negative about your cancer care?'). It is this freedom that gives us the opportunity to observe patient concerns that may not be covered by closed questions, and allows patients to provide additional detail that may help contextualise their responses to quantitative measures. There is a question, therefore, as to whether a taxonomy developed from one data set risks obscuring useful original data from future sets (i.e. new findings that do not map to the existing taxonomy).

This is a serious concern in using text processing systems of any kind, instead of coding the data in an entirely bottom-up fashion. However, steps can be taken that, if applied in a consistent and rigorous way, will likely minimise the risk of losing original data in future surveys.

1) ***All comments would still be read by researchers*** – learning-based sorting only takes place at the highest level of taxonomy, and these will be cleaned prior to more detailed coding. This means that researchers will have the opportunity to see all comments and thus to code comments that are novel and/or do not map directly to the existing taxonomy.

2) ***The taxonomy can be developed over time in response to novel findings –*** the taxonomy itself can be adjusted in response to future findings emerging from future work. In turn, the training data for algorithms can be expanded to include new categories, as well as new material for improving training in existing categories.

Are there any other limitations to the practical application of text mining to sorting of free-text comments from patients?

Necessary expertise

Learning-based text mining (in our case, using the RTextTools package) requires specialist knowledge of both the systems for implementing them (in this case, the R Statistical Software) as well as theoretical knowledge of the approaches of different algorithms, in order to assess their individual suitability for given tasks and interpret the resulting data. While it is possible to apply and use this package with only limited knowledge of these areas, it is recommended strongly that a specialist in text mining and machine learning is consulted at all stages of the application.

Learning-based vs. rule-based text mining.

What the findings of our application indicate, is that the algorithms were particularly successful when categories were defined by clear and narrow sets of markers, such as particular words or

partial words. Sorting of this type could also be achieved by other processes, such as a rules-based approach to text mining, in which formal rules could be written by researchers that perform the same function (similar to a more sophisticated form of web searching). The advantage over a learning-based approach is that the process becomes controllable and transparent (i.e. we can see the process by which comments are categorised and amend them by changing or adding new rules). This extra level of control also offers the possibility of leveraging expert knowledge (e.g. from consultants, nurses, hospital doctors and other healthcare staff) to inform the rules that are used to sort the information. This would be especially useful in areas with 'fuzzy' or broad terminology, and for identifying novel themes in the data. By comparison, a learning-based approach using algorithms can be thought of as a 'black box', i.e. while we may have theoretical knowledge of the kinds of approach an individual algorithm may take, we cannot inspect directly the specific models or solutions built for each category (nor can we change or amend them directly). Consultation with specialists in machine learning and text engineering indicates that a rules-based text mining system is preferable for this type of sorting. Applications of this approach to sorting of patient comments from free-text are currently under way at the University of Southampton in partnership with Nominet UK.

5 - Conclusion

Learning-based text mining has the potential to save time and resources in analysing free-text data from patients. The possibility of using this approach, and the quality of the results that it produces, are dependent upon the size and quality of the training data sets available for sorting the free-text material. The results of the 'gold-standard' manual approach to thematic analysis of free-text data from the WCPES have produced both a taxonomy and training data set that can facilitate analysis of free-text material from cancer patients in the future. Attention to the points raised in this report with respect to checking of algorithm results, as well as the need to ensure complete coverage of the data and remain alert to novel findings which may not map to the existing taxonomy, can help mitigate some of the potential limitations associated with this approach. In addition, it is likely that a rules-based approach to text mining can enhance this process by providing a more accurate system for identifying comments. This type of system would also be amenable to direct control by the researcher, and thus able to incorporate knowledge from expert informants in a manner not practical with a learning-based approach.

References

BOYATZIS, R. E. 1998. *Transforming qualitative information: Thematic analysis and code development,* Thousand Oaks, CA, US, Sage Publications, Inc.

BRACHER, M., WAGLAND, R. & CORNER, J. 2014. Exploration and analysis of free-text comments from the 2013 Wales Cancer Patient Experience Survey (WCPES). Southampton, UK: University of Southampton.

CARLETTA, J. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics,* 22.

COLLINGWOOD, L. & WILKERSON, J. 2012. Tradeoffs in Accuracy and Efficiency in Supervised Learning Methods. *Journal of Information Technology & Politics,* 9**,** 298-318.

FAWCETT, T. 2006. An introduction to ROC analysis. *Pattern Recognition Letters,* 27**,** 861-874.

HEARST, M. 2003. What Is Text Mining? Berkeley, CA, USA: UC Berkeley.

JURKA, T. P., COLLINGWOOD, L., BOYDSTUN, A. E., GROSSMAN, E. & VAN ATTEVELDT, W. 2012. *RTextTools: Automatic Text Classification via Supervised Learning. R package version 1.3.9.*

JURKA, T. P., COLLINGWOOD, L., BOYDSTUN, A. E., GROSSMAN, E. & VAN ATTEVELDT, W. 2013. RTextTools: A Supervised Learning Package for Text Classification. *The R Journal,* 5**,** 7.

Appendix A – Test results for learning-based text mining in sorting of free-text responses to the WCPES.

Category	Test n	SVM Precision	SVM Recall	SVM F-score
X3.8.Oncology	234	0.94	0.97	0.955
X3.18.Nursing	2147	0.92	0.915	0.915
X3.11.Radiotherapy	502	0.9	0.905	0.9
X3.9.Pain.Management	164	0.765	0.765	0.75
X3.6.Hospital.Doctors.NOS	952	0.93	0.935	0.93
X3.2.Chemotherapy	612	0.935	0.935	0.935
X3.19.Surgery	1081	0.91	0.905	0.905
X2.4.Wait.On.Day	377	0.875	0.89	0.88
X3.3.Consultants.SpecialistsNOS	932	0.76	0.64	0.675
X3.7.Investigations	950	0.855	0.86	0.855
X4.4.Travel	322	0.875	0.875	0.87
X4.2.Food.Catering	306	0.79	0.79	0.79
X3.16.Emotional.Social.MH.support	272	0.725	0.725	0.71
X2.5.Out.of.hours.Weekend.NOS	138	0.725	0.725	0.71
X1.1.Positive.Clean	1708	0.795	0.79	0.785
X2.2.Com.Pat.Prov	2026	0.73	0.73	0.725
X2.1.Com.Inter.Intra.Agency	476	0.73	0.73	0.725
X3.15.After.care	398	0.795	0.795	0.795
X2.3.Wait.App	1340	0.79	0.79	0.79
X1.2.Improve.Clean	4626	0.835	0.83	0.83
X4.1.Environment	480	0.85	0.865	0.85
X3.4.GP	626	0.825	0.825	0.825
X2.6.Staff.Levels.NOS	260	0.77	0.77	0.77
Mean scores		**0.827173913**	**0.824347826**	**0.820652174**
Mean SD		**0.073203112**	**0.085057525**	**0.082259351**

Table 4 - Precision, recall and f-scores for SVM algorithm.

Category	Test n	SLDA Precision	SLDA Recall	SLDA F-score
X3.8.Oncology	234	0.77	0.81	0.735
X3.18.Nursing	2147	0.88	0.88	0.875
X3.11.Radiotherapy	502	0.9	0.905	0.9
X3.9.Pain.Management	164	0.62	0.62	0.62
X3.6.Hospital.Doctors.NOS	952	0.885	0.885	0.88
X3.2.Chemotherapy	612	0.885	0.885	0.88
X3.19.Surgery	1081	0.83	0.825	0.825
X2.4.Wait.On.Day	377	0.785	0.795	0.79
X3.3.Consultants.SpecialistsNOS	932	0.715	0.8	0.745
X3.7.Investigations	950	0.815	0.81	0.8
X4.4.Travel	322	0.53	0.53	0.525
X4.2.Food.Catering	306	0.54	0.545	0.51
X3.16.Emotional.Social.MH.support	272	0.44	0.44	0.43
X2.5.Out.of.hours.Weekend.NOS	138	0.44	0.44	0.43
X1.1.Positive.Clean	1708	0.735	0.725	0.72
X2.2.Com.Pat.Prov	2026	0.6	0.605	0.6
X2.1.Com.Inter.Intra.Agency	476	0.6	0.605	0.6
X3.15.After.care	398	0.745	0.745	0.745
X2.3.Wait.App	1340	0.725	0.725	0.72
X1.2.Improve.Clean	4626	0.795	0.78	0.775
X4.1.Environment	480	0.815	0.805	0.805
X3.4.GP	626			
X2.6.Staff.Levels.NOS	260	0.77	0.77	0.77
Mean scores		**0.719090909**	**0.724090909**	**0.712727273**
Mean SD		**0.141830183**	**0.142617006**	**0.143450068**

Table 5 - Precision, recall and f-scores for SLDA algorithm.

Category	Test n	RF Precision	RF Recall	RF F-score
X3.8.Oncology	234	0.94	0.97	0.955
X3.18.Nursing	2147	0.93	0.93	0.93
X3.11.Radiotherapy	502	0.92	0.92	0.92
X3.9.Pain.Management	164	0.675	0.65	0.62
X3.6.Hospital.Doctors.NOS	952	0.925	0.925	0.925
X3.2.Chemotherapy	612	0.89	0.885	0.885
X3.19.Surgery	1081	0.9	0.9	0.895
X2.4.Wait.On.Day	377	0.935	0.935	0.935
X3.3.Consultants.SpecialistsNOS	932	0.445	0.5	0.47
X3.7.Investigations	950	0.87	0.875	0.87
X4.4.Travel	322	0.855	0.845	0.845
X4.2.Food.Catering	306	0.805	0.825	0.775
X3.16.Emotional.Social.MH.support	272	0.8	0.75	0.71
X2.5.Out.of.hours.Weekend.NOS	138	0.8	0.75	0.71
X1.1.Positive.Clean	1708	0.825	0.825	0.825
X2.2.Com.Pat.Prov	2026	0.77	0.77	0.765
X2.1.Com.Inter.Intra.Agency	476	0.77	0.77	0.765
X3.15.After.care	398	0.795	0.795	0.795
X2.3.Wait.App	1340	0.875	0.86	0.865
X1.2.Improve.Clean	4626			
X4.1.Environment	480	0.85	0.85	0.85
X3.4.GP	626			
X2.6.Staff.Levels.NOS	260	0.81	0.81	0.81
Mean scores		**0.827857143**	**0.825714286**	**0.815238095**
Mean SD		**0.111091982**	**0.10766018**	**0.117372017**

Table 6 - Precision, recall and f-scores for TREE algorithm.

Category	Test n	TREE Precision	TREE Recall	TREE F-score
X3.8.Oncology	234	1	1	1
X3.18.Nursing	2147	0.94	0.94	0.94
X3.11.Radiotherapy	502	0.935	0.94	0.935
X3.9.Pain.Management	164	0.95	0.93	0.935
X3.6.Hospital.Doctors.NOS	952	0.925	0.93	0.925
X3.2.Chemotherapy	612	0.92	0.92	0.915
X3.19.Surgery	1081	0.915	0.915	0.915
X2.4.Wait.On.Day	377	0.875	0.89	0.88
X3.3.Consultants.SpecialistsNOS	932	0.855	0.88	0.865
X3.7.Investigations	950	0.84	0.845	0.84
X4.4.Travel	322	0.83	0.82	0.81
X4.2.Food.Catering	306	0.825	0.85	0.805
X3.16.Emotional.Social.MH.support	272	0.785	0.79	0.785
X2.5.Out.of.hours.Weekend.NOS	138	0.785	0.79	0.785
X1.1.Positive.Clean	1708	0.775	0.775	0.775
X2.2.Com.Pat.Prov	2026	0.77	0.77	0.765
X2.1.Com.Inter.Intra.Agency	476	0.77	0.77	0.765
X3.15.After.care	398	0.77	0.745	0.74
X2.3.Wait.App	1340	0.74	0.74	0.735
X1.2.Improve.Clean	4626	0.715	0.715	0.715
X4.1.Environment	480	0.705	0.685	0.69
X3.4.GP	626	0.68	0.68	0.68
X2.6.Staff.Levels.NOS	260	0.655	0.63	0.61
Mean scores		**0.824347826**	**0.823913043**	**0.817826087**
Mean SD		**0.095528486**	**0.099976776**	**0.10098184**

Table 7 - - Precision, recall and f-scores for RF algorithm.

Category	Test n	Mean category precision (all algorithms)	Mean category recall (all algorithms)	Mean category f-scores (all algorithms)
X3.8.Oncology	234	0.9125	0.9375	0.91125
X3.18.Nursing	2147	0.9175	0.91625	0.915
X3.11.Radiotherapy	502	0.91375	0.9175	0.91375
X3.9.Pain.Management	164	0.7525	0.74125	0.73125
X3.6.Hospital.Doctors.NOS	952	0.91625	0.91875	0.915
X3.2.Chemotherapy	612	0.9075	0.90625	0.90375
X3.19.Surgery	1081	0.88875	0.88625	0.885
X2.4.Wait.On.Day	377	0.8675	0.8775	0.87125
X3.3.Consultants.SpecialistsNOS	932	0.69375	0.705	0.68875
X3.7.Investigations	950	0.845	0.8475	0.84125
X4.4.Travel	322	0.7725	0.7675	0.7625
X4.2.Food.Catering	306	0.74	0.7525	0.72
X3.16.Emotional.Social.MH.support	272	0.6875	0.67625	0.65875
X2.5.Out.of.hours.Weekend.NOS	138	0.6875	0.67625	0.65875
X1.1.Positive.Clean	1708	0.7825	0.77875	0.77625
X2.2.Com.Pat.Prov	2026	0.7175	0.71875	0.71375
X2.1.Com.Inter.Intra.Agency	476	0.7175	0.71875	0.71375
X3.15.After.care	398	0.77625	0.77	0.76875
X2.3.Wait.App	1340	0.7825	0.77875	0.7775
X1.2.Improve.Clean	4626	0.781666667	0.775	0.773333333
X4.1.Environment	480	0.805	0.80125	0.79875
X3.4.GP	626	0.7525	0.7525	0.7525
X2.6.Staff.Levels.NOS	260	0.75125	0.745	0.74
Mean scores		**0.79865942**	**0.798478261**	**0.790905797**
Mean SD		**0.080307761**	**0.083827946**	**0.086059449**

Table 8 - Mean category precision, recall and f-scores for all algorithms.

Category	Test set n	SVM mean accuracy	SVM lower fold accuracy	SVM upper fold accuracy	SVM SD
X1.2.Improve.Clean	4626	0.806729753	0.782881002	0.829321663	0.016241413
X2.1.Com.Inter.Intra.Agency	2147	0.772755801	0.6	0.863636364	0.076808116
X2.2.Com.Pat.Prov	2026	0.786423097	0.730569948	0.817204301	0.027485752
X1.1.Positive.Clean	1708	0.803959195	0.736842105	0.850299401	0.037273076
X2.3.Wait.App	1340	0.821420925	0.763358779	0.885350318	0.036773753
X2.4.Wait.On.Day	1081	0.877117233	0.829787234	0.925	0.030297368
X2.5.Out.of.hours.Weekend.NOS	952	0.817385947	0.727272727	0.909090909	0.057448196
X2.6.Staff.Levels.NOS	950	0.799560375	0.678571429	0.935483871	0.100551549
X3.2.Chemotherapy	932	0.960761617	0.9375	0.984615385	0.016046199
X3.3.Consultants.SpecialistsNOS	626	0.893335287	0.826086957	0.943925234	0.035359111
X3.4.GP	612	0.78709643	0.707692308	0.847457627	0.048256788
X3.7.Investigations	502	0.864653604	0.816326531	0.929292929	0.033879857
X3.8.Oncology	480	0.933656315	0.826086957	1	0.064300855
X3.9.Pain.Management	476	0.93245098	0.8125	1	0.074290559
X3.6.Hospital.Doctors.NOS	398	0.891187065	0.824324324	0.944954128	0.0328412
X3.11.Radiotherapy	377	0.90297181	0.847826087	0.936507937	0.032396407
X3.15.After.care	322	0.758325542	0.514285714	0.903225806	0.121725862
X3.16.Emotional.Social.MH.support	306	0.772570162	0.535714286	0.84	0.090674392
X3.18.Nursing	272	0.927075562	0.905555556	0.966183575	0.018114425
X3.19.Surgery	260	0.885678288	0.8125	0.92248062	0.033209224
X4.1.Environment	234	0.77149093	0.714285714	0.880952381	0.060472134
X4.2.Food.Catering	164	0.861082337	0.806451613	0.933333333	0.047416589
X4.4.Travel	138	0.871970906	0.740740741	0.958333333	0.068178376
Mean scores		**0.847811268**	**0.759876522**	**0.91333257**	**0.050436574**
All-category mean SD		*0.061809641*	*0.10446984*	*0.053585268*	*0.028018055*

Table 9 - k-fold cross validation data for SVM algorithm.

Category	Test set n	SLDA mean accuracy	SLDA lower fold accuracy	SLDA upper fold accuracy	SLDA SD
X1.2.Improve.Clean	4626	0.763234324	0.732334047	0.789583333	0.018101885
X2.1.Com.Inter.Intra.Agency	2147	0.767454874	0.681818182	0.829787234	0.054540187
X2.2.Com.Pat.Prov	2026	0.728901461	0.702970297	0.766839378	0.0212545
X1.1.Positive.Clean	1708	0.751751942	0.701086957	0.846153846	0.043504668
X2.3.Wait.App	1340	0.749490923	0.671755725	0.797202797	0.037579739
X2.4.Wait.On.Day	1081	0.749112251	0.52	0.931034483	0.749112251
X2.5.Out.of.hours.Weekend.NOS	952	0.658791728	0.307692308	0.833333333	0.184264916
X2.6.Staff.Levels.NOS	950	0.72199852	0.592592593	0.807692308	0.069668459
X3.2.Chemotherapy	932	0.920774732	0.866666667	0.965517241	0.032046871
X3.3.Consultants.SpecialistsNOS	626	0.880370071	0.846153846	0.91	0.022001407
X3.4.GP	612	0.888191168	0.869565217	0.921568627	0.016201694
X3.7.Investigations	502	0.790133304	0.72972973	0.862068966	0.045245323
X3.8.Oncology	480	0.692459595	0.347826087	0.894736842	0.200796991
X3.9.Pain.Management	476	0.610395328	0.428571429	0.8125	0.133603598
X3.6.Hospital.Doctors.NOS	398	0.861741818	0.833333333	0.885416667	0.019674371
X3.11.Radiotherapy	377	0.879601379	0.824561404	0.924528302	0.034973309
X3.15.After.care	322	0.72490078	0.595238095	0.863636364	0.072200849
X3.16.Emotional.Social.MH.support	306	0.683310761	0.52	0.852941176	0.116240247
X3.18.Nursing	272	0.894310006	0.862944162	0.933333333	0.023577029
X3.19.Surgery	260	0.798973293	0.735042735	0.843137255	0.034182229
X4.1.Environment	234	0.573084109	0.422222222	0.790697674	0.1251309
X4.2.Food.Catering	164	0.69366615	0.390243902	0.862068966	0.162770168
X4.4.Travel	138	0.704167677	0.473684211	0.866666667	0.161017026
Mean scores		*0.760296356*	*0.637218833*	*0.860454121*	*0.103377766*
All-category mean SD		*0.093278482*	*0.180229876*	*0.053704653*	*0.152756924*

Table 10 - k-fold cross validation data for SLDA algorithm.

Category	Test set n	RF mean accuracy	RF lower fold accuracy	RF upper fold accuracy
X1.2.Improve.Clean	4626	DNF	DNF	DNF
X2.1.Com.Inter.Intra.Agency	2147	0.791937548	0.68	0.931818182
X2.2.Com.Pat.Prov	2026	0.799901952	0.767772512	0.843575419
X1.1.Positive.Clean	1708	0.815210598	0.78974359	0.826815642
X2.3.Wait.App	1340	0.836266631	0.76744186	0.880597015
X2.4.Wait.On.Day	1081	0.886654397	0.815789474	0.96969697
X2.5.Out.of.hours.Weekend.NOS	952	0.801665016	0.545454545	1
X2.6.Staff.Levels.NOS	950	0.813825792	0.722222222	0.896551724
X3.2.Chemotherapy	932	0.936566593	0.894736842	0.962962963
X3.3.Consultants.SpecialistsNOS	626	0.882607799	0.845238095	0.929292929
X3.4.GP	612	0.895640892	0.847058824	0.923076923
X3.7.Investigations	502	0.869012076	0.811111111	0.921348315
X3.8.Oncology	480	0.862059011	0.727272727	1
X3.9.Pain.Management	476	0.905965285	0.769230769	1
X3.6.Hospital.Doctors.NOS	398	0.885094024	0.844444444	0.930232558
X3.11.Radiotherapy	377	0.909643316	0.86	0.959183673
X3.15.After.care	322	0.804132996	0.7	0.88
X3.16.Emotional.Social.MH.support	306	0.784823908	0.68	0.866666667
X3.18.Nursing	272	0.922539806	0.896551724	0.959798995
X3.19.Surgery	260	0.877588944	0.846153846	0.927272727
X4.1.Environment	234	0.827094903	0.773584906	0.875
X4.2.Food.Catering	164	0.844381223	0.76	0.90625
X4.4.Travel	138	0.903836935	0.866666667	0.96875
Mean scores		*0.857111348*	*0.78229428*	*0.925404123*
All-category mean SD		*0.046829652*	*0.084110061*	*0.04978471*

Table 11 - k-fold cross validation data for RF algorithm.

Category	Test set n	TREE mean accuracy	TREE lower fold accuracy	TREE upper fold accuracy	TREE SD
X1.2.Improve.Clean	4626	0.70144005	0.678646934	0.722222222	0.013753271
X2.1.Com.Inter.Intra.Agency	2147	0.758772599	0.62	0.872340426	0.073944954
X2.2.Com.Pat.Prov	2026	0.722524511	0.673267327	0.75	0.021183165
X1.1.Positive.Clean	1708	0.778050279	0.743902439	0.830409357	0.027668239
X2.3.Wait.App	1340	0.764604731	0.726027397	0.805084746	0.026171699
X2.4.Wait.On.Day	1081	0.826032802	0.75	0.926829268	0.061938785
X2.5.Out.of.hours.Weekend.NOS	952	0.874169164	0.714285714	1	0.10442322
X2.6.Staff.Levels.NOS	950	0.758368868	0.625	0.96	0.099231506
X3.2.Chemotherapy	932	0.95661381	0.926470588	0.984375	0.02051514
X3.3.Consultants.SpecialistsNOS	626	0.910479434	0.846938776	0.93902439	0.03175478
X3.4.GP	612	0.708194367	0.629032258	0.8	0.052882501
X3.7.Investigations	502	0.849268957	0.795918367	0.895833333	0.031254686
X3.8.Oncology	480	0.95226603	0.9	1	0.033368419
X3.9.Pain.Management	476	0.967691388	0.9	1	0.03683331
X3.6.Hospital.Doctors.NOS	398	0.878002407	0.797619048	0.930555556	0.047122843
X3.11.Radiotherapy	377	0.909367554	0.86	0.976190476	0.032181695
X3.15.After.care	322	0.740435894	0.648648649	0.875	0.072739642
X3.16.Emotional.Social.MH.support	306	0.764214921	0.636363636	0.84	0.065749905
X3.18.Nursing	272	DNF	DNF	DNF	DNF
X3.19.Surgery	260	0.894447472	0.851851852	0.931372549	0.024575966
X4.1.Environment	234	0.758470729	0.675	0.815789474	0.043857964
X4.2.Food.Catering	164	0.825439106	0.666666667	0.962962963	0.092236386
X4.4.Travel	138	0.835957673	0.727272727	0.930232558	0.065273751
Mean scores		*0.82430967*	*0.745132381*	*0.897646469*	*0.049030083*
All-category mean SD		*0.08406785*	*0.099561809*	*0.083237105*	*0.026811242*

Table 12 - k-fold cross validation data for TREE algorithm.

Category	Test set n	Mean category accuracy (all algorithms)
X1.2.Improve.Clean	4626	0.757134709
X2.1.Com.Inter.Intra.Agency	2147	0.772730206
X2.2.Com.Pat.Prov	2026	0.759437755
X1.1.Positive.Clean	1708	0.787243003
X2.3.Wait.App	1340	0.792945803
X2.4.Wait.On.Day	1081	0.834729171
X2.5.Out.of.hours.Weekend.NOS	952	0.788002964
X2.6.Staff.Levels.NOS	950	0.773438389
X3.2.Chemotherapy	932	0.943679188
X3.3.Consultants.SpecialistsNOS	626	0.891698148
X3.4.GP	612	0.819780714
X3.7.Investigations	502	0.843266985
X3.8.Oncology	480	0.860110238
X3.9.Pain.Management	476	0.854125745
X3.6.Hospital.Doctors.NOS	398	0.879006328
X3.11.Radiotherapy	377	0.900396015
X3.15.After.care	322	0.756948803
X3.16.Emotional.Social.MH.support	306	0.751229938
X3.18.Nursing	272	0.914641791
X3.19.Surgery	260	0.864171999
X4.1.Environment	234	0.732535168
X4.2.Food.Catering	164	0.806142204
X4.4.Travel	138	0.828983298
Mean scores		0.822277329
All-category mean SD		*0.058845618*

Table 13 - k-fold cross validation mean category accuracy for all algorithms.

		Best performing algorithm			
Category	**Test n**	**Precision**	**Recall**	**F-score**	
X1.2.Improve.Clean	4626	0.835	0.83	0.83	
X2.1.Com.Inter.Intra.Agency	2147	0.77	0.77	0.765	
X2.2.Com.Pat.Prov	2026	0.77	0.77	0.765	
X1.1.Positive.Clean	1708	0.825	0.825	0.825	
X2.3.Wait.App	1340	0.875	0.86	0.865	
X2.4.Wait.On.Day	1081	0.935	0.935	0.935	
X2.5.Out.of.hours.Weekend.NOS	952	0.785	0.79	0.785	
X2.6.Staff.Levels.NOS	950	0.81	0.81	0.81	
X3.2.Chemotherapy	932	0.935	0.935	0.935	
X3.3.Consultants.SpecialistsNOS	626	0.855	0.88	0.865	
X3.4.GP	612	0.825	0.825	0.825	
X3.7.Investigations	502	0.87	0.875	0.87	
X3.8.Oncology	480	1	1	1	
X3.9.Pain.Management	476	0.95	0.93	0.935	
X3.6.Hospital.Doctors.NOS	398	0.93	0.935	0.93	
X3.11.Radiotherapy	377	0.935	0.94	0.935	
X3.15.After.care	322	0.795	0.795	0.795	
X3.16.Emotional.Social.MH.support	306	0.785	0.79	0.785	
X3.18.Nursing	272	0.94	0.94	0.94	
X3.19.Surgery	260	0.915	0.915	0.915	
X4.1.Environment	234	0.85	0.865	0.85	
X4.2.Food.Catering	164	0.825	0.85	0.805	
X4.4.Travel	138	0.875	0.875	0.87	
Mean scores		**0.864783**	**0.866957**	**0.862391304**	
SD		**0.06658**	**0.065065**	**0.067096768**	

		Best performing ensemble		Best algorithm vs. best ensemble	
Category	**Test n**	**Coverage**	**Recall**	**Precision/Coverage difference**	**Recall difference**
X1.2.Improve.Clean	2	1	0.82	0.165	-0.01
X2.1.Com.Inter.Intra.Agency	3	0.96	0.78	0.19	0.01
X2.2.Com.Pat.Prov	3	0.94	0.81	0.17	0.04
X1.1.Positive.Clean	3	0.91	0.82	0.085	-0.005
X2.3.Wait.App	3	0.91	0.84	0.035	-0.02
X2.4.Wait.On.Day	2	1	0.86	0.065	-0.075
X2.5.Out.of.hours.Weekend.NOS	2	1	0.64	0.215	-0.15
X2.6.Staff.Levels.NOS	2	1	0.77	0.19	-0.04
X3.2.Chemotherapy	3	1	0.93	0.065	-0.005
X3.3.Consultants.SpecialistsNOS	3	0.92	0.95	0.065	0.07
X3.4.GP	1	1	0.74	0.175	-0.085
X3.7.Investigations	3	0.96	0.9	0.09	0.025
X3.8.Oncology	3	0.91	1	-0.09	0

X3.9.Pain.Management	3	0.81	0.85	-0.14	-0.08
X3.6.Hospital.Doctors.NOS	3	1	0.93	0.07	-0.005
X3.11.Radiotherapy	3	0.96	0.96	0.025	0.02
X3.15.After.care	3	0.9	0.83	0.105	0.035
X3.16.Emotional.Social.MH.support	2	1	0.64	0.215	-0.15
X3.18.Nursing	4	0.93	0.93	-0.01	-0.01
X3.19.Surgery	3	0.97	0.92	0.055	0.005
X4.1.Environment	2	1	0.82	0.15	-0.045
X4.2.Food.Catering	2	1	0.74	0.175	-0.11
X4.4.Travel	2	1	0.84	0.125	-0.035
Mean scores	**3**	**0.96**	**0.84**	**0.095217391**	**-0.026956522**
SD		0.048544	0.092689	0.09046352	0.056736215

Table 14 - Best performing single algorithm vs best performing ensemble data.

Appendix B – Full taxonomy of categories for coding free-text material from WCPES.

Label	n
X1.1.Positive.Clean	3818
X1.2.Improve.Clean	2313
X1.3.Other.Clean	1183
X1.4.NOS.Total	1428
X1.5.NOS.Improve.Total	969
X1.6.NOS.Positive.Total	581
X3.18.Nursing	1074
X2.1.1.Com.Inter.Intra.Agency.Improve	197
X2.1.1.1.Com.Inter.Intra.Agency.Improve.NOS	165
X2.1.2.Com.Inter.Intra.Agency.Positive.NOS	44
X2.2.Com.Pat.Prov	1013
X2.2.1.Com.Pat.Prov.Improve	558
X2.2.1.1.Com.Pat.Prov.Improve.NOS	287
X2.2.1.1.1.Com.Pat.Prov.Info.Improve.NOS	142
X2.2.1.1.1.1.Com.Pat.Prov.Info.Treat.Improve.NOS	60
X2.2.1.2.Com.Pat.Prov.Manner.Improve.NOS	90
X2.2.1.2.1.Com.Pat.Prov.ManDiag.Improve.NOS	54
X2.2.2.Com.Pat.Prov.Positive	550
X2.2.2.1.Com.Pat.Prov.Positive.NOS	287
X2.2.2.1.1.Com.Pat.Prov.Info.Positive.NOS	75
X2.2.2.1.2.Com.Pat.Prov.Manner.Positive.NOS	216
X2.2.2.1.2.1.Com.Pat.Prov.Manner.Pers.NOS	194
X2.2.2.1.2.2.Com.Pat.Prov.Manner.Prof.NOS	74
X2.3.Wait.App	670
X2.3.1.Wait.App.Improve	366
X2.3.1.1.Wait.App.Improve.NOS	335
X2.3.2.Wait.App.Positive	333
X2.3.2.1.Wait.App.Positive.NOS	249
X3.19.Surgery	541
X2.4.1.Wait.On.Day.Improve	159
X2.4.1.1.Wait.On.Day.Improve.NOS	152
X2.4.2.Wait.On.Day.Positive.NOS	31
X3.6.Hospital.Doctors.NOS	476
X2.5.1.Out.of.Hours.Weekend.Improve.NOS	60
X2.5.2.Out.of.Hours.Weekend.Positive.NOS	8
X3.7.Investigations	475
X2.6.1.Staff.Levels.Improve.NOS	129
X2.6.2.Staff.Levels.Positive.NOS	1
X3.1.Anaesthesia	22

X3.1.1.Anaes.Improve	6
X3.1.2.Anaes.Positive	16
X3.3.Consultants.SpecialistsNOS	466
X3.2.1.Chemo.Improve	85
X3.2.1.1.Chemo.Com.Improve	27
X3.2.1.1.1.Chemo.Info.Improve	26
X3.2.1.1.Chemo.Improve.NOS	58
X3.2.2.Chemo.Positive	233
X3.2.2.1.Chemo.Com.Positive	61
X3.2.2.1.1.Chemo.Info.Positive	15
X3.2.2.1.2.Chemo.Manner.Positive	49
X3.2.2.2.Chemo.Positive.NOS	174
X3.4.GP	401
X3.3.1.Con.Spec.Improve	72
X3.3.1.1.Con.Spec.App.Speed.Improve	11
X3.3.1.2.Con.Spec.Com.Improve	45
X3.3.1.2.1.Con.Spec.Info.Improve	25
X3.3.1.2.2.Con.Spec.Manner.Improve	22
X3.3.2.Con.Spec.Positive	408
X3.3.2.1.Con.Spec.App.Speed.Positive	11
X3.3.2.2.Con.Spec.Com.Positive	133
X3.3.2.2.1.Con.Spec.Access.Positive	10
X3.2.2.2.2.Con.Spec.Info.Positive	45
X3.2.2.2.3.Con.Spec.Manner.Positive	101
X3.2.Chemotherapy	306
X3.4.1.GP.Improve	246
X3.4.1.1.GP.Care.Pdiag.Improve	69
X3.4.1.1.1.GP.Cond.Know.Improve	18
X3.4.1.1.2.GP.Serv.Prov.Improve	8
X3.4.1.2.GP.Diag.Improve	154
X3.4.1.2.1.GP.Diag.Com.Improve	13
X3.4.1.2.2.GP.Diag.Speed.Improve	39
X3.4.1.2.3.GP.Misdiag.Improve	35
X3.4.1.2.4.GP.Referral.Improve	80
X3.4.1.2.GP.Improve.NOS	32
X3.4.2.GP.Postive	161
X3.4.2.1.GP.Diag.Positive	51
X3.4.2.1.1.GP.Referral.Positive	41
X3.4.2.2.GP.Pdiag.Care.Positive	43
X3.4.2.3.GP.Positive.NOS	69
X3.5.Haematology	24
X3.5.1.Haem.Improve	2
X3.5.2.Haem.Positive	23
X3.15.After.care	290
X3.6.1.Hosp.Doc.Improve.NOS	73

X3.6.1.1.Hosp.Doc.Com.Improve.NOS	48
X3.6.1.1.1.Hosp.Doc.Info.Improve.NOS	25
X3.6.1.1.2.Hosp.Doc.Lang.Improve.NOS	4
X3.6.1.1.3.Hosp.Doc.Manner.Improve.NOS	32
X3.6.1.2.Hosp.Doc.Levels.Improve.NOS	16
X3.6.2.Hosp.Doc.Positive.NOS	411
X3.6.2.1.Hosp.Doc.Com.Positive.NOS	144
X3.6.2.1.1.Hosp.Doc.Info.Positive.NOS	11
X3.6.2.1.2.Hosp.Doc.Manner.Positive.NOS	135
X3.11.Radiotherapy	251
X3.7.1.Invest.Improve	288
X3.7.1.1.Invest.Improve.NOS	102
X3.7.1.2.Invest.Speed.Improve	132
X3.7.1.3.Invest.Initial.Improve	56
X3.7.1.4.Invest.Mis.Improve	36
X3.7.1.5.Invest.Wait.Results.Improve	76
X3.7.1.6.Invest.Diag.Wait.NOS	41
X3.7.1.7.Invest.Follow.Results.Improve.NOS	4
X3.7.1.8.Invest.Treat.Result.Improve.NOS	9
X3.7.2.Investigations.Positive	198
X3.7.2.1.Invest.Wait.Results.Positive	5
X3.7.2.2.Invest.Positive.NOS	81
X3.7.2.3.Invest.Speed.Positive	57
X3.7.2.4.Invest.Screening.Positive	62
X3.7.2.4.1.Invest.Screen.Bowel.Positive	24
X3.7.2.4.2.Invest.Screen.Breast.Positive	28
X4.1.Environment	240
X3.8.1.Onc.Improve	31
X3.8.1.1.Onc.Com.Improve	20
X3.8.1.1.1.Onc.Info.Improve	13
X3.8.1.1.2.Onc.Manner.Improve	7
X3.8.2.Onc.Positive	90
X3.8.2.1.Onc.Com.Positive	16
X3.8.2.1.1.Onc.Access.Positive	2
X3.8.2.1.2.Onc.Info.Positive	7
X3.8.2.1.3.Onc.Manner.Positive	10
X2.1.Com.Inter.Intra.Agency	238
X3.9.1.Pain.Man.Improve	73
X3.9.1.2.Pain.Chronic.Improve	11
X3.9.1.3.Pain.Disch.Improve	7
X3.9.1.4.PainWaitImprove	28
X3.9.2.PainManagePositive	10
X3.10.Physiotherapy	32
X3.10.1.Physio.Improve	12
X3.10.2.Physio.Positive	12

X2.4.Wait.On.Day	188
X3.11.1.Rad.Improve	67
X3.11.1.1.Rad.Com.Improve	29
X3.11.1.1.1.Rad.Info.Improve	24
X3.11.1.1.2.Rad.Manner.Improve	6
X3.11.1.2.Rad.Improve.NOS	40
X3.11.2.Rad.Positive	191
X3.11.2.1.Rad.Com.Positive	73
X3.11.2.1.1.Rad.Info.Positive	20
X3.11.2.1.2.Rad.Manner.Positive	65
X3.11.2.2.RadPositiveNOS	120
X3.12.Respiratory	19
X3.12.1.Resp.Improve	6
X3.12.2.Resp.Positive	14
X3.13.Urology	39
X3.13.1.Uro.Improve	7
X3.13.2.Uro.Positive	34
X3.14.A.E	41
X3.14.1.A.EImprove	33
X3.14.2.A.EPositive	8
X4.4.Travel	161
X3.15.1.Aftercare.Improve	199
X3.15.1.1.Aftercare.E.S.MH.Improve	19
X3.15.1.2.Aftercare.Improve.NOS	157
X3.15.1.3.Aftercare.Invest.Follow.Improve	26
X3.15.2.Aftercare.Positive	97
X3.15.2.1.Aftercare.Positive.NOS	82
X3.15.2.2.Aftercare.Invest.Follow.Positive	17
X4.2.Food.Catering	153
X3.16.1.EmSocMH.Improve	94
X3.16.2.Emotional.Social.MHPositive	43
X3.17.Palliative.Care	16
X3.17.1.PalliativeCareImprove	4
X3.17.2.PalliativeCarePositive	12
X3.16.Emotional.Social.MH.support	136
X3.18.1.Nurs.Improve	388
X3.18.1.1.Nurs.Avail.Improve.NOS	31
X3.18.1.2.Nurs.Com.Improve	70
X3.18.1.2.1.Nurs.Com.Improve.NOS	48
X3.18.1.2.1.1.Nurs.Info.Improve.NOS	4
X3.18.1.2.1.2.Nurs.Manner.Improve.NOS	46
X3.18.1.2.2.Nurs.Info.Improve	7
X3.18.1.2.3.Nurs.Manner.Improve	64
X3.18.1.2.Nurs.Breast.Improve	16

X3.18.1.2.1.Nurs.Breast.Avail.Improve	8
X3.18.1.3.Nurs.CNS.Improve	10
X3.18.1.3.1.Nurs.CNS.Avail.Improve	7
X3.18.1.4.Nurs.District.Improve	28
X3.18.1.4.1.Nurs.Dist.Avail.Improve	11
X3.18.1.5.Nurs.Key.Improve	20
X3.18.1.5.1.Nurs.Key.Avail.Improve	18
X3.18.1.6.Nurs.MacMil.Improve	18
X3.18.1.6.1.Nurs.MacMil.Avail.Improve	9
X3.18.1.7.NursSpecialImprove.NOS	15
X3.18.1.7.1.NursSpecialAvailImprove	13
X3.18.1.8.NursImproveNOS	127
X3.18.1.8.1.NursCareImproveNOS	70
X3.18.1.9.NursOutOfHoursImprove	24
X3.18.1.10.NursLevelsImprove	124
X3.18.2.Nurs.Positive	785
X3.18.2.1.Nurs.Com.Positive	245
X3.18.2.1.1.Nurs.Com.Positive.NOS	2
X3.18.2.1.2.Nurs.Info.Positive	65
X3.18.2.1.3.Nurs.Info.Positive.NOS	38
X3.18.2.1.4.Nurs.Manner.Positive	201
X3.18.2.1.4.1.Nurs.Manner.Positive.NOS	152
X3.18.2.2.Nurs.Breast.Positive	51
X3.18.2.2.1.Nurs.Breast.Manner.Info.Positive	18
X3.18.2.3.Nurs.Chemo.Positive	68
X3.18.2.3.1.Nurs.Chemo.Info.Positive	8
X3.18.2.3.2.Nurs.Chemo.Manner.Positive	25
X3.18.2.4.Nurs.CNS.Positive	44
X3.18.2.4.1.Nurs.CNS.Manner.Info.Positive	18
X3.18.2.5.Nurs.Dist.Positive	48
X3.18.2.5.1.Nurs.Dist.Manner.Positive	10
X3.18.2.6.Nurs.Key.Positive	18
X3.18.2.7.Nurs.MacMil.Positive	44
X3.18.2.7.1.NursMacMilManner.InfoPositive	14
X3.18.2.8.NursSpecialPositive.NOS.	70
X3.18.2.8.1.NursSpecialInfoPositive.NOS.	6
X3.18.2.8.2.NursSpecialMannerPositive.NOS.	19
X3.18.2.9.NursPositiveNOS	402
X2.6.Staff.Levels.NOS	130
X3.19.1.Surg.Improve	181
X3.19.1.1.Surg.Cancel.Delay.Improve	18
X3.19.1.2.Surg.Com.Improve	67
X3.19.1.2.1.Surg.Info.Improve	35
X3.19.1.2.2.Surg.Lang.Improve	3
X3.19.1.2.3.Surg.Manner.Improve	16

X3.19.1.3.Surg.Improve.NOS	17
X3.19.1.4.Surg.Follow.Improve	17
X3.19.1.5.Surg.PostOp.Improve	53
X3.19.1.5.1.Surg.PostOp.Pain.Improve	16
X3.19.1.5.2.Surg.PostOp.Rec.Improve	48
X3.19.1.6.Surg.PreOpImprove	17
X3.19.1.7.Surg.Proced.Improve	15
X3.19.2.Surg.Positive	393
X3.19.2.1.Surg.Appoint.Speed.Positive	78
X3.19.2.2.Surg.Com.Positive	62
X3.19.2.2.1.Surg.Info.Positive	19
X3.19.2.2.2.Surg.Manner.Positive	46
X3.19.2.3.Surg.Positive.NOS	186
X3.19.2.4.Surg.PostOp.Positive	46
X3.19.2.4.1.Surg.PostOp.Rec.Positive	38
X3.19.2.5.Surg.PreOp.Positive	20
X3.19.2.6.Surg.Proced.Positive	61
X3.8.Oncology	117
X4.1.1.Env.Cleaning.Staff	22
X4.1.2.Env.Improve	182
X4.1.2.1.Env.Bed.Levels.Improve	58
X4.1.2.2.Env.Hosp.Clean.Improve	22
X4.1.2.3.Env.Hosp.Toilet.Improve	16
X4.1.2.4.Env.Hosp.Privacy.Improve	22
X4.1.3.Env.Positive	53
X4.1.3.1.Env.Bed.Levels.Positive	1
X4.1.3.2.Env.Hosp.Clean.Positive	18
X3.9.Pain.Management	82
X4.2.1.Food.Cat.Improve	128
X4.2.2.Food.Cat.Positive	26
X4.3.Finances	36
X4.3.1.Finances.Improve	34
X4.3.2.Finances.Positive	3
X2.5.Out.of.hours.Weekend.NOS	69
X4.4.1.Travel.Improve	122
X4.4.1.1.Amb.Trans.Improve	9
X4.4.1.2.Parking.Improve	28
X4.4.2.Travel.Positive	45
X4.4.2.1.Amb.Trans.Positive	23
X4.4.2.2.Parking.Positive	2

Table 15 - WCPES taxonomy of free-text responses with n of true positive results.