Development and testing of a text-mining approach to analyse patients' comments on their experiences of colorectal cancer care

Word count: Abstract: 275; Manuscript 4864 words

ABSTRACT

Background: Quality of cancer care may greatly impact upon patients' health-related quality of life (HRQoL). Free-text responses to patient-reported outcome measures (PROMs) provide rich data but analysis is time and resource-intensive. This study developed and tested a learning-based text-mining approach to facilitate analysis of patients' experiences of care and develop an explanatory model illustrating impact upon HRQoL.

Methods: Respondents to a population-based survey of colorectal cancer survivors provided free-text comments regarding their experience of living with and beyond cancer. An existing coding framework was tested and adapted, which informed learning-based text mining of the data. Machine-learning algorithms were trained to identify comments relating to patients' specific experiences of service quality, which were verified by manual qualitative analysis. Comparisons between coded retrieved comments and a HRQoL measure (EQ5D) were explored.

Results: The survey response rate was 63.3% (21,802/34,467), of which 25.8% (n=5634) participants provided free-text comments. Of retrieved comments on experiences of care (n=1688), over half (n=1045, 62%) described positive care experiences. Most negative experiences concerned a lack of post-treatment care (n=191, 11% of retrieved comments), and insufficient information concerning self-management strategies (n=135, 8%) or treatment side effects (n=160, 9%). Associations existed between HRQoL scores and coded algorithm-retrieved comments. Analysis indicated that the mechanism by which service quality impacted upon HRQoL was the extent to which services prevented or alleviated challenges associated with disease and treatment burdens.

Conclusions: Learning-based text mining techniques were found useful and practical tools to identify specific free-text comments within a large dataset, facilitating resource-efficient qualitative analysis. This method should be considered for future PROM analysis to inform policy and practice. Study findings indicated that perceived care quality directly impacts upon HRQoL

Key words: text-mining; PROMs; quality of life; colorectal cancer; machine learning; machine learning algorithms; thematic analysis; thematic content analysis; qualitative methods.

INTRODUCTION

Patient-reported outcome measures (PROMs) and experience measures (PREMs) are increasingly being used in Europe and North America to ascertain patients' views, including those with cancer, concerning symptoms, functional status, and health-related quality of life (HRQoL)[1,2]. This has been motivated by an understanding that quality and effectiveness of care is best determined from the patient's perspective[3,4]. Although often under-utilised, free-text comments from patients can complement quantitative measures by providing information on experiences not covered by the specified measures, and/or give more detail that may help contextualise responses to closed questions. In the UK, the NHS Cancer Reform Strategy[5] and Outcomes Strategy for Cancer[6] identified the important role of patient-reported intelligence in measuring and improving clinical quality, and national surveys have been undertaken to determine the quality of experience of cancer patients and survivors[7,8]. The PROM programme was recently extended to the first national cancer site specific survey, covering colorectal cancer (CRC)[9].

Researchers commonly include open-ended questions at the end of PROMs for respondents to leave comments[10]. Previous analysis of such responses has highlighted the physical, emotional and social challenges to health-related quality of life (HRQoL) often faced by cancer survivors[11]. Nevertheless, resource implications exist for free-text analysis, both in fiscal and temporal terms due to the large volumes of data generated[10,12]. Consequently, raw free-text data from large scale surveys, such as the national cancer patients experience survey (CPES) in England, are often not analysed in any systematic way and potential insights consequently lost.

For this reason, automated data sorting into broad categories prior to more detailed qualitative analysis is beneficial. Learning-based approaches to text mining, using 'supervised' machine-learning algorithms, derive data regarding specific topics of interest from large sets of textual data, typically through identification of patterns and trends using statistical pattern learning[13]. Algorithms are 'supervised' in that they are presented with a set of pre-coded ('labelled') data as belonging to different categories (i.e. different aspects of cancer patient experience), from which they learn to recognise patterns of text within a larger dataset. Text-mining has been used previously to process large amounts of online data from social networks[14,15,16,17]. A simpler type of text-mining (keyword-in-context analysis) has also been used in online and media portrayals of heath topics[18,19,20] and small scale surveys[21]. The Information Strategy for NHS England encourages analysis of data from patients[22].The novel application of 'supervised', learning-based text-mining can help facilitate this aim by enhancing greatly the ability of researchers to work with large amounts of free-text data, thus allowing issues of concerns raised by patients to shape recommendations for service improvements.

Previous work has shown patients with colorectal cancer experience much lower emotional and social functioning and greater financial difficulties than the general population, especially amongst younger patients[23]; difficulties socialising due to physical problems including unreliable stoma or altered bowel movements[24]; delays in return to work that exacerbate financial problems and social isolation[25,26]; anxiety, depression and fear of recurrence[24,27]; physical treatment side-effects such as impaired cognition, pain, fatigue, changed bowel habit, and sexual dysfunction[24]; psychological distress that manifests amongst patients with recently formed ostomies[28], often diminishing body image and confidence in sexual attractiveness[29,30].

HRQoL is a multi-dimensional concept that includes physical, mental, emotional and social functioning, and focuses on the impact health status has on quality of life. The quality of care

received (e.g. whether good or bad), as experienced and perceived by patients before, during and post-treatment, may have either a beneficial or detrimental impact upon a patient's HRQoL[11]. The aims of this investigation were to; (1) develop a learning-based text-mining approach to facilitate analysis of patients' free-text responses to a national CRC specific PROM survivorship survey, relating to their experiences of care quality; and (2) to develop a model from that analysis to illustrate the impact that subjective experiences of the quality of received care has upon HRQoL.

METHODS

Study design

A population-based postal survey undertaken in 2013 of all individuals aged ≥16 years in England who survived 12-36 months following diagnosis of colorectal cancer in 2010 or 2011.

Cohort identification

Individuals were identified via the National Cancer Registry Service (NCRS). Cases were excluded if they were not known to have a UK address.

Questionnaire design and content

Questionnaires included questions on socio-demographics, treatment, disease status, physical activity, long-term conditions (LTCs), EQ-5D[31], Social Difficulties Inventory (SDI)[32], Functional Assessment of Cancer Therapy (FACT) (colorectal cancer specific outcomes)[33]. The free-text comments box was placed at the end of the questionnaire, following the closed questions, with the header: '*If you have anything else you would like to tell us about living with and beyond cancer, please do so here:*'

Survey process

Individuals were sent questionnaires from the Cancer Centre identified as having treated them. Individuals consented to participation by returning completed questionnaires. Two reminders were sent to non-responders. A dedicated free telephone helpline was provided to resolve queries.

Ethics and governance

Approval was given by the National Information Governance Board (NIGB). to perform the survey (ref:ECC 5-02(FT8)/12). All analyses presented here were conducted on anonymised respondent data.

ANALYSIS

There were three phases to analysis (figure 1): firstly, primarily deductive development of a thematic framework to categorise comprehensively the survey comments; secondly, application of machine-learning algorithms to identify patients' comments concerning their experience of care quality using the 'RTextTools'[34] package for 'R Statistical Computing' software[35]; and thirdly, qualitative analysis of retrieved comments in the NVivo qualitative data analysis package[36].

Phase 1: A random sample of comments (n=400) was triple coded by three researchers (AR, KH, RW), (two female/one male), each with more than ten years' experience with qualitative research. Data were deductively coded against an abridged version of an existing thematic framework designed for the PROM pilot study[11]. The framework allows comments to be coded as positive or negative experiences of specific areas of care (i.e. timely/delayed diagnosis; good/inadequate post-treatment care) and whether specific forms of information to prepare patients were lacking (i.e. lack of information on treatment side-effects; psychological impact of cancer and treatment; self-management strategies). Consistency of inter-rater coding (Cohen's Kappa) ranged from substantial (0.64) to excellent (0.87).

Inconsistencies between the existing framework and data were discussed between researchers, with disagreements resolved by a fourth researcher (JC).

Phase 2: Algorithms were then trained and tested to identify comments within the whole dataset that specifically related to positive and negative experiences of care quality. The quality of the algorithm results depend upon three factors[37]. Firstly, the quality of the data provided may make it more 'difficult' or 'easy' it is to identify patterns. Secondly, each algorithm is governed by different sequential sets of rules for identifying semantic or grammatical relationships within the text, and particular algorithms may suit some datasets better than others may. Thirdly, the larger the training sets used the more accurate are algorithms likely to be at identifying similar comments within the wider dataset, but trade-offs with time and human coding are necessary to ensure the method is resource-efficient[37].

To optimise the number of coded comments available to train algorithms, a second random sample of comments (n=400) (random sample 2) was coded to provide a combined term document matrix (TDM) of 800 comments, resulting in a two-dimensional grid of rows (samples) and columns (terms) representing the frequency that a term appears in the samples. The following 'R' settings in the Term Document Matrix {tm} package were used in creating the TDM: removeNumbers=TRUE, stemWords=TRUE, weighting=tm::weightBin. No further processing was applied to this document[38].This TDM was then used as a template for mining the remaining comments in the database. A 50% sample was drawn randomly from the 800 to train algorithms, the remainder used to test them. A 10-fold cross validation (a technique in which an original sample is randomly split into ten subsamples, with training conducted in nine datasets, testing on one, and the process repeated ten times) was then used to assess algorithm performance[38]. Algorithm performance is measured as sensitivity (true positives/(true positives + false negatives)), precision (true positives/(true positives+ false positives)) and by the f-score ((2*sensitivity*precision)/ (sensitivity + precision))[39]. Sensitivity describes the ability to identify all relevant comments of a given

category; precision defines the ability to exclude non-relevant comments. The f-score describes overall performance, representing the harmonic mean of precision and sensitivity[39]. Comparative analysis using t-tests was then conducted between categories of individuals' comments and their single index EQ5D score (summarising five domains: mobility; self-care; usual activities; pain/discomfort; anxiety/depression), to identify associations between them.

Phase 3: All retrieved comments were then read and double coded by two researchers (AR, RW) to determine their relevance for each category of care experience, with a third researcher (JC) supporting interpretive analysis. Individual respondents often provided comments that were coded into more than one category, and some reported both positive and negative experiences, which were coded accordingly. Any disagreements between coders were resolved through discussion. Comments coded as irrelevant often described a patient's treatment journey but without conveying sentiments regarding their quality (e.g. whether experiences were positive or negative).

Patterns and relationships were explored between reported care quality in individuals' comments and their HRQoL, as measured by their summary EQ5D scores. Comments coded in each category were examined for references made to other themes and any reported relationships between them. These were explicit within many comments, but latent in others. Emerging patterns were then tested against all comments in relevant categories and a tentative model generated to illustrate where data from a number of participants indicated such patterns existed (figure 2)[41]. The concepts of validity and trustworthiness within qualitative research are understood in terms of the credibility and dependability of the analytical process[42,43]. The emerging model was tested against all individual negative and positive comments to challenge its capacity to represent participants' experiences. Selected quotes are used to illuminate particular types of experience.

FINDINGS

Of 21,802 survey participants, a quarter (n= 5634, 25.8%) added comments in the free-text box. Table 1- compares socio-demographic characteristics of participants who provided comments and those who did not. All demographic variables were significantly associated with whether an individual provided comments except for tumour site. Those more likely to comment were older participants (p<.001 for trend across age groups), females (p=.001), less socially deprived (p<.001 for trend across deprivation quintiles), those diagnosed in 2010 (p=.023), and those with higher Duke's stage at diagnosis (p=.014 for trend across stages).

Length of comments ranged between one and 225 Words. Wide variation existed in the number of comments from participants who had received care from the 145 different healthcare providers (English NHS Trusts), with a range of 3–140 (mean 38; median 35). We first report on the development of the text-mining approach used to analyse the free-text dataset and the number of comments retrieved, and then report findings from the free-text analyses.

Development of text-mining approach

In the first analytical phase, the coding framework previously developed within the pilot PROM study[11] was found to comprehensively code care experience comments in the colorectal cancer dataset. Of the 800 manually coded comments, 248 were coded as reporting positive care experiences (*Excellent/good treatment/ caring staff*), and 85 as reporting negative experiences (*poor staff attitude/lack of care*). Extrapolation to the remaining comments in the dataset (n=4834) suggested that algorithms should be expected to find approximately 1436 positive comments (29.7% of all comments) and 491 negative comments (10.1%) in total (Table 2).

In Phase 2 initial testing indicated that some comment categories were easier to detect than others, and that algorithms trained to identify both positive and negative experience coded

comments outperformed separate algorithms for each. Table 3 indicates the precision, sensitivity and f-scores calculated for each of the seven algorithms used. Highest overall sensitivity was found for the SVM (support vector machines) algorithm (78%), with precision considerably higher (83.5%), and an overall f-score of 80%. Based on the training and testing we would have expected approximately 1360 comments to be retrieved from the remaining comments (n=4834) through this one algorithm.

As each algorithm identifies comments by using different sets of rules to recognise semantic and grammatical patterns, they are each likely to retrieve varying numbers of comments regarding any specific theme[34]. Only 313 comments were identified by all seven algorithms. However, 2645 records had at least one algorithm indicating they contained an care experience comment. Therefore, to optimise the number of retrieved comments from the yet unclassified dataset (n=4834), we retrieved those that had been identified separately by the four best performing algorithms (SVM; RF (Random Forests); TREE (Decision Trees); GLMNET (Generalised linear models network)). This retrieval strategy identified 2076 comments, of which 1688 (81.3%) were found to be relevant, and 388 (18.7%) irrelevant. Relevant comments comprised 1045 coded as positive, fewer comments (n=391, 8.4%) than expected from manual coding extrapolations, and 643 coded as negative, more comments (n=152, 3.2%) than expected (Table 2). The content of positive comments was usually much less specific than for negative comments, and algorithms could only effectively identify two positive comment subcategories, with 109 (6%) comments describing timely diagnosis and 289 (17%) describing post-treatment care in positive terms. Although positive comments did exist within other subcategories (e.g. coordination of care), they were too few to adequately train algorithms to identify them in the dataset. Algorithms were more successful in identifying subcategories for negative comments. The highest proportions of negative experiences related to inadequate post-treatment care (n=191, 11%), and to lack of information provided by staff concerning treatment side-effects (n=160, 9%) and self-management strategies (n=135, 8%).

Validation of results

A generalised approach to compare the performance of algorithms is the 10-fold cross validation which divides the data sample into 10 blocks or folds and builds ten evaluations on each fold in order to estimate the error variance. The feature to perform this validation in the RTools was used to validate the results obtained with the original 50/50 approach adopted. Results of the performance variations between the two methods described in Table 3 show that most algorithms remained unchanged except for the Maxentropy which produced dramatically different results (0.014 10-fold compared to 0.67 in the original approach), a 25% reduction in the score for GLMNET, and 20% improvement for Logitboost, which became the best performing algorithm. The use of four best performing algorithms to retrieve comments from the unclassified dataset, which performed well in the 10-fold cross validation, improved the retrieval strategy and controlled the limitations of the 50/50 approach.

Comparing coded experiences of care with reported summary EQ5D scores

Mean summary EQ5D scores were explored to determine whether individuals' self-reported HRQoL could be correlated with their reported experiences of care quality. The mean summary EQ5D score for all survey respondents who left comments ('All comments') on any subject was 0.78, but was higher (0.85) for those respondents who provided 'positive comments' and lower (0.74) for those providing 'negative comments' (Table 2). T-tests demonstrated these to be significant associations (p=<.001), indicating that those reporting positive experiences of care had correspondingly higher HRQoL than the mean for all free-text respondents, and those reporting negative experiences. Associations were also found between mean EQ5D scores for 'All comments' and the comment subcategories: timely diagnosis (p=<.001); positive post-treatment care (p=<.001); inadequate post-treatment care (p=.002).

Experiences of care quality and their reported impact upon quality of life

Given the quantitative evidence that lower HRQoL scores were associated with negative comments of care quality, in phase 3 researchers analysed comments to better understand this relationship: how did participants with lower or higher reported HRQoL describe their care? Any explanatory theory must include reference to mechanisms by which relationships between variables are generated[40]. A tentative model was developed to illustrate relationships between themes emerging from the data (figure 2), in which evidence from comments suggested the link between HRQoL and different experiences of care quality. The top half of the figure represents those aspects of care associated by respondents with positive experiences of care, and which reportedly minimised or addressed challenges faced along the treatment pathway. The lower part of the figure identifies issues related to experiences of care that had a negative impact upon respondents' HRQoL.

Timeliness of diagnosis

Text-mining identified 109 (6.4%) comments describing their path to diagnosis as timely, compared with 36 comments (2.1%) describing it as delayed. Comments indicated that participants recognised a timely diagnosis had contributed to successful treatment outcomes, thus positively impacting on HRQoL:

> *The early diagnosis of cancer and treatment has been essential to my excellent recovery. It was discovered after giving blood. I have returned to work a year ago and I have had no time off at all since despite going back early.* (Male, age 45-55)

Other participants reported having taken advantage of the opportunity presented by the mailed national bowel screening, allowing them to become involved actively in diagnosis. Participants who experienced diagnostic delays reported GPs as attributing symptoms to conditions other than colorectal cancer, or did not consider bowel cancer due to the patient's relative youth. Comments suggest experiences of delayed or mis-diagnosis creates a perception of longer or more complex treatment.

I receive help from my current GP but if my previous GP had not ignored all the clear signs of my cancer that I presented with over a period of one year I would not have had to undergo all the treatment etc. that followed the eventual diagnosis. (Male age 55-64)

Absence of coordinated/integrated patient care

Algorithms identified 78 (4.6%) comments relating to lack of coordinated care. Comments described delays that had a detrimental impact on available treatment options, discontinuity between GP practices and the hospital at the end of treatment, 'confusion' regarding which health professionals had responsibility to provide certain aspects of on-going care, absent care plans and limited access to community-based clinicians. Thus, the emotional impact of cancer and its treatment were not resolved, and sometimes represented a source of increasing stress over time. However, a lack of coordination and communication between Health professionals and respondents could lead to delays that had a detrimental impact on available treatment options:

I was told by my surgeon that I would have a follow-up appointment 3/12 after operation and she would then refer me to an oncologist to see if I needed any [adjuvant] treatment. At that point the surgeon was sick and it was cancelled. 5/12 after the op after much agitation by my GP I was seen by another rectal cancer surgeon who assured me that I had been referred to the oncologist. Four days later I was seen by the oncologist who said he had had no referral and it was too late to think about treatment. (Female, age 55-64)

Algorithms failed to identify positive comments of coordinated care, suggesting few existed. Nevertheless, examples were found amongst positive comments that described good communication between staff from different departments across the treatment journey, often facilitated by a specialist nurse.

> *The role of the colorectal nurse in providing on-going contact and reassurance is absolutely vital in helping patients understand their condition, what is happening to them and to bolster their morale before during and after surgery.* (Male, age 44-55)

Lack of patient preparation for cancer and its treatment

Algorithms identified comments in three separate sub-categories relating to inadequate patient preparation. Some participants reported insufficient information concerning treatment side-effects (n=160 comments, 9.4%), the possible psychological impact of cancer and treatment (43 comments, 2.5%) and potential self-management strategies (135 comments, 7.9%) to ameliorate these. Comments reporting inadequate patient preparation often described how this could exacerbate physical and emotional problems. For example, some participants reported being insufficiently informed about available cancer therapies, with some individuals subsequently uneasy about their treatment decisions.

> *The one area which has given me major problems has been the severity of the discomfort I have experienced since the stoma reversal. More information in that area would have been very helpful. My quality of life has suffered more since the stoma reversal than at any other time since the initial surgery. Had I realised how severe the reaction would be I might have elected to retain the stoma.* (Male, age 54-65)

Comments indicated that without preparation, the negative impact of emotional problems experienced following cancer treatment could be more severe as they were less equipped to deal with them. From amongst the positive comments, examples could be found where individuals reported having been prepared by services for potential problems, and were thus better able to cope with these challenges.

Restricted opportunities for emotional support

Algorithms identified 78 comments (4.6%) that described a lack of emotional support, including sign-posting to cancer support groups, talking therapies and counselling. The need for emotional support transcended the patient journey from diagnosis, through treatment and post-treatment. When support was not available emotional and psychological problems could worsen. Ultimately, as indicated in Figure 2, if support needs were not addressed as they arose during treatment a greater need for care post-treatment might arise.

> *I did and still do feel 'abandoned' following surgery and treatment for colon cancer. I appreciate that the oncology and surgical departments are very busy but I would have liked some form of counselling following discharge. The anxiety doesn't go away, it just gets worse.* (Female, age 65-74)

Although the algorithms were unable to identify positive descriptions of patient preparation, amongst positive comments examples were found of participants who had been provided with emotional support as part of their care and reported its importance to their HRQoL.

Quality of care post-treatment

Algorithms identified sub-categories of comments relating to both good (289 comments, 17.1%) and inadequate post-treatment care (191 comments, 11.3%). Retrieved comments related primarily to accessibility of health professionals after finishing treatment, but also regular monitoring and follow-up consultations to detect recurrence. Both elements appeared to have a direct impact upon HRQoL. Participants often contrasted '*poor at best*' or '*non-existent'* care following treatment with very good '*hot house'* care during treatment.

> *The effects of my treatment still affects me now but there is nowhere to go to access any support. This is so important as no one explained the emotional impact of living beyond bowel cancer after treatment or what side effects to expect of which there are many causing undue stress. A lot more support is needed for post cancer treatment.* (Female, age 44-55)

When participants felt supported and monitored beyond the treatment phase they reported satisfaction and confidence in their care *(e.g.* '*My needs are met'* and '*reassurance that problems will be dealt with'*). Thus, psychological problems such as fears of cancer recurring were partly allayed and participants were able to '*move on'/* '*plan for the future'*. However, such challenges were not addressed for those participants who felt abandoned by services due to a perceived lack of coordination between primary and secondary care. The lack of post-treatment support reported by some participants meant an added burden for carers.

> *Would have liked more help and support after my surgery. Have been left to cope on my own. Would not have been able to live without my wife who took on everything. Not visited by district nurse on return from hospital. Lost records. Not called back for further check-up. Wife had to keep ringing.* (Male, age 64-75)

Discussion

This study had two objectives: to develop and test a text-mining approach to facilitate analysis of free-text comments within the national colorectal PROM survivorship survey; and to develop an explanatory model to illustrate the impact of experiences of care quality upon HRQoL. Regarding the first study objective, analysis of this large dataset of comments was facilitated by the application of learning-based text-mining techniques. Formal analysis of free-text survey comments has previously generated important insights in to the experience of participants[10,11,12]. However, to our knowledge this is the first time text-mining techniques have been used to facilitate a resource-efficient analysis of participants' free-text responses to a PROM exploring experiences of cancer survivors. What is significant about this study is the utilisation of a three stage analytical process that reduced the number of comments that required analysis to those specifically related to issues of interest, thus rendering a large number of free-text responses manageable. First, a thematic framework was evaluated for qualitatively coding data, in this instance an existing framework developed within a pilot study[11]; secondly, learning-based text-mining was used to identify and

retrieve a subset of comments from a larger dataset based upon that framework; and thirdly, that subset was subjected to further thematic analysis to determine patterns within it.

Comprehensive reading of all comments within the dataset remains the 'gold standard' method for analysing free-text comments, and is currently the only way to ensure all relevant comments are coded and analysed. Based upon extrapolation from the manually coded sample we estimated more positive and fewer negative comments within the whole dataset than were found. A high proportion of positive comments were non-specific, and the algorithms were unable to identify positive comments for some sub-categories for which they could identify negative comments. A risk exists, therefore, that analysis may be biased towards negative experiences. Sensitivity bias may also have been present. It is unknown whether algorithms were able to retrieve all relevant comments from the dataset, as comprehensive manual coding of all comments was not conducted. This may be a limitation to be balanced against the value of resource-efficient analysis. Nevertheless, while some relevant comments (relating to experience rather than other issues) were missed, sufficient numbers were retrieved to provide insight into participants' experiences and with few time and resource demands. Moreover, 88% of retrieved comments were subsequently found relevant indicating some efficiency of method. Efforts now need to be made in future studies to improve text-mining retrieval rates and determine greater accuracy of retrieval. Nevertheless, the text-mining method used in this study would be easily transferable to other patient experience and outcome surveys, both in the UK and elsewhere, allowing the issues of most concern to patients as expressed in their comments to be included in analysis and to influence policy recommendations.

The ultimate purpose of PROM surveys is to facilitate care quality improvement by benchmarking outcomes for patients over time and assisting comparisons between providers[2]. More comments were positive (57%) than negative (35%). Nevertheless, it was not possible to conduct health care provider profiling using these comments as the

proportion of responses describing experiences of care quality varied greatly across the 145 English NHS trusts from which participants received treatment and care, with very few or no responses from some trusts. Moreover, the free-text question used in the survey was very broad, with participants describing many issues unrelated to quality of care. A differently designed question might elicit responses focussing on care quality to facilitate provider profiling if this was the focus of the survey. Survey designers thus need to be clear of the purpose of including free-text comments, the type of information they seek from participants and have an explicit strategy for analysis.

Regarding the second objective, our analysis focussed upon the positive and negative experiences participants reported of the quality of care they received. While many comments were short and vague offering limited insights, many others contained rich data. More than half retrieved comments (n=1045, 57%) described positive experiences, though negative comments more often contained specific details of the quality of care they experienced and what went wrong for respondents. Comments described physical, psychological and social challenges faced by individuals with colorectal cancer that have been identified in previous studies[23,24,27], though comments also indicated the prevalence of such issues. Respondents who provided comments describing positive experiences of their care also reported significantly higher HRQoL than those who described negative experiences. However, while these correlations were highly statistically significant, no causal direction could be confirmed. Other potentially important factors are participants' demographic characteristics, almost all of which were found significantly associated with whether participants provided comments. Demographic variables may also impact upon the types of comments individuals provided, whether positive or negative. This issue should be addressed in further research.

Qualitative analysis explored the comments for relationships and patterns between themes that might explain the associations between coded comments and reported HRQoL[41]. A

tentative model was developed to illustrate these findings (Figure 2). As the model suggests, if respondents reported they had not experienced good care in the form of timely diagnosis, coordinated care, adequate patient preparation and emotional support throughout their treatment journey, the challenges they faced would not be adequately addressed, with increasing negative impact on HRQoL over time. Previous experiences of poor care during treatment might also exacerbate later stresses associated with inadequate post-treatment care. Furthermore, if respondents reported that the challenges they faced had not been adequately addressed, a greater burden was placed upon informal carers and unresolved problems might increase demand for post-treatment services. Our findings therefore emphasise connections between HRQoL and the specific aspects of quality care that were most important to respondents. They also identify areas of care health providers should prioritise to ensure patients experience both improved HRQoL and health outcomes.

Conclusions

This study has shown text-mining techniques successfully identify free-text comments relating to specific themes of interest to policy-makers, care-providers and researchers within a large dataset. The method used in this study facilitates qualitative analysis economic in resources and time and would be transferable to any other national patient experience and outcome surveys. Analysis of comments within this study provided insights into significant associations found between participants' comments relating to the quality of their care and HRQoL.

References

1. Lipscomb J, Gotay CC, Snyder CF Patient-reported outcomes in cancer: a review of recent research and policy initiatives. CA Cancer J Clin 2007; 57: 278–300.
2. Black N. Patient reported outcome measures could help transform healthcare. BMJ 2013 (Clinical research ed); 346: f167.
3. Institute of Medicine. Crossing the Quality Chasm: A new health system for the 21st Century. Washington, DC: National Academy Press, 2001.
4. Darzi A. High Quality care for all: NHS Next Stage Review final report. London: Department of Health, 2008.
5. Department of Health. Cancer Reform Strategy, 2007. webarchive.nationalarchives.gov.uk/+/www.dh.../Cancer/DH_091120
6. Department of Health. Improving Outcomes: A Strategy for Cancer (January 2011). https://www.gov.uk/government/uploads/system/.../dh_123394.pdf
7. Department of Health. Cancer Patient Experience Survey 2011/12. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/212860/Cancer-Patient-Experience-Survey-National-Report-2011-12 pdf.
8. Glaser A, Fraser L, Corner J, et al. Patient-reported outcomes of cancer survivors in England 1-5 years after diagnosis: a cross-sectional study. BMJ Open 2013; 3:e002317.
9. Downing A, Morris E, Richards M, et al. Health-related quality of survival after cancer in England: a patient-reported outcomes study of 21,000 individuals diagnosed with colorectal cancer. J Clin Oncol 2014; *In Press*.
10. O'Cathain A, Thomas KJ. "Any other comments?" Open questions on questionnaires – a bane or a bonus to research? BMC Med Res Methodol 2004; 4: 25.
11. Corner J, Wagland R, Glaser A, et al. Qualitative analysis of patients' feedback from a PROMs survey of cancer patients in England. BMJ Open 2013; 3:e002316.
12. Howell E, Martin J. Inpatient Survey 2007: Analysis of the patients' free-text comments. Picker Institute Europe http://www.nhssurveys.org/survey/736 (accessed 020114).
13. Hearst M (2003) What Is Text Mining? http://people.ischool.berkeley.edu/~hearst/text-mining.html
14. Li N, Wu DD. Using text mining and sentiment analysis for online forums hotspot detection and forecast. Decision Supp Systems 2010; 48: 354–368.
15. Cheung KW, Kwok JT, Law MH, et al. Mining customer product ratings for personalized marketing. Decision Supp Systems 2003; 35(2): 231-243.
16. Pang B, Lee L (2008). Opinion mining and sentiment analysis found. Trends Inf Retr 2008; 2 (1-2):1-138.
17. Alpers GW, Winzelberg AJ, Classen C, et al. Evaluation of computerized text analysis in an Internet breast cancer support group. Computers in Human Behavior 2005; 21: 361–376.
18. Greaves F, Ramirez-Cano D, Millett C, et al. Use of sentiment analysis for capturing patient experience from free-text comments posted online. J Med Internet Res 2013; 15(11): e239.
19. Seale C, Ziebland S, Charteris-Black J. Gender, cancer experience and internet use: a comparative keyword analysis of interviews and online cancer support groups. Soc Sci Med 2006; 62: 2577–2590.

20. Seale S (2006) Gender accommodation in online cancer support groups. Health 2006; 10(3): 345–360.
21. Moore AP, Milligan P, Rivas C, Goff LM. Sources of weaning advice, comparisons between formal and informal advice, and associations with weaning timing in a survey of UK first-time mothers. Public Health Nutr: 2012; 15(9): 1661–1669.
22. Department of Health. The power of information: Putting all of us in control of the health and social care information we need. London: Department of Health, 2012.
23. Arndt V, Merx H, Stegmaier C, et al. Quality of life in patients with colorectal cancer 1 year after diagnosis compared with the general population: a population based study. J Clin Oncol 2004; 22: 4777-4784.
24. Ness RM, Holmes A, Klein R, et al. Outcome states of colorectal cancer: Identification and description using patient focus groups. American J Gastroenterol 1998; 93:1491-1497.
25. Sanchez KM, Richardson JL, Mason HRC. The return to work experiences of colorectal cancer survivors. AAOHN Journal 2004; 52: 500-510.
26. Gordon LG, Lynch BM, Beesley VL et al. The Working After Cancer Study (WACS): a population-based study of middle-aged workers diagnosed with colorectal cancer and their return to work experiences. BMC Public Health 2011; 11(1):604. http://www.biomedcentral.com/1471-2458/11/604.
27. Northouse LL, Schafer JA, Tipton J et al. The concerns of patients and spouses after the diagnosis of colon cancer: a qualitative analysis. J Wound Ostomy Continence Nurs 1999; 26: 8-17.
28. Brown H, Randle J. Living with a stoma: a review of the literature. J Clin Nurs 2005; 14: 74-81.
29. Jenks JM, Morian KH, Tomaselli N. The influence of ostomy surgery on body image in patients with cancer. Applied Nursing Research 1997; 10: 174-180.
30. Persson E, Hellstrom A. Experiences of Swedish men and women 6 to 12 weeks after ostomy surgery. J Wound Ostomy Continence Nurs 2002; 29: 103-108.
31. EuroQol Group. EuroQol - a new facility for the measurement of health-related quality of life. Health Policy 1990; 16:199-208.
32. Wright E, Kiely M, Johnston C et al. Development and evaluation of an instrument to assess social difficulties in routine oncology practice. Qual Life Res 2005;14(373): 386.
33. Webster K, Cella D, Yost K. The functional assessment of chronic illness therapy (FACIT) measurement system: properties, applications and interpretation. Health Qual Life Outcomes 2003; 1:79
34. Jurka TP, Yoshimasa T. maxent: Low-memory Multinomial Logistic Regression with Support for Text Classification. URL http://CRAN. R-Project. org/package= maxent, R package version 2012; 1(2).
35. R Core Team. The R project for statistical computing. Revised on October 25 (2011). www.r-project.org/
36. Gibbs GR. Qualitative data analysis: explorations with NVivo. Buckingham: Open University, 2012.
37. Collingwood L, Wilkerson J. Trade-offs in Accuracy and Efficiency in Supervised Learning Methods. Journal of Information Technology & Politics 2012; 9(3); 298-318.
38. Timothy P. Jurka, Loren Collingwood, Amber E. Boydstun, Emiliano Grossman and Wouter van Atteveldt (2015). RTextTools: Automatic Text Classification via

Supervised Learning. R package version 1.4.2. http://cran.r-project.org/web/packages/RTextTools/RTextTools.pdf
39. Sokolova M, Japkowicz N, Szpakowicz S. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In AI 2006: Advances in Artificial Intelligence, pp1015-1021. Springer Berlin Heidelberg, 2006.
40. Sasaki Y. The truth of the F-measure. Teach Tutor mater, 1-5, 2007.
41. Silverman D. Interpreting Qualitative Data: Methods for Analysing Talk, Text and Interaction. London: SAGE Publications, 1993.
42. Seale C. The Quality of Qualitative Research. London: SAGE Publications, 1999..
43. Bryman A, Burgess RG. Analyzing qualitative data. London: Routledge, 1994.

Figure 1: Three phases of text-mining assisted qualitative analysis

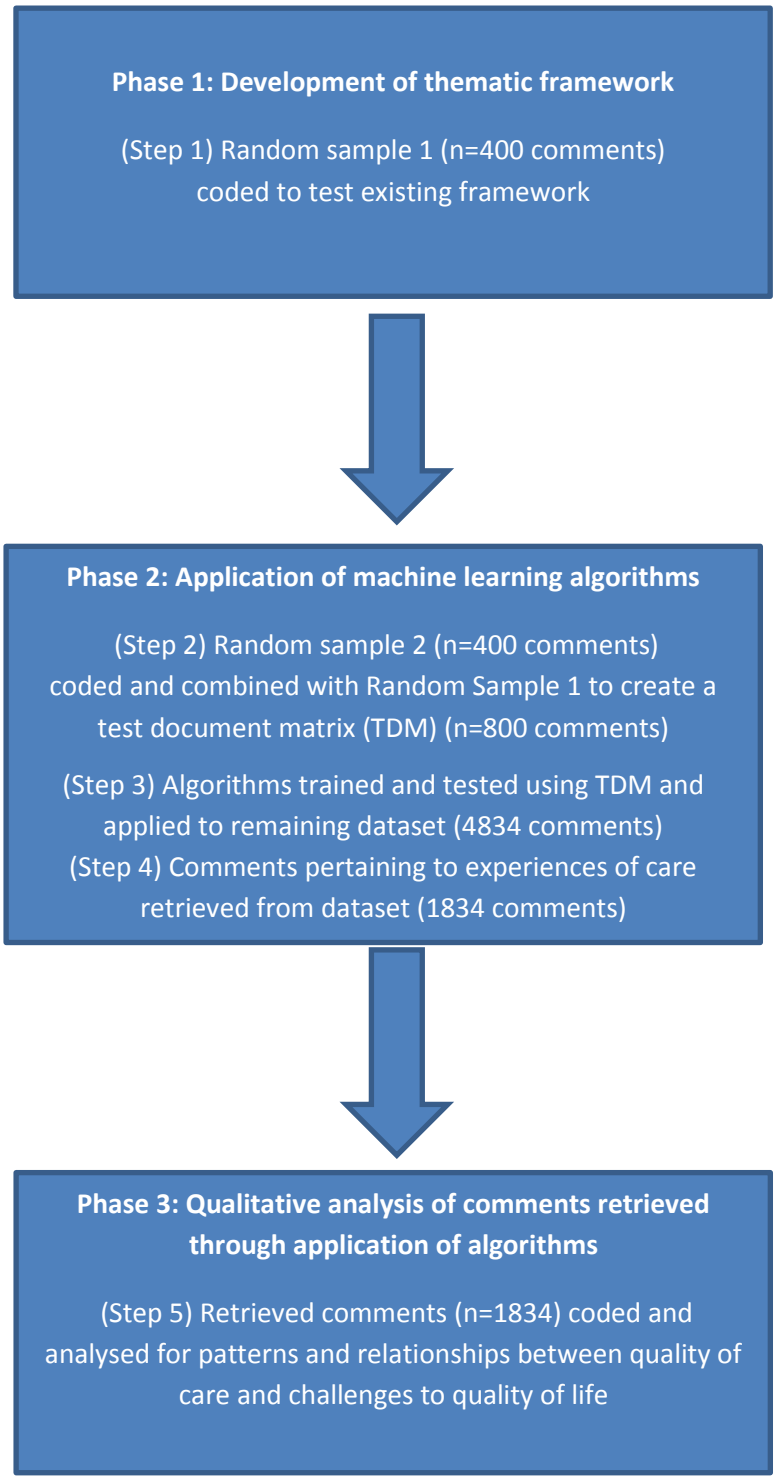

Figure 2: Model of factors influencing quality of patient experience

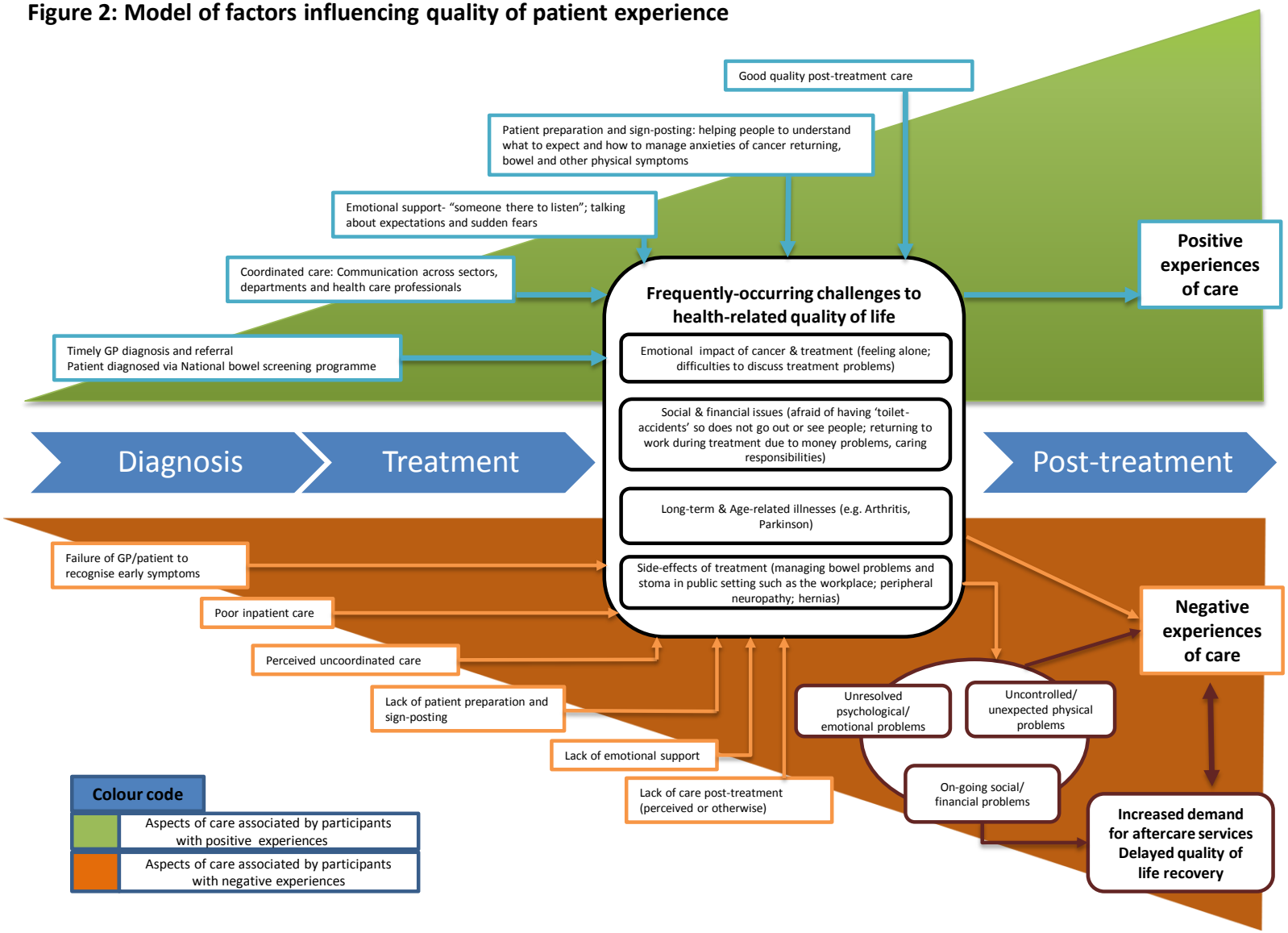

Table 1: Socio-demographic and disease related characteristics of free-text respondents

Characteristic		No. received survey (n=34,467)	No. completed survey (n=21,802)	Survey response rate	No. provided comments (n=5,634)	Free-text question response rate	Comparison of those who did and did not provide comments
Sex	Male	19,580	12,683	64.8%	3,180	25.1%	
	Female	14,887	9,119	61.3%	2,454	26.9%	p=.002[1]
Age at Diagnosis	<55	3,645	2,040	56.0%	582	28.5%	
	55-64	7,611	5154	67.7	1,418	27.5%	
	65-74	11,559	7824	67.7	1,975	25.2%	
	75-84	9,302	5,633	60.6	1,368	24.3%	
	85+	2,359	1,151	49.0	291	25.3%	p<.001[3]
Ethnic group	White	24,781	16,079	64.9	4,257	26.5%	
	Mixed	75	40	53.3	9	22.5%	
	Asian	407	171	42.0	40	23.4%	
	Black	332	143	43.1	24	16.8%	
	Other	255	124	48.6	28	22.6%	
	Unknown	8,617	5,245	60.9	1,276	24.3%	p=.001[1,2]
Index of Multiple Deprivation	1 – least deprived	7,812	5,484	70.2	1,595	29.1%	
	2	8,090	5,360	66.3	1,460	27.2%	
	3	7,391	4,742	64.2	1,197	25.2%	
	4	6,267	3,658	58.4	860	23.5%	
	5 – most deprived	4,907	2,558	52.1	522	20.4%	p<.001[3]
Year of diagnosis	2010	16,491	10,523	63.8	2,793	26.5%	
	2011	17,976	11,279	62.7	2,841	25.2%	p=.023[1]
Tumour site	Colon	21,696	13,577	62.6	3,544	26.1%	
	Rectosigmoid	2,319	1,512	65.2	389	25.7%	
	Rectum	10,452	6,713	64.2	1,701	25.3%	p=.502[1]
Duke's stage of disease at diagnosis	A	3,536	2,435	68.9	602	24.7%	
	B	7,710	5,149	66.8	1,335	25.9%	
	C	6,705	4,482	66.8	1,178	26.3%	
	D	1,766	1,045	59.2	305	29.2%	
	Unknown	14,750	8,691	58.9	8,691	25.5%	p=.014[3]

[1] p-value corresponds to chi-square test

[2] chi-square test compares white versus non-white versus unknown

[3] p-value corresponds to chi-square test for trend

Table 2: Manually coded positive and negative comments, estimated and actual algorithm-retrieved comments and mean EQ5D scores

Category	Sub-category	Manually coded comments (n=800)		Estimated algorithm-retrieved comments from remaining dataset (n=4834)		Actual algorithm-retrieved comments from remaining dataset (n=4834)		Actual MLA retrieved comments (n=1688)		Mean EQ5D score	Mean difference between all comments and category/sub-category (EQ5D score)
		(n=)	%	(n=)	%	(n=)	%	(n=)	%	1.0 or less	t (p=, CI)
All comments		**800**	**100**	**4834**	**100**	**4834**	**100**	**1688**	**100**	**0.78**	
Positive and negative comments combined		**333**	**37**	**1714**	**35.4**	**1688**	**34.9**	**1688**	**100**		
Positive experiences		248	31	1436	29.7	1045	21.6	1045	61.9	0.85	0.07 (p=<.001 95% CI = 0.05 - 0.09)
	Timely diagnosis							109	6.4	0.90	0.12 (p=<.001 95% CI = 0.09 - 0.16)
	Good quality post-treatment care							289	17.1	0.88	0.09 (p=<.001 95% CI = 0.08 - 0.12)
Negative experiences		85	10.6	491	10.1	643	13.3	643	38.0	0.74	0.05 (p=<.001 95% CI = 0.02 - 0.07)
	Delayed diagnosis							36	2.1	0.73	0.05 (p=0.298 95% CI = 0.05 - 0.15)
	Inadequate post-treatment care							191	11.3	0.71	0.07 (P=.002 95% CI = 0.02 – 0.11)
	Poor in-patient care							45	2.6	0.74	0.03 (p=.423 95% CI = 0.05 – 0.12)
	Lack of coordinated care							78	4.6	0.76	0.02 (p=.536 95% CI = 0.04 – 0.08)
	Lack of emotional support							78	4.6	0.87	0.04 (p=.100 95% CI = 0.01 – 0.10)
	Lack of information on treatment side-effects							160	9.4	0.74	0.04 (p=.034 95% CI = 0.00 – 0.08)
	Lack of information concerning possible psychological impact of cancer and treatments							43	2.5	0.69	0.09 (P=.073 95% CI = 0.01 – 0.19)
	Lack of information on self-management strategies							135	7.9	0.79	0.01 (P=0.581 95% CI = 0.03 – 0.06)
	Lack of GP involvement							52	3.0	0.73	0.05 (P=.167 95% CI = 0.02 - 0.13)

Note: Individual participants often provided free-texts comments that were coded into more than one category.

Table 3: Algorithm performance identifying comments relating to positive and negative experience: 50:50 split between training and testing samples approach and 10-fold cross validation

Algorithm	Approach 1: 50% comments used for training and 50% for testing			Approach 2: 10 fold cross-validation	
	Precision	Sensitivity	F-Score	Mean performance score across 10 folds	SD
Support Vector machines (SVM)	0.835	0.780	0.800	0.834	0.027
Random Forests	0.825	0.765	0.780	0.839	0.028
Decisions Trees	0.735	0.710	0.720	0.770	0.050
(GLMNET)	0.750	0.700	0.710	0.523	0.084
Bagging	0.725	0.700	0.710	0.811	0.039
Maxentropy	0.670	0.670	0.670	0.014	0.009
Logitboost	0.710	0.655	0.655	0.876	0.037

Note: GLMNET and Bagging have the same F-score, but precision was higher for GLMNET