EasyChair Preprint

№ 1103

Application of Natural Language Processing to Determine User Satisfaction in Public Services

Radoslaw Kowalski, Marc Esteve and Slava Jankin Mikhaylov

June 6, 2019

Improving Public Services by Mining Citizen Feedback: An Application of Natural Language Processing

Radoslaw Kowalski	Marc Esteve	Slava J. Mikhaylov
University College London	University College London and ESADE	Hertie School of Governance
radoslaw.kowalski.14@ucl.ac.uk	marc.esteve@ucl.ac.uk	jankin@hertie-school.org

Abstract

Research on user satisfaction has increased substantially in recent years. Studies to date tend to test for significance of pre-defined factors thought to have an influence with no scalable means to verify the validity of the assumptions made. Digital technology has enabled new methods to collect user feedback, for example through online forums where service users post comments. Topic models can help analyze large volumes of such feedback and are proposed as a feasible solution to aggregate user opinions for use in the public sector. Insights can contribute to a more inclusive decision-making process in public services. This novel approach is applied to process reviews of publicly-funded primary care practices in England. Findings from the analysis of over 200,000 reviews indicate that the quality of interactions with staff and bureaucratic exigencies are the key drivers of user satisfaction. Moreover, patient satisfaction is strongly influenced by factors not considered in state-of-the-art patient surveys. These results highlight the potential benefits that text mining and machine learning for the public administration field.

Keywords:

natural language processing, citizen feedback, public service performance, citizen satisfaction, big data

Introduction

Democratic governance is possible and effective when citizens' opinions are included in public decisions (Fung 2015; Feldman 2014; Mahmoud and Hinson 2012). At the same time, citizens' opinions are hard to capture. They tend to have little to do with the formal measures of organizational performance used within organizations (Harding 2012; Ma 2017; Moynihan, Herd, and Harvey 2014; Sanders and Canel 2015) or the opinions of organizational managers (Andersen and Hjortskov 2016; Sanders and Canel 2015). Existing research on citizen satisfaction shows that this is determined by several factors, such as how they use public services (Brown 2007; Im et al. 2012; Ladhari and Rigaux-bricmont 2013; Pierre and Røiseland 2016; Van Ryzin and Charbonneau 2010), in what way they are involved with their provision (Sanders and Canel 2015; Scott and Vitartas 2008; Taylor 2015) as well as according to their held-out knowledge, beliefs (Barrows et al. 2016; Brown 2007; Harding 2012; Ladhari and Rigaux-bricmont 2013) and emotions (Lawton and Macaulay 2013; Ma 2017). Continuous analysis of those preferences can help ensure that managers of public institutions make decisions aligned with the public need (Walker and Boyne 2009).

Digital technologies have led to the creation of a host of new opportunities for the collection of citizen feedback (Kong and Song 2016). On the one hand, these new data resources can be very insightful because they contain full citizen opinions about public services compared to traditional survey methods that probe select range of issues. User comments are widely utilized for this reason in private sector organizations (Qi et al. 2016), so far with scant examples within the public sector (Hogenboom et al. 2016). There are also problems with using these new data resources. First, they can be too large to read and analyze manually (Kong and Song 2016). Second, the obtainable data may predominantly consist of unstructured text, which is hard to summarize with statistical techniques (Kong and Song 2016). Finally, it can be difficult to

pinpoint the sample biases because authors' identities are uncertain (Yang 2010). The volume and structure of text feedback, e.g. in the form of reviews, makes it difficult to understand the causes of user satisfaction from public services. Simultaneously, existing tools developed for private organizations may not be adequate for use in the public sector. Public organizations require insights into service user preferences in situations where citizens are "forced customers" (Di Pietro, Mugion, and Renzi 2013) and where public organizations must fulfill objectives unrelated to service demand or profitability (Brownson et al. 2012).

This study addresses the shortcoming in quantifying user satisfaction expressed in unstructured text feedback. Unstructured and anonymous opinions can help provide a substantial answer to the research question: "What are the determinants of user satisfaction in public services?" Large quantities of reviews can be summarized with natural language processing (NLP) models, such as topic models in order to obtain actionable insights (Blei, Ng, and Jordan 2003; Hogenboom et al. 2016). Insights from topic modeling can be compared against other analyses such as surveys to systematically evaluate the validity and reliability of text-derived insights. The article offers two contributions to the public management field: 1) it evaluates a comprehensive model of determinants of citizen satisfaction constructed from a large corpus of written feedback, and 2) offers a method to analyze big data to allow inclusion of citizen voice in reforms of public services. The contributions stem from the implementation of NLP to solve a public management analytical problem.

User Satisfaction for Inclusive Public Policy

The inclusion of the service user voice in decisions about public services requires a robust understanding of whether, how and why they are satisfied. It is then possible to take citizen preferences into account when making political or public policy decisions. As noted above,

citizen satisfaction is known to correlate (but often non-linearly) with a number of factors including socio-economic status, education, and employment history (Christensen and Laegreid 2005; Harding 2012; Jlike, Meuleman, and Van de Walle 2014; Yang 2010), demographic background (Yang 2010), and available knowledge (Hong 2015; Im et al. 2012; James and Moseley 2014; Lavertu 2014; Villegas 2017). While researchers have uncovered multiple possible of user satisfaction from public services, it often remains unclear how those determinants relate to one another in a specific context, and whether the interactions between determinants are the same irrespective of context and the passage of time (Song and Meier 2018). Moreover, it is often unclear whether the aspects of user satisfaction of interest to researchers and/or commissioners of research constitute a complete list of issues (Lavertu 2014; Roberts et al. 2014). Factors outside the scope of already well-known determinants of satisfaction may bias insights from commissioned studies in unpredictable ways. The avenues of how and why it happens are often entirely unclear (Pierre and Røiseland 2016).

Similarly, researchers can choose from a wide range of theories when designing their opinion research, which makes it difficult to construct a robust, holistic understanding of what matters most to users of public services across studies. For example, analysts may emphasize the impacts of available information (James and Moseley 2014; Marvel 2016), self-centered utility maximization (Jensen and Andersen 2015), emotions (Ladhari and Rigaux-bricmont 2013), sense of identity (Jlike, Meuleman, and Van de Walle 2014), unconscious tendency towards conformity (Sanders and Canel 2015), or the level of physical involvement with the services under review (Loeffler 2016). In the end it can be uncertain how does subconscious identification as a member of a group (Sanders and Canel 2015) intertwine with, for instance, self-interest (Jensen and Andersen 2015) to lead to a specific set of reasons as to why a given service user (dis)likes a specific public service. Similarly, it is not certain why achievements

in improving official performance measures are often incongruent with citizens' satisfaction levels (Brenninkmeijer 2016). Narratives used by citizens to explain their (dis)satisfaction may be unknown even when behavior is well-understood (Müssener et al. 2016). Politicians and policymakers may struggle to include citizen perspective in decisions even when studies of user opinion are abundantly available.

The available literature indicates that there is a gap in understanding the relative importance and relationships between the determinants of service user satisfaction, combined with an absence of means to assess whether some factors influencing user satisfaction are omitted in citizen satisfaction evaluations. Written comments of citizens about public services are a big data resource that can help address some of the gaps in understanding of user satisfaction and make insights more useful for guiding policy-making. Citizen comments contain holistic insight into reasons for citizens' satisfaction and can help establish the importance on all issues relative to one another. Machine learning can be a useful tool to effectively summarize text comments and retrieve relevant insights.

User Feedback as a Measure of Satisfaction

Consideration of public opinion is a prerequisite of successful democratic governance (Feldman 2014) and is necessary to solve problems regarding service output performance (Fung 2015; Mahmoud and Hinson 2012). Physical participation of citizens in public decision-making is one way for authorities to engage and understand the service user perception of public services (Fung 2015). The approach can help bring change to institutions and increase public satisfaction from public services (Moon 2015). At the same time, direct public participation in decisions is not always easy to implement in complex policy areas. In an applied context, it may also politicize otherwise quick administrative decisions with poor

marginal returns for the additional effort put into decision-making (Bartenberger and Sześciło 2016). Moreover, in many institutional contexts it is difficult to capture enough interest from service users to keep them regularly involved in decision-making (Fung 2015; Greer et al. 2014). Liu (2016) argues, with hands-on examples, that the understanding of service user preferences could improve with information technologies and lead to new modes of decision-making.

The representation of the service user voice through data collection and summarization can replace direct citizen participation in situations where the latter is not feasible. Experiments or qualitative researches are one way to study public opinion (e.g. James and Moseley 2014; Mahmoud and Hinson 2012). Those research methods, however, tend to be one-off with the aim of understanding specific problems with public services. The high running costs involved may be among the reasons why reviewed studies did not mention the use of experiments or qualitative research approaches for the day-to-day inclusion of the public's voice in decisions about public services. Surveys are a widely used alternative way to measure user satisfaction and assess service providers (Van de Walle and Van Ryzin 2011; Olsen 2015) but they are also a method with its own problems. There are no systematic tools to adapt survey's structure or scope to changing conditions (Burton 2012). Furthermore, inability to carry out frequent surveys also makes them unsuitable for daily monitoring of opinions to observe in real-time organizational change (Burton 2012; Walker and Boyne 2009). Feedback received through restricted lists of survey questions tends to also oversimplify the reasons for user satisfaction (Amirkhanyan, Kim, and Lambright 2013; Jlike, Meuleman, and Van de Walle 2014), may be biased by survey structure (Van de Walle and Van Ryzin 2011) and the final survey outputs may blur distinctions between similarly scoring service providers (Voutilainen et al. 2015). Therefore, both practitioners and academics encourage the introduction of other forms of data

to gauge the determinants of user satisfaction regarding public services more effectively (Amirkhanyan, Kim, and Lambright 2013; Andersen, Heinesen, and Pedersen 2016; Brenninkmeijer 2016; Lavertu 2014).

Alternative forms of user satisfaction measurement should be able to map dynamic changes in what organizational performance means across contexts and over time. Data insights should also holistically capture and represent what is meant by service users and other relevant individuals such as political decision-makers and public servants. Conceptualization of public service performance as a ever-changing and differently defined from person to person can help avoid the reproduction of deficiencies in evidence-based policy making. Those deficiencies include the suppression of the less powerful voice of service users within the performance measurement process (Mergel, Rethemeyer, and Isett 2016) and the measurement of user satisfaction with methods that quickly lose their relevance, requiring effort to develop a replacement (Gao 2015). Data resources that start to be available have the potential to help improve public services by enabling dynamic monitoring of performance (Rogge, Agasisti, and De Witte 2017). For example, network signals and written feedback have already proved their usefulness in service improvements such as e-government, traffic control, and crime detection (Rogge, Agasisti, and De Witte 2017). At the same time, the new technological possibilities require further effort in order to utilize new data within the public policy domain. The sheer volume of data may be challenging to handle (Grimmer and Stewart 2013) and decision-makers may not be fully able to collect, process, visualize, and interpret them (Brenninkmeijer 2016; Lavertu 2014; Rogge, Agasisti, and De Witte 2017). Furthermore, public policy researchers highlight the ethical issues inherent in handling personal data, including respect for individual privacy and security as well as concerns around the quality of democratic processes (Mergel, Rethemeyer, and Isett 2016). The tools developed to handle complex data from service users

should be designed with the intention to address those concerns while offering added value for delivery of public services.

Written reviews of public services are one data resource that captures the voice of service users and include it in public decision-making. Online written reviews can help to address privacy issues since they can be posted anonymously. At the same time, they may still be a valid resource for decision-makers within public institutions, despite complex sample biases (Grimmer and Stewart 2013). This is because they can be validated against state-of-the-art structured forms of user feedback, such as carefully drafted surveys with large numbers of reviewers (Grimmer and Stewart 2013; Rogge, Agasisti, and De Witte 2017). Furthermore, the requirements of basic literacy in any language combined with access to the internet can make online forums a channel wherein almost every public service user could contribute and inform research and practice. The ease of use of online forums results in written reviews being a potential means for ensuring the equitable distribution of services (Kroll 2017), and for addressing concerns about democratic deficit in public decision-making (Mergel, Rethemeyer, and Isett 2016). Moreover, organizations assessed based on user reviews may be relatively less able to manipulate performance scores, a common problem with evaluations of performance in public institutions at present (Hood and Dixon 2015, 265–267). In addition, the likelihood of decision-makers making poor decisions due to over-reliance on very narrow understandings of service quality is reduced (Luciana 2013). Thus, online reviews could be helpful in understanding and including citizen feedback in decisions about how to provide public services.

Data

In the present article, the evaluation of the link between satisfaction surveys and unstructured reviews is carried out on a dataset of online reviews about publicly funded primary care (GP) services in England. Reviews were downloaded in .xml format from a web service of National Health Service (NHS)[1] and transformed into a .csv table format used for modeling with the R programming language. The reviews were posted from July 2013 to January 2017, concerning almost 7,700 GP practices. 208,287 reviews were fully filled out and included in this study (about 89% of all reviews). The reviews were 5-6 sentences long on average, with a median length of five sentences.

The reviews corpus was pre-processed, following the standard practice (Grimmer and Stewart 2013). We lowercased and stemmed the tokens (words); and removed numbers, punctuation, stop words, tokens shorter than three characters, and tokens that appeared fewer than 10 times and more than 100,000 times in the corpus. Pre-processing removed 46,277 terms that occurred 89,374 times in GP reviews. The final corpus contained 9,148 terms that occurred over 8.5 million times in the dataset.

Each month, anonymous users posted between 3,000 and 5,000 written comments accompanied by 5-point Likert-scale star ratings of six aspects of their GP service experience. The Likert-scale star ratings related to survey statements: 1) "Are you able to get through to the surgery by telephone?", 2) "Are you able to get an appointment when you want one?", 3) "Do the staff treat you with dignity and respect?", 4) "Does the surgery involve you in decisions about your care and treatment?", 5) "How likely are you to recommend this GP surgery to

[1] More about NHS Choices, the NHS organization responsible for feedback data management: http://www.nhs.uk/aboutNHSChoices/aboutnhschoices/Pages/what-we-do.aspx, viewed on 17 September 2017

friends and family if they needed similar care or treatment?", and 6) "This GP practice provides accurate and up to date information on services and opening hours". Variability in author comments and star ratings did not occur as a result of variance in how authors interpreted the questions becaus20e the formatting of the Likert-scale questions was stable across the period when comments were posted.

It should be noted that there are no available socio-demographic attributes for users posting the data, so the sample could be skewed towards certain demographics. Anyone can comment on the website and evaluate GP practices. Qualitative reading of the comments reveals that most comments are posted by patients or patients' carers, relatives and friends, especially in situations when a significant positive or negative experience has moved them emotionally. Lack of access to internet or computer skills among patients does not prevent some groups of patients from sharing opinions but it is less likely. Apart from that, administrators at NHS Choices manually remove malicious or otherwise inappropriate messages from the server and ensure that unfavorable but legitimate reviews remain consistently in the dataset across England[2].

Topic Modeling

A key challenge in using written reviews for inclusive public policy is how to process them in a way that is scalable and meaningful for public decision-makers. Fortunately, machine learning models, such as topic models, are already well known to simplify insights from written reviews into easy-to-understand summaries in near real-time, regardless of their quantity (Blei, Ng, and Jordan 2003; Griffiths and Steyvers 2004). An advantage of these over user surveys is that they can automatically adapt to changes in what citizens write (Blei and Lafferty 2006;

[2] See for further details http://www.nhs.uk/aboutNHSChoices/aboutnhschoices/termsandconditions/Pages/commentspolicy.aspx

Dai and Storkey 2015) without prior assumptions or constraints about which service aspects reviewers can express their satisfaction (Blei, Ng, and Jordan 2003). This is especially useful when manual labelling of written documents is not feasible due to their high volume, or when new documents are continually added to the dataset and require processing. Several studies have attempted an analysis of written user feedback from services using machine learning algorithms for organizational improvement (Gray 2015; Rogge, Agasisti, and De Witte 2017). However, none has established a firm relationship as to how key themes identified in online written reviews with topic modeling relate to the established measures of user satisfaction, such as satisfaction surveys. The knowledge gap must be filled before online written reviews can be used reliably as a measure of user satisfaction that supports the provision of public services (Grimmer and Stewart 2013; Rogge, Agasisti, and De Witte 2017). Furthermore, the relationship between survey outcomes and the content of written reviews can help researchers understand how reviewer narratives relate to dimensions of satisfaction with public services included in the survey.

Written user comments were analyzed using structural topic modeling (STM) implemented with the *stm* software package for R programming language[3]. A set of key topics from the database of written documents is identified and proportional presence of each topic in each document is estimated (Blei 2012). Topics derived from reviews in this study may be about thanking doctors, complaining about reception staff, or commenting about the quality of GP facilities.

Proportions of topics in comments are calculated based on how each word included in each comment is likely to belong to each topic. We follow topic model description from Blei, Ng,

[3] Further details about the *stm* software library used in the R programming language for model implementation is available at: https://CRAN.R-project.org/package=stm, viewed on 17 September 2017

and Jordan (2003). The probabilities of each word belonging to each topic are estimated during model training. The algorithm begins model training with a random allocation of topics to every document in the corpus, in a form of a probability distribution. Values for all topics in a comment are probabilities between 0 to 1 of them occurring in the document, Topic probabilities given document sum to 1. Next, for each word in every document, the algorithm picks a topic from the probability distribution of topics assigned to the document. After passing through all documents, each word has some probability of belonging to each topic, a likelihood of a topic being chosen given a word. Then, the algorithm attempts to reproduce original text documents by picking random words from topics according to the topic-word probability distributions and given the probability of each topic in each document. The mismatch in picked words and the word content of the original documents constitutes the loss of the model which is minimized iteratively during model training. Structural topic model (STM) is used in the analysis here (Roberts et al. 2014).

The model requires a human analyst to pick the number of topics to be uncovered within the dataset. We follow Roberts et al. (2015) and select the optimal number of topics as a balance between exclusivity and semantic coherence from models which range from 3 to 100 topics. Our analysis shows that 20 topics is the optimal setting for our objective to evaluate how text comments can be analyzed with machine learning for use in public policy research. Models with fewer than 20 topics suffered from lower exclusivity of topics, which means topics are less likely to represent distinct meanings. Models with more than 20 topics, on the other hand, did not improve in terms of semantic coherence or exclusivity of topics over the 20-topic model while containing more complex insights. Greater complexity of the topic model was not necessary for answering the research question. Appendix A in supplementary materials discusses the selection process in more detail. The 20 topics from the selected model are listed

in Table 1. Appendix B provides details on the topic labeling exercise, and additional information on the topics' content and frequency in our data.

[TABLE 1]

A map of topic correlations (Figure 1) is a convenient way to summarize topic modeling results[4]. It allows to make comparisons between topics that have been calculated based on the similarity of words between pairs of topics. The greater the distance and the thinner the connecting line between two topics, the less they tend to occur together within reviews. Clusters of related topics are represented by node colors. In this case, red topics represent negative experiences, green topics cluster positive experiences, and orange topics group themes without a strong positive or negative sentiment. Topic clusters have been calculated with a sentiment analysis model trained to predict star rating (further details are in Appendix D). Furthermore, node size for topics corresponds to their popularity across patient reviews. Larger nodes stand for more common topics.

Figure 1: Topic map for the 20-topic STM model

[FIGURE 1]

Notes: (1) Topic map illustrates, on a 2-dimensional plane, how similar 20 topics generated with the STM topic model from NHS GP practice reviews are to one another. Distances between topics are proportional to the differences of the words they contain. The most similar topics in terms of the words they contain tend to be close to one another. (2) Nodes represent individual topics. The bigger the node, the more prevalent the given topic within the dataset. (3) The stronger the line connecting a pair of topics, the greater the similarity between the two topics. (4) Node colors indicate clusters to which topics have been assigned. The green cluster contains topics (marked with "∧") related to positive evaluation of GP service quality. The red cluster groups negative evaluations of GP service quality (marked with

[4] Topic map has been generated with Gephi, a software package for network modelling. For further information about Gephi, please visit: http://gephi.org, viewed on 17 September 2017

"∨"). The orange cluster groups themes (marked with "=") related which tend to be more neutral. Labels have been assigned with a sentiment analysis model.

Figure 1 maps positive topics on the left side of the map. They are most different from topics containing negative GP service evaluations at the top-right of the map. The second greatest difference is between topics that cluster words used to express personal thoughts and feelings (top of the map) and topics that contain words used in third person narratives or passive voice (bottom of the map). The most common topics include expressions of gratitude and complaints about the difficulty/impossibility of accessing the services.

Explaining User Satisfaction with Feedback

As discussed above, the GP reviews in our dataset also come with Likert-scale survey responses, a common and accepted measure of user satisfaction (Hartley and Betts 2010). We use them here as a well-established template measure to relate to our metrics of user satisfaction from topic modeling. First, we estimate Random Forest (RF) models where the proportional presence of topic reviews are independent variables and six Likert-scale ratings are treated as dependent variables.

RF is a machine learning algorithm that builds decision trees on randomly sub-sampled data with a smaller subset of randomly sampled predictors. A large number of trees is grown, and then all trees are combined and averaged for use as a trained RF model. RF takes advantage of both weak and strong predictor variables, where weak ones are those that make predictions only slightly better than a random guess of an outcome. The model is easier to interpret than other popular machine learning algorithms, and can capture non-linear relationships between predictors and predicted variables. One benefit of using RF models here is that by design they deal with multicollinearity between predictor variables and thus allow for an unambiguous

identification of the importance of topics identified with the STM analysis in predicting Likert-scale ratings. For further details on RF models see, for example, Hastie, Tibshirani and Friedman (2001, 587–603).

Our multiclass RF model predicts the outcome variables with accuracy ranging from 0.48 on "phone access ease" to 0.77 on "likely to recommend" dimensions. Precision and recall measures vary across "star" levels and dimensions, with the F1-score ranging between close to zero to 0.85. This variation is partly driven by difference in sample sizes across different models (as can be seen from the confusion matrices in Appendix E). Overall, we are capturing some of the relationship between unstructured data (reviews summarized with topic models) and structured data (Likert-scale "star" ratings).

Figure 2 presents the results of the RF model in terms of the importance ranking of independent variables for predicting each individual Likert-scale outcome variable[5]. RF outcomes indicate that topics generated from online reviews are related to Likert-scale responses provided by service users. Furthermore, satisfaction from multiple aspects of the GP service is related to similar themes present in the reviews. It suggests that user satisfaction can be improved among multiple dimensions by adopting a single approach of addressing important, common problems and enhancing the key positive experiences.

The topics from 20-topic STM model were labelled according to the most common words present in each of them. Topics indicating positive experiences were the strongest predictor of satisfaction with topic 7 ("recommend") as the most important, followed by similar topics 9 "thanks" and 8 "helpful". The most common words in topic 7 include: thank, recommend, support and kind (see Appendix B for details). The topics' contents indicate that caring staff behavior towards patients has the highest influence on how positively patients evaluate GP

[5] We show only the top 10 most important predictors to simplify the presentation in the plots.

services. Similarly, the opposite approach - rejection of patients - is the most significant drag on patient evaluations of their experience. Topics 14 “discourage registration” and 18 “no appointments” group opinions expressing disappointment lack of access to the services because of disrespectful treatment of patients (topic 14) or possibly demand outstripping supply of services (topic 18). The top topics show that patients seek treatment from caring professionals. More neutral experiences represented by topics such as “proper treatment”, “diagnosed and sorted” and “unwelcoming” tend to be good predictors of a neutral sentiment (Figure 1). They have a weaker impact on Likert-scale ratings. Among negative experiences, the quality of medical care is less of an issue to patients than non-medical issues. Procedural problems with making an appointment are a strong negative impact on evaluations of GP services. Patients finding it hard to use telephone, online and on-site booking of appointments suggest that the NHS is not a fully efficient organization. Procedural problems increase the cost of providing GP services, especially administrative costs, without adding any value. They also worsen the atmosphere in GP practices, which is suggested by topic 20 “rude reception” appearing consistently among the top 10 topics predicting star ratings (Figure 2).

Figure 2: Random forest model results - importance ranking for topics on six dimensions of GP service quality

[FIGURE 2]

Notes: (1) Random Forest model outcomes illustrate with horizontal bars the importance of topics (independent variables) for correct prediction of star ratings (dependent variables) given in response to the six Likert-scale survey statements. Star ratings are treated as categorical data. (2) Topic importance represents the average improvement in classification at the moment when a topic is used in the Random Forest model as an independent variable. Model improvement is measured with residual sum of squares. (3) Each sub-figure includes the most important 10 topics for predicting the dependent variable. The omitted 10 topics had scores similar to the included least important topics.

Overall, our analysis suggests that access to healthcare services has the highest impact on patient experience out of all issue areas that do not relate to the quality of service offered by doctors and nurses. Improvements to this dimension of the GP service could boost patient satisfaction with potential for cost-cutting opportunities in the NHS. It is also plausible to argue that if GP staff and patients spent less time on administrative efforts, patient satisfaction would likely improve on the most important aspects of satisfaction from GP services through an enhanced interaction of patients with medical and non-medical GP staff. Improving waiting times themselves for an already scheduled appointment is less important for patient satisfaction than ensuring that patients can schedule an appointment when they need it. Improving ease of booking an appointment is also more financially feasible to achieve on the national scale than overall shortening of appointment waiting times. Importantly, issues summarized with topic 2 "not enough time" were not featured in the most comprehensive GP Patient Survey[6] run by the NHS to gauge patients' opinions on GP services. The subject grouped words expressed to comment about the brevity of the appointment. Such issue omission in a national survey is unwelcome and worrying because it may lead to the inaccurate assessments of factors which affect patient satisfaction.

The generated insights point to a similar but wider range of patient issues than in the patient surveys but they also need to be treated with care. Among the insights, for example, it is evident that topic 10 "unprofessional care" is among the less important predictors affecting overall patient satisfaction. While on the national scale they are less salient issues, they likely have a significant effect in specific local contexts or for less numerous groups of individuals who are

[6] More information on the GP Patient Survey is available at: https://gp-patient.co.uk/SurveysAndReports, accessed 5 March 2018.

particularly concerned about those issues. Furthermore, there may be many issues which did not make it to the top 20 main topics extracted from the dataset of over 200,000 reviews which are very important to smaller groups of individuals.

Overall, Likert-scale evaluations are firmly related to topics about medical and administrative service experience. Relationships between service users and GP staff, accessibility of the services and the care and professionalism from GP staff towards users, are among the most important factors relating to satisfaction from GP services. Less important are waiting times for already scheduled appointments or instances of perceived medical mistreatment. More general opinions have a still lesser importance for the Likert-scale ratings of patients, probably because the sentiment of statements grouped into those topics tend to be mixed. Those include “time expressions” (topic 1) and “comparisons” (topic 6).

Insights into determinants of patient satisfaction, obtained through use of machine learning without any assumption about what is important for patients, may be useful for government efforts to increase patient satisfaction.

Robustness Analysis

Fixed-effects models were used to establish if, after controlling for other relevant variables, the statistically significant correlation between topic identified in text comments and star ratings still holds. For simplicity, topic proportions have been grouped into negative, neutral and positive clusters, in line with the color coding scheme from Figure 1 in the main paper. Percentage presence of positive and negative topics in comments were used as independent variables in fixed-effects models. Neutral topics have not been used in models to avoid a multicollinearity problem (topic proportions in reviews always sum to 1). Patient reviews were grouped according to month of posting and according to the NHS commissioning. Grouping

data eases computation of the fixed-effect models. Administrative data about GP practices was used to link GP comments to Clinical Commissioning Groups (CCG, a mid-level unit of NHS administration) which manage disbursement of funding for each GP practice[7]. Regional management style of NHS managers disbursing funds to GP practices and the month of posting reviews may affect satisfaction of patients. Two control variables were used as well: GP practice size expressed as the number of registered patients in a practice and average deprivation of patients. Counts of patients registered from each area of England (LSOA, Lower Layer Super Output Area – about 300 households per area)[8] were merged with data on levels of deprivation at each LSOA[9] to calculate the 2 control variables. Dataset mergers resulted in the inclusion of 205,214 reviews. We removed 3,073 reviews due to any missing attributes. Reviews of new, closed down, and/or less popular GP practices were more likely to be removed. On average, there were 17.7 reviews per CCG and month, in months when GPs funded by a given CCG received feedback. The panel dataset has 11594 cells for 209 CCGs over 60 months. There were almost 10 000 patients registered in GP practices on average, and average IMD deprivation score is at 5.37.

GP practices, avg. deprivation, avg. practice size (+min max).

The results of the linear two-way (CCG and month) fixed effects model are presented in Table 2[10]. They suggest that what patients write is significantly correlated to how they also rate

[7] Source: http://content.digital.nhs.uk/catalogue/PUB18468, last visited on 1st August 2017
[8] Source: https://data.gov.uk/dataset/numbers-of-patients-registered-at-a-gp-practice-lsoa-level, last visited on 1st August 2017
[9] Source: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015, last visited on 1st August 2017
[10] All fixed-effects models were calculated with R programming language, using *plm* package.

their experience after considering the available control variables. The cluster of positive topics predicts higher star ratings the more it is present in reviews, and the cluster of negative topics predicts lower star ratings the more it is present in reviews. While we do not have access to any external data for validation of the results, this expected direction of coefficients on positive and negative topic cluster variables when controlling for other covariates can be viewed as a weak form of validation. Another finding is that levels of deprivation in areas served by GP practices, combined with GP practice sizes, do not meaningfully change the relationship between star ratings and topics.

As part of the robustness analysis we replicated the key analysis in the paper using alternative number of estimated STM topics. In addition to our main 20-topic model, we also estimated 5-, 10-, 30-, and 40-topic models. The results are presented in Appendix C in supplementary materials.

[TABLE 2]

Limitations

The study's limitations relate to the relatively low response rate from patients. GP practices received approximately 27 reviews on average over a period of almost four and a half years. This makes comparison between individual practices infeasible. Instead, we have to limit comparison to mid-level administrative areas (CCGs). Given the effectiveness of the modelling approach, sufficient frequency of posting feedback is the prime limitation to real-time performance evaluations or evaluations on more granular level. Apart from that, the biases in the sample of patient experiences analyzed with the topic model are unknown and hard to predict (e.g., Xiang et al. 2017).

In addition, the data summarization deployed with the topic model has a few known methodological weaknesses (Grimmer and Stewart 2013). These include: 1) possible misalignment between topic proportional presence in reviews and topic importance for users, 2) unavoidable uncertainty over how many topics to generate to best represent reviews, as well as 3) crude assumptions made about natural language in the design of the topic model.

Therefore, it is advisable to compare topic model results obtained from online reviews with a representative and systematic survey of service user opinions about their service experience. The comparison could help to establish the representativeness of topic modeling outcomes. In the instance of the NHS, the GP Patient Survey is the most systematic and regularly collected opinion survey about GP services in England available (Cowling, Harris, and Majeed 2015). It could be used to validate topic model outcomes. Validated topic models can in turn help decrease in the frequency and cost of data collection with mass patient surveys by obtaining proxy survey values from text comments.

Discussion and conclusion

Text comments posted by citizens and processed with machine learning are a possible avenue for addressing the deficit of citizens' contribution towards innovation process in public services (O'Leary 2016; De Vries, Bekkers, and Tummers 2016). Decisions about reform of public services are increasingly backed with data (Hood and Dixon 2015). Cues can come from small-sample qualitative studies (Salt, Rowles, and Reed 2012), citizen surveys (Van Ryzin and Charbonneau 2010) or easy-to-access quantitative measures of citizen behavior such as the number of visits at public service providers (Hood and Dixon 2015). Qualitative studies offer comprehensive insights but a small sample size and high costs of data collection can make them unfeasible for decision-making. Surveys are reliable but can only cover narrow sub-

samples of citizen experiences. Any understanding of public issues from survey data limits and biases insights by over-emphasizing what is known. Behavioral data, in turn, give little or no information on the reasons for certain behaviors, which means any decisions on such insights are prone to also produce undesirable side-effects. Systematic and exhaustive inclusion of citizen voices with machine learning is a desirable improvement over weaknesses of predominant ways to include citizens' voice in political and public policy decisions. Large data can be processed to identify a full spectrum of citizens' concerns. Better understanding of the narratives which drive patient behaviors can be used for instance to help deliver more cost-effective healthcare (Vlaev et al. 2016; Eton 2017; McClellan 2011). Half of all deaths in USA have been self-inflicted, which is indicative of the proportion of costs in healthcare that could be avoided with behavior change (Vlaev et al. 2016).

The feedback insights we obtained can be produced quickly from very large samples of patient opinion. Moreover, the obtained insights shed light on issue areas which were wholly excluded from the most comprehensive survey data on GP service experience. For example, the issue of management style in the GP practice (topic 4) is comparable in salience to the availability of appointments. Of the two, only the latter issue was included in GP Patient Survey. Insights from surveys that omit salient issues reported by patients may lead to sub-optimal decisions in the NHS attempting to improve patients' service experience.

The examination of what makes patients happy or unhappy as showcased here, despite unknown sample biases, may help administrators in the National Health Service identify and learn from successful GP practices across England. Patient feedback can be clustered according to the NHS institution to which it relates, giving insight into patterns of satisfaction and GP management styles across the country. It may also be useful to identify GP services which suffer from factors such as poor telephone access, thus requiring improvement. The reviews

themselves can also be clustered according to their topical structure and Likert-scale satisfaction levels to understand the prevalent narratives of users about their service experience. It would be important to better understand where, how and why this occurs. It is hard to identify and analyse those reviews when they are in large volume without means for making a quantitative representation of their content.

The insights generated in this study also point to key challenges facing public institutions that could be overcome nationally by the NHS for the sake of efficiency for all patients, rather than at GP or CCG level. We observe that many patients express frustration with the difficulty of making GP appointments. If comments are frequent enough, machine learning models may also be able to generate near real-time insights into patient satisfaction on the level of individual GP practices or doctors. It can also be known how policies affect patients over time, whether some areas suffer from significant shifts in perceived GP service quality, and how the impact of NHS decisions varies in different locations.

Finally, insights from this study may help to inform public preferences regarding NHS services. The public should be able to obtain information about current NHS challenges through the lens of actual GP reviews as they are written by patients, as opposed to a limited range of hard figures prepared by the public service provider. Quantitative summaries of written feedback at national or regional level give extra advantage to members of the public who require improvements to be made.

In summary, researchers and public managers can use machine learning for text analysis to benefit from inquiries into user satisfaction from public services. For example, in the instance of public healthcare in England, topic model outcomes obtained from online reviews suggest that patients tend to comment proficiently about their difficulties in accessing GP services, but this is not the most important predictor of satisfaction with the health services. Instead, how

GP staff treat patients is what most critically determines whether users rate their experience highly or not, then followed by patients' experience with reception staff and lack of appointments. Potentially, a change in communication style by NHS staff, aided by a more convenient online booking service and streamlined bureaucratic procedures, could help to lift patient satisfaction in relation to GP services despite difficulties in getting a GP appointment. The tools and insights of this study can be publicly available, in this way responding to the demand for more inclusive decisions about public service provision (O'Leary 2016).

References

Amirkhanyan, A. A., Kim, H. J. & Lambright, K. T. (2013). The performance puzzle: Understanding the factors influencing alternative dimensions and views of performance. *Journal of Public Administration Research and Theory*, 24(1), 1–34.

Andersen, L. B., Heinesen, E. & Pedersen, L. H. (2016). Individual performance: From common source bias to institutionalized assessment. *Journal of Public Administration Research and Theory*, 26(1), 63–78.

Andersen, S. C. & Hjortskov, M. (2016). Cognitive biases in performance evaluations. Journal of Public Administration Research and Theory, 26(4), 647–662.

Barrows, S., Henderson, M., Peterson, P. E. & West, M. R. (2016). Relative performance information and perceptions of public service quality: Evidence from American school districts. Journal of Public Administration Research and Theory, 26(3), 571–583.

Bartenberger, M. & Dawid, S. (2016). The benefits and risks of experimental co-production: The case of urban redesign in Vienna. Public Administration, 94(2), 509–525.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77–84.

Blei, D. M. & Lafferty, J. D. (2006). Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning*, 113–120.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research,* 3, 993–1022.

Brenninkmeijer, A. (2016). Interfaces: How to connect effectively with citizens. *Public Administration Review,* 77(1), 10–11.

Brown, T. (2007). Coercion versus choice: Citizen evaluations of public service quality across methods of consumption. *Public Administration Review,* 67(3), 559–572.

Brownson, R. C., Allen, P., Duggan, K., Stamatakis, K. A. & Erwin, P. C. (2012). Fostering more effective public health by identifying administrative evidence-based practices: A review of the literature. *American Journal of Preventive Medicine,* 43(3), 309–319.

Burton, T. T. (2012). Technology: Enabler or inhibitor of improvement? *Process Excellence Network* accessed at http://www.processexcellencenetwork.com/business-process-management-bpm/articles/technology-enabler-or-inhibitor-of-improvement/.

Christensen, T. & Laegreid, P. (2005). The relative importance of service satisfaction, political factors, and demography. *Public Performance and Management Review,* 28(4), 487–511.

Cowling, T. E., Harris, M. J. & Majeed, A. (2015). Evidence and rhetoric about access to UK primary care. *British Medical Journal,* 350, h1513–h1513.

Dai, A. M. & Storkey, A. J. (2015). The supervised hierarchical dirichlet process. *IEEE Transactions on Pattern Analysis and Machine Learning,* 37(2), 243-255.

De Vries, H., Bekkers, V. & Tummers, L. (2016). Innovation in the public sector: A systematic review and future research agenda. *Public Administration,* 94(1), 146–166.

Di Pietro, L., Mugion, R. & Renzi, M. F. (2013). An integrated approach between lean and customer feedback tools: An empirical study in the public sector. *Total Quality Management and Business Excellence,* 24(7-8), 899–917.

Eton, D. T., et al. (2017). Healthcare provider relational quality is associated with better self-management and less treatment burden in people with multiple chronic conditions. *Patient Preference and Adherence,* 11, 1635–1646.

Feldman, D. L. (2014). Public value governance or real democracy. *Public Administration Review,* 74(4), 504–505.

Fung, A. (2015). Putting the public back into governance: The challenges of citizen participation and its future. *Public Administration Review,* 75(4), 513–522.

Gao, J. (2015). Pernicious manipulation of performance measures in China's cadre evaluation system. *The China Quarterly,* 223, 618–637.

Gray, M. (2015). The social media effects of a few on the perceptions of many. *Public Administration Review,* 75(4), 607–608.

Greer, S. L., Wilson, I., Stewart, E. & Donnelly, P. D. (2014). 'Democratizing' public services? Representation and elections in the Scottish NHS. *Public Administration,* 92(4), 1090–1105.

Griffiths, T. L. & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America,* 101(S1), 5228–5235.

Grimmer, J. & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis,* 21(3), 267-297.

Harding, J. (2012). Choice and information in the public sector: A higher education case study. *Social Policy and Society,* 11(2), 171–182.

Hartley, J. & Lucy, R. B. (2010). Four layouts and a finding: the effects of changes in the order of the verbal labels and numerical values on Likert-type scales. *International Journal of Social Research Methodology*, 13(1), 17–27.

Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The elements of statistical learning*. New York, NY: Springer.

Hogenboom, F., Frasincar, F., Kaymak, U., de Jong, F. & Caron, E. (2016). A survey of event extraction methods from text for decision support systems. *Decision Support Systems*, 85, 12–22.

Hong, S. (2015). Citizen participation in budgeting: A trade-off between knowledge and inclusiveness? *Public Administration Review,* 75(4), 572–582.

Hood, C. & Dixon, R. (2015). What we have to show for 30 years of new public management: Higher costs, more complaints. *Governance,* 28(3), 265–267.

Im, T., Cho, W., Porumbescu, G. & Park, J. (2012). Internet, trust in government, and citizen compliance. *Journal of Public Administration Research and Theory,* 24(3), 741–763.

James, O. & Moseley, A. (2014). Does performance information about public services affect citizens' perceptions, satisfaction and voice behaviour? Field experiments with absolute and relative performance information. *Public Administration,* 92(2), 493–511.

Jensen, U. T. & Andersen, L. B. (2015). Public service motivation, user orientation, and prescription behaviour: Doing good for society or for the individual user? *Public Administration*, 93(3), 753–768.

Jlike, S., Meuleman, B. & Van de Walle, S. (2014). We need to compare, but how? Measurement equivalence in comparative public administration. *Public Administration Review*, 75(1), 36–48.

Kelly, J. M. (2005). The dilemma of the unsatisfied customer in a market model of public administration. *Public Administration Review,* 65(1), 76–84.

Kong, H-S. & Song, E-J. (2016). A study on customer feedback of tourism service using social big data. *Information,* 19: 49–54.

Kroll, A. (2017). Can performance management foster social equity? Stakeholder power, protective institutions, and minority representation. *Public Administration,* 95(1), 22–38.

Ladhari, R. & Rigaux-Bricmont, B. (2013). Determinants of patient satisfaction with public hospital services. *Health Marketing Quarterly,* 30(4), 299–318.

Lavertu, S. (2014). We all need help: 'Big data' and the mismeasure of public administration. *Public Administration Review,* 76(6), 864–872.

Lawton, A. & Macaulay, M. (2013). Localism in practice: Investigating citizen participation and good governance in local government standards of conduct. *Public Administration Review,* 74(1), 75–83.

Liu, H. K. (2016). Bring in the crowd to reinventing government. *Journal of Public Administration Research and Theory,* 26(1), 177–181.

Loeffler, E. (2016). Coproduction of public outcomes: Where do citizens fit in? *Public Administration Review,* 76(3), 436–437.

Luciana, A. (2013). Organizational learning and performance. A conceptual model. *Proceedings of the 7th International Management Conference*, 547–556.

Ma, L. (2017). Performance management and citizen satisfaction with the government: Evidence from Chinese municipalities. *Public Administration,* 95(1), 39–59.

Mahmoud, M. A. & Hinson, R. E. (2012). Market orientation in a developing economy public institution: Revisiting the Kohli and Jaworski's framework. *International Journal of Public Sector Management,* 25(2), 88–102.

Marvel, J. D. (2016). Unconscious bias in citizens' evaluations of public sector performance. *Journal of Public Administration Research and Theory,* 26(1), 143–158.

McClellan, M. (2011). Reforming payments to healthcare providers: The key to slowing healthcare cost growth while improving quality? *Journal of Economic Perspectives*, 25(2), 69–92.

Mergel, I., Rethemeyer, K. R. & Isett, K. (2016). Big data in public affairs. *Public Administration Review*, 76(6), 928–937.

Moon, S. J. (2015). Citizen empowerment: New hope for democratic local governance. *Public Administration Review,* 75(4), 584.

Moynihan, D. P., Herd, P. & Harvey, H. (2014). Administrative burden: Learning, psychological, and compliance costs in citizen-state interactions. *Journal of Public Administration Research and Theory*, 25(1), 43–69.

Müssener, U., Bendtsen, M., McCambridge, J. & Bendtsen, P. (2016). User satisfaction with the structure and content of the NEXit intervention, a text messaging-based smoking cessation programme. *BMC Public Health,* 16, 1179.

O'Leary, I. (2016). Big data in public affairs. *Public Administration Review*, 76(6), 928–937.

Olsen, A. L. (2015). Citizen (dis)satisfaction: An experimental equivalence framing study. *Public Administration Review,* 75(3), 469–478.

Pierre, J. & Røiseland, A. (2016). Exit and voice in local government reconsidered: A 'choice revolution'? *Public Administration,* 94(3), 738–753.

Qi, J., Zhang, Z., Jeon, S. & Zhou, Y. (2016). Mining customer requirements from online reviews: A product improvement perspective. *Information and Management,* 53(8), 951–963.

Roberts, M. E., Steward, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Kushner-Gadarian, S., Albertson, B. & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science,* 58(4), 1064–1082.

Roberts, M. E., Steward, B. M. & Tingley, D. (2015). Navigating the local modes of big data: The case of topic models. In Alvarez M. R. (Ed.), *Computational Social Science,* 51–97. New York, NY: Cambridge University Press.

Rogge, N., Agasisti, T. & De Witte, K. (2017). Big data and the measurement of public organizations' performance and efficiency: The state-of-the-art. *Public Policy and Administration,* 32, 263–281.

Salt, E., Rowles, G. D. & Reed, D. B. (2012). Patient's perception of quality patient–provider communication. *Orthopaedic Nursing,* 31(3), 169–176.

Sanders, K. & Canel, M. J. (2015). Mind the gap: Local government communication strategies and Spanish citizens' perceptions of their cities. *Public Relations Review,* 41, 777–784.

Scott, D. & Vitartas, P. (2008). The role of involvement and attachment in satisfaction with local government services. *International Journal of Public Sector Management,* 21(1), 45–57.

Song, M. & Meier, K. J. (2018). Citizen satisfaction and the kaleidoscope of government performance: How multiple stakeholders see government performance. *Journal of Public Administration Research and Theory,* 28(4), 489–505.

Taylor, C. D. (2015). Property tax caps and citizen perceptions of local government service quality: Evidence from the Hoosier Survey. *American Review of Public Administration*, 45(5), 525–541.

Van de Walle, S. & Van Ryzin, G. G. (2011). The order of questions in a survey on citizen satisfaction with public services: Lessons from a split-ballot experiment. *Public Administration*, 89(4), 1436–1450.

Van Ryzin, G. G. & Charbonneau, E. (2010). Public service use and perceived performance: An empirical note on the nature of the relationship. *Public Administration*, 88(2), 551–563.

Villegas, J. A. (2017). Perception and performance in effective policing. *Public Administration Review*, 77(2), 240–241.

Vlaev, I., King, D., Dolan, P. & Darzi, A. (2016). The theory and practice of "nudging": changing health behaviors. *Public Administration Review*, 76(4), 550–561.

Voutilainen, A., Pitkaaho, T., Kvist, T. & Vehvilainen-Julkunen, K. (2015). How to ask about patient satisfaction? The visual analogue scale is less vulnerable to confounding factors and ceiling effect than a symmetric Likert scale. *Journal of Advanced Nursing*, 72(4), 946–957.

Walker, R. M. & Boyne, G. A. 2009. Introduction: Determinants of performance in public organizations. *Public Administration*, 87(3), 433–439.

Xiang, Z., Du, Q., Ma, Y. & Fan, W. (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. *Tourism Management*, 58, 51–65.

Yang, Y. (2010). Adjusting for perception bias in citizens' subjective evaluations. *Public Performance and Management Review*, 34(1), 38–55.

Table 1: Topic labels: 20-topic STM model labeled by the authors.

Topic 1	Topic 2	Topic 3	Topic 4
time expressions	not enough time	proper treatment	poor management
Topic 5	Topic 6	Topic 7	Topic 8
diagnosed and sorted	comparisons	recommend	helpful
Topic 9	Topic 10	Topic 11	Topic 12
thanks	unprofessional care	unwelcoming	poor phone access
Topic 13	Topic 14	Topic 15	Topic 16
prescription problem	discourage registration	great	lack manners
Topic 17	Topic 18	Topic 19	Topic 20
hard appointments	no appointments	late appointments	rude reception

Note:

Table 2: Two-way fixed-effects models

	Phone access ease	Appoint-ment ease	Dignity and respect	Involved in care decisions	Likely to recommend	Up-to-date GP details
Positive topics	2.30 ***	3.23 ***	4.05 ***	4.61 ***	5.14 ***	3.32 ***
	(0.16)	(0.18)	(0.19)	(0.19)	(0.21)	(0.16)
Negative topics	-3.36 ***	-3.77 ***	-1.89 ***	-1.12 ***	-3.18 ***	-2.15 ***
	(0.18)	(0.18)	(0.20)	(0.19)	(0.21)	(0.16)
Average deprivation (IMD) score	0.03 **	0.03 **	0.03 *	0.04 **	0.03*	0.05 ***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Number of patients	-0.00 ***	-0.00 ***	0.00*	0.00	-0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
CCG FE	YES	YES	YES	YES	YES	YES
Month FE	YES	YES	YES	YES	YES	YES
R2	0.46	0.54	0.43	0.40	0.58	0.40
Adj R2	0.45	0.53	0.42	0.40	0.57	0.39
Num. Obs.	11594	11594	11594	11594	11594	11594

Notes: Outcomes of two-way fixed-effects models take into account variance in the review data that results from differences between Clinical Commissioning Groups (NHS units responsible for funding allocations to GP practices) and monthly time periods when the reviews were posted. Likert-scale star ratings are the dependent variables. Topic proportions within documents are the independent variables. Topic proportions have been clustered into positive, negative, and neutral – in line with the schema in Figure 1. The neutral cluster has been excluded to avoid perfect multicollinearity. The models included two control variables. The average index of multiple deprivation (IMD) score (1 is the best and 100 is the worst) of

*patients using GP services, as well as a count of how many patients are registered at a reviewed GP practice (a proxy value correcting for GP size). Robust standard errors for coefficients are reported in brackets. Significance: *** p < 0.001, ** p < 0.01, * p < 0.05*

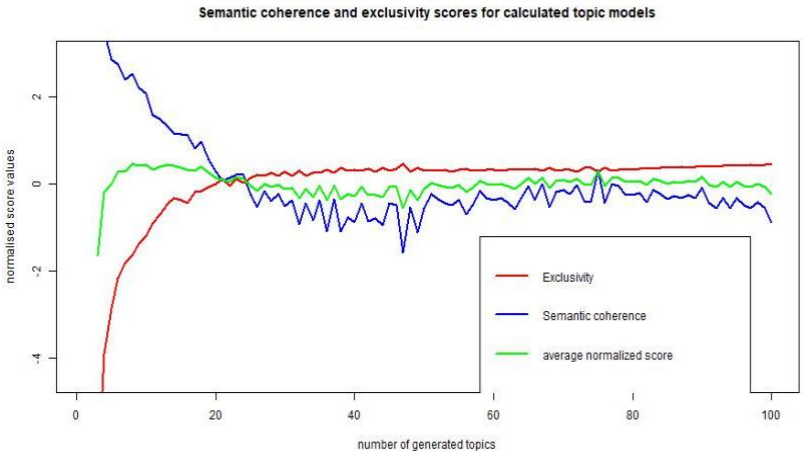
Supplementary Materials

Appendices A – E

Appendix A: Selecting the number of topics for STM analysis

Topic models containing from 3 up to 100 topics were calculated from pre-processed data and compared in order to identify the optimal number of topics for modeling. Following Roberts et al. (2015), 97 topic models were evaluated with semantic coherence (the rate at which the topic's most common words tend to occur together in the same reviews) and exclusivity (the rate at which most common terms are exclusive to individual topics) scores. The model with 20 topics had one of the best combination of semantic coherence and exclusivity scores out of all models. More complex models which had more topics tended to have lower semantic coherence scores and did not meaningfully improve over the quality of the 20-topic STM model (see Figure A1).

Figure A1: Semantic coherence and exclusivity scores for calculated topic models



Notes: (1) The illustration portrays semantic coherence (the rate at which each topic's most common words tend to occur together in the same reviews) and exclusivity (the rate at which most common terms are exclusive to individual topics) for topic models with up to 100 generated topics. Higher semantic coherence and exclusivity scores tend to correlate with higher perceived quality of generated topics. (2) Scores were normalized by dividing individual model scores by average scores for all models.

Appendix B: Explanation of topic labeling

The 20 topics generated with the chosen STM topic model have been labeled according to the most frequently occurring words in topics, as well as the written reviews which are representative of each topic. Table B1 below lists the seven most frequently occurring terms for each topic, while Table B2 includes the labels assigned to each topic together with a review representing the topic. Representative reviews have been identified by the high proportion of terms within reviews classified into a given topic.

Table B1: Most prominent words for STM model with 20 topics

Topic 1 Top Words: last, week, two, month, first, time, now	**Topic 11 Top Words:** like, say, feel, know, realli, just, want
Topic 2 Top Words: need, see, time, one, problem, can, make	**Topic 12 Top Words:** call, told, phone, back, answer, ring, got
Topic 3 Top Words: medic, health, issu, visit, treatment, concern, condit	**Topic 13 Top Words:** prescript, inform, repeat, request, medic, contact, order
Topic 4 Top Words: practic, patient, manag, seem, quot, nhs, poor	**Topic 14 Top Words:** ask, regist, letter, wrong, complet, anoth, told
Topic 5 Top Words: hospit, pain, refer, referr, prescrib, suffer, symptom	**Topic 15 Top Words:** good, well, year, seen, servic, great, also
Topic 6 Top Words: servic, use, move, gps, area, new, difficult	**Topic 16 Top Words:** patient, staff, recept, deal, member, person, peopl

Topic 7 Top Words: practic, recommend, excel, profession, nurs, famili, year	**Topic 17 Top Words:** day, get, book, work, system, tri, avail
Topic 8 Top Words: alway, help, staff, friend, recept, listen, polit	**Topic 18 Top Words:** get, even, never, dont, cant, will, just
Topic 9 Top Words: care, thank, receiv, support, provid, team, kind	**Topic 19 Top Words:** wait, time, hour, minut, walk, seen, late
Topic 10 Top Words: test, nurs, went, blood, result, said, check	**Topic 20 Top Words:** receptionist, rude, speak, talk, person, one, way

Table B2: Topic labels with representative reviews

Topic 1	Topic 2	Topic 3
time expressions	not enough time	proper treatment
"Been with this surgery since I moved to Huddersfield almost 25 years ago. Nothing has changed in that time and there is a reason for that. Still offering 2 periods of open surgery 3 days per week, still the same quality of care. Just a shame they have to take holidays and we lose them for a couple of weeks every year."	"Tell one doctor your problems and they usually solve them, although sometimes we think we should have a little more time. When you get older you may have more problems which can not be solved in 10 minutes and have to make another appointment."	"I have been a regular visitor to the Practice for over 10 years due to ongoing health issues (Hypertension, cholesterol) The management programme put in place by the Practice and regular reviewing of the programme has ensured that my conditions are well controlled and do not inhibit my life in any way."
Topic 4	Topic 5	Topic 6
poor management	diagnosed and sorted	comparisons
"In March 16 we were promised that we would have permanent GPs by June 16. In Jan 17, we only have 1permanent GP for 2 practices. Overuse of locums, no consistency, rarely see same locum twice, no consistency. 2 permanent GPs were	"throat problem referred to hospital assessed by consultant on the 10 day following the surgery list. Followed up with advice from GP. HIP pain referred for x ray - phoned hospital x ray completed same day. Followed up with chat with GP."	"I recently moved here from a large metropolitan city in the north west the surgery I used there was perfect for me for the 10 years I lived there so I was concerned about moving to a new surgery in a new town that would live up to what I had, with the Orchard practice it proved within the a few visits this a great practice and with a friendly team of staff"

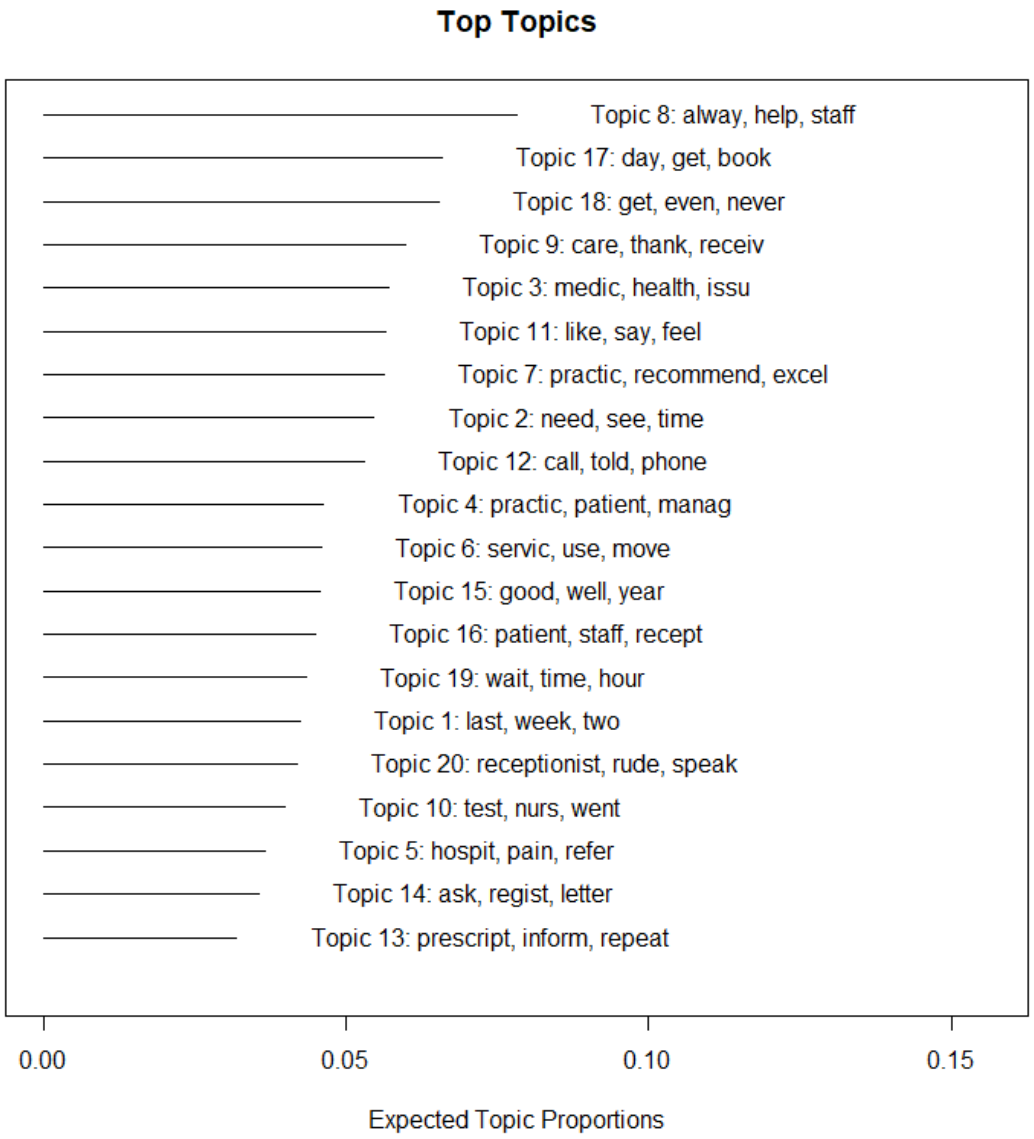
employed, but both resigned withina couple of months! Terrible practice, I will be moving to another practice!"		
Topic 7	Topic 8	Topic 9
recommend	helpful	thanks
"All the doctors practice nurses and clerical staff are extremely caring efficient and helpful in every possible way I highly recommend this practice."	"The reception staff are extremely helpful! They always treat you with respect and are always happy to go out of their way to help you out."	"I am not a patient but the care shown to my mother and father in law is the best.hes 92 she is 81 father in law been Very Ill this year and care and support has been amazing from the whole team thank you all"
Topic 10	Topic 11	Topic 12
unprofessional care	unwelcoming	poor phone access
"went for blood test. when I back for the results the thyroid function had not been checked so had to make another appointment for blood test."	"Anytime I go there I feel really uncomfortable, maybe because of the secretary that makes you feel stupid everything you ask them, and they make you feel like we are doing a favor to you. The doctor is Ok but they should be more approachable (they dont say hi when you go in) also they are like they are doing a favor to you and you shouldnt be there."	"21/03/16 called surgery at 0801 told I was no 11 on hold ok. 0834 told I was no 1 then was cut off. called straight back told I was no 19 on hold ok 0854 told I was no2 0855 I was cut off"
Topic 13	Topic 14	Topic 15
prescription problem	discourage registration	great
"Failed to action a request faxed to them by my consultant. They very often change the prescription service without informing me. Changed from collection to direct to pharmacy. They recently sent my repeat prescription to the wrong pharmacy."	"We were unable to register at this practice because our driving licences (re-issued earlier this year when we moved to Bromley) were not accepted as proof of address. Instead, we were told to present a bank statement or utility bill. We regard this as an arbitrary and unreasonable decision and have since registered	"used same place for many years and i have brought my children here too and both young adults, plesant and always clean and tidy and very helpful staff, keep up the great work"

	at another surgery where our driving licences were accepted without question."	
Topic 16	Topic 17	Topic 18
lack manners	hard appointments	no appointments
"reception staff are bad mannered to patients , cancelling appointments with very little notice , namely reception staff have little interest in patients needs. no member of staff at this surgery has the slightest interest."	"Though the doctor at the surgery is very good, its almost impossible to get an appointment. Allows telephone bookings only, and lines open only when the doctor is in the surgery. No system for booking in advance / early, either by phone or by any other means."	"Theres no point in being registered at this surgery! Can never get through and when you finally do theres never any appointments anyway, absolutely useless."
Topic 19	Topic 20	
late appointments	rude reception	
"Always late running not very good explanation Given no apology given even after waiting for 1 hour after appointment time seen late all the time some times upto 1an half hour late"	"The receptionist is very rude. No manners at all. Very lazy. I have also heard them speak to other people in this manor but I dont think people have complained. The doctor is good but the receptionist extremely rude."	

The features extracted from text reviews with the STM topic model relate to a range of patient experiences. Some relate to whether or not GP staff were helpful and nice, to cases of perceived misdiagnosis and difficulties in obtaining a GP appointment over phone or otherwise. Topic 6 also grouped comparative assessments of GP services, and topic 1 clustered terms used to express passage of time. It appears that some topics could be broken up into sub-topics. For example, topic 1 “time expressions” appears to cluster both reviews rich in expressions of time periods as well as reviews which include meaningful expressions of GP service experience over longer periods of time. The topics had a varying prevalence across the GP reviews dataset (see Figure B1 below), from over 3% of all tokens in the dataset to almost 8%. The model clustered reviews according to the choices of vocabulary

used by reviewers. Topic 8 "helpful" was the most prevalent, followed by topic 17 "hard appointments". Topics about the difficulty of obtaining or scheduling an appointment (12, 14, 17, 18, 19) featured prominently as a group, cumulatively constituting about 26% of all content in reviews on average. Figure B1 presents proportion of appearance in the corpus for each topic in our estimation.

Figure B1: Topic proportions in the GP reviews dataset



Appendix C: Examination of STM models with 5, 10, 30, and 40 topics

As part of the robustness analysis alternative STM models (with 5, 10, 30 and 40 topics) have been investigated for evaluate relative performance of the 20-topic model. Overall, our analysis shows that models with fewer topics retain thematic duplicates if some general theme is very common in reviews (see Tables C1-C4). Even the STM model with five topics is comprised of two covering the issue of rudeness and the interrelated issue of accessing the services (Table C1).

Table C1: Most prominent words for LDA model with 5 topics

Topic 1 Top Words: practic, patient, medic, servic, health, year, issu
Topic 2 Top Words: alway, help, staff, care, nurs, year, practic
Topic 3 Top Words: ask, told, prescript, said, test, hospit, went
Topic 4 Top Words: receptionist, one, like, rude, staff, recept, don't
Topic 5 Top Words: get, call, time, day, phone, wait, see

Table C2: Most prominent words for LDA model with 10 topics

Topic 1 Top Words:	**Topic 6 Top Words:**

time, use, work, servic, patient, new, telephon	call, told, wait, back, week, minut, got
Topic 2 Top Words: practic, medic, health, patient, issu, nhs, experi	**Topic 7 Top Words:** prescript, repeat, month, medic, ask, request, didn't
Topic 3 Top Words: time, one, never, see, like, problem, say	**Topic 8 Top Words:** staff, good, recept, practic, servic, patient, year
Topic 4 Top Words: alway, help, care, friend, nurs, thank, recommend	**Topic 9 Top Words:** get, day, phone, book, tri, can, time
Topic 5 Top Words: test, hospit, nurs, blood, result, visit, pain	**Topic 10 Top Words:** receptionist, rude, peopl, patient, recept, person, speak

Table C3: Most prominent words for LDA model with 30 topics

Topic 1 Top Words: week, two, nurs, month, first, last, clinic	**Topic 11 Top Words:** alway, help, staff, friend, nurs, great, recept	**Topic 21 Top Words:** never, ever, bad, place, look, absolut, one
Topic 2 Top Words: staff, recept, peopl, rude, patient, person, member	**Topic 12 Top Words:** feel, like, treat, way, make, understand, made	**Topic 22 Top Words:** time, seen, problem, long, take, see, walk
Topic 3 Top Words: test, blood, result, done, check , nurs, pressur	**Topic 13 Top Words:** care, thank, kind, support, team, much, famili	**Topic 23 Top Words:** good, experi, realli, keep, also, sometim, busi
Topic 4 Top Words: see, one, will, thing, ill, though, sure	**Topic 14 Top Words:** said, went, didnt, ask, took, daughter, got	**Topic 24 Top Words:** issu, inform, contact, manag, requir, medic, regard
Topic 5 Top Words: practic, servic, excel, profession, recommend, high, effici	**Topic 15 Top Words:** medic, condit, serious, life, health, symptom, treatment	**Topic 25 Top Words:** prescript, repeat, request, medic, order, pharmaci, collect
Topic 6 Top Words:	**Topic 16 Top Words:**	**Topic 26 Top Words:**

seem, patient, practic, manag, poor, quot, amp	patient, review, better, find, may, read, think	get, can, work, need, system, abl, cant
Topic 7 Top Words: just, dont, even, want, know, say, tell	**Topic 17 Top Words:** call, told, back, got, today, morn, rang	**Topic 27 Top Words:** receptionist, wait, hour, minut, anoth, late, room
Topic 8 Top Words: visit, recent, advic, within, attend, quick, given	**Topic 18 Top Words:** day, book, week, avail, emerg, open, tri	**Topic 28 Top Words:** phone, tri, answer, ring, minut, line, get
Topic 9 Top Words: year, ive, gps, mani, past, last, now	**Topic 19 Top Words:** servic, patient, centr, use, access, park, consid	**Topic 29 Top Words:** patient, practic, nhs, consult, provid, continu, number
Topic 10 Top Words: regist, move, new, sinc, chang, area, now	**Topic 20 Top Words:** hospit, pain, refer, referr, specialist, sent, examin	**Topic 30 Top Words:** ask, complet, form, name, refus, letter, complaint

Table C4: Most prominent words for LDA model with 40 topics

Topic 1 Top Words: inform, regist, letter, contact, complet, form, address	**Topic 11 Top Words:** patient, gps, good, mani, work, well, other	**Topic 21 Top Words:** wait, minut, late, room, min, sit, turn	**Topic 31 Top Words:** patient, quot, access, appear, communic, lack, general
Topic 2 Top Words: check, clinic, nurs, first, time, babi, attend	**Topic 12 Top Words:** one, occas, now, need, see, time, last	**Topic 22 Top Words:** feel, much, experi, like, realli, say, way	**Topic 32 Top Words:** ask, said, didnt, went, tell, couldnt, got
Topic 3 Top Words: can, sometim, one, find, quit, time, good	**Topic 13 Top Words:** dont, know, want, just, like, ever, bad	**Topic 23 Top Words:** prescript, repeat, request, order, pharmaci, collect, readi	**Topic 33 Top Words:** medic, without, month, despit, review, chang, prescrib
Topic 4 Top Words: made, visit, explain, concern, felt, feel, discuss	**Topic 14 Top Words:** nurs, great, happi, found, quick, good, well	**Topic 24 Top Words:** book, day, system, work, avail, onlin, can	**Topic 34 Top Words:** time, see, need, emerg, long, seen, urgent

Topic 5 Top Words: get, phone, tri, ring, line, morn, answer	**Topic 15 Top Words:** year, sinc, now, old, children, littl, drs	**Topic 25 Top Words:** time, can, need, often, lot, fault, find	**Topic 35 Top Words:** peopl, time, thing, take, sure, one, need
Topic 6 Top Words: thank, receiv, famili, year, husband, support, care	**Topic 16 Top Words:** alway, help, staff, best, polit, friend, love	**Topic 26 Top Words:** nhs, manag, complaint, respons, regard, read, comment	**Topic 36 Top Words:** week, time, two, hour, get, wait, see
Topic 7 Top Words: health, issu, condit, serious, problem, life, sever	**Topic 17 Top Words:** care, treat, respect, team, support, kind, level	**Topic 27 Top Words:** call, told, back, day, next, today, rang	**Topic 37 Top Words:** ive, never, cant, actual, even, absolut, one
Topic 8 Top Words: get, seem, difficult, imposs, make, can, almost	**Topic 18 Top Words:** problem, better, need, time, park, although, also	**Topic 28 Top Words:** staff, recept, patient, member, deal, person, front	**Topic 38 Top Words:** servic, offer, telephon, use, consult, abl, within
Topic 9 Top Words: see, walk, left, anoth, centr, even, though	**Topic 19 Top Words:** receptionist, rude, speak, attitud, unhelp, person, extrem	**Topic 29 Top Words:** pain, son, daughter, infect, gave, took, prescrib	**Topic 39 Top Words:** answer, open, number, someon, queue, close, phone
Topic 10 Top Words: practic, year, regist, amp, anyon, join, recommend	**Topic 20 Top Words:** test, hospit, blood, result, refer, referr, follow	**Topic 30 Top Words:** excel, profession, recommend, friend, high, effici, servic	**Topic 40 Top Words:** move, new, area, year, look, live, hous

Themes which may be of interest to NHS decision makers but are more specific to individuals, such as experiences of acute health problems, the handling of repeat prescriptions or comments about hospital referrals, disappear from the topic lists as models are trained to produce fewer topics. For example, topics 5, 27, 38 and 39 from the 40-topic STM model all have portions of their vocabularies related to phone calls made by patients (Table C4). The model with 20 topics (Table B1) compresses portions of those subjects together with a ‘poor telephone access’ theme. Similarly, comments present in the 40-topic model, such as topic 20 about hospital referrals topic 23 about repeat prescriptions disappear altogether in models with fewer topics.

Linear regressions, lasso models and cross-validation calculations have also been carried out for the same set of models as in the main paper. The results were compared (Tables C5 and C6). Cross-validation errors for linear regressions and lasso yield almost identical prediction errors. This is because the Lasso regression's optimal shrinkage parameter was almost 0, which meant that the Lasso penalty did not meaningfully exclude or reduce any of the predictors. All predictive models perform better than the baseline, i.e. predicting a star rating using average star rating.

Table C5: 5-fold cross-validation errors for linear regression models

# of topics	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Mean
5	1.206	1.207	1.265	1.422	1.376	1.398	1.312
10	1.140	1.148	1.105	1.336	1.247	1.306	1.214
20	1.096	1.078	1.078	1.255	1.153	1.232	1.148
30	1.098	1.107	1.099	1.272	1.181	1.247	1.167
40	1.070	1.066	1.060	1.252	1.128	1.231	1.135
Standard deviations of star ratings	1.484	1.615	1.587	1.604	1.841	1.546	1.613

Notes: In the illustration below, star ratings are the dependent variables. Topic proportions in documents are the independent variables. The lower the mean squared prediction error, the better the model. Green indicates the best model.

Table C6: 5-fold cross-validation errors for lasso models

# of topics	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Mean
5	1.206	1.207	1.265	1.422	1.376	1.398	1.312
10	1.140	1.148	1.105	1.336	1.247	1.306	1.214
20	1.096	1.078	1.078	1.255	1.153	1.232	1.149
30	1.098	1.107	1.099	1.272	1.181	1.247	1.167
40	1.070	1.066	1.060	1.252	1.129	1.231	1.135
Standard deviations	1.484	1.615	1.587	1.604	1.841	1.546	1.613

of star ratings							

Notes: In the illustration below, star ratings are the dependent variables. Topic proportions in documents are the independent variables. The lower the mean squared prediction error, the better the mode. Green indicates the best model.

It is better to avoid comparisons between topics from different models based on their top words at face value. Topics with seemingly overlapping meanings have very different coefficient values in regression models with the same dependent variables. For example, topics 5, 27, 38 and 39 from the 40-topic STM model, which all relate to telephone access, have varying coefficient values in lasso model outcomes (Table C7) while in the 20-topic model there is "poor telephone access" topic which does not properly represent such differentiation (Table C8).

Table C7: 40-topic STM – Predictors for lasso models where star ratings are the dependent variable

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
topic 1	-2.31	-2.77	-4.03	-4.01	-3.02	-4.49
topic 2	0.72	1.26	0.37	0.77	1.73	0.29
topic 3	3.60	4.13	5.60	6.55	10.18	6.91
topic 4	3.57	3.90	3.64	4.02	5.78	4.49
topic 5	-5.62	-2.54	-1.01	-0.89	-1.24	-0.84
topic 6	2.27	3.09	2.77	3.52	5.50	2.87
topic 7	0.00	-0.41	-2.17	-2.52	-1.21	-0.49
topic 8	-8.90	-12.09	-6.72	-7.48	-11.97	-6.80
topic 9	-3.48	-6.73	-5.06	-6.08	-6.23	-3.95
topic 10	1.80	2.77	1.11	1.35	3.64	1.44
topic 11	1.58	1.49	2.37	1.93	4.09	2.23
topic 12	-4.97	-7.68	-6.39	-7.74	-10.61	-6.90
topic 13	-3.53	-3.68	-6.08	-7.09	-4.52	-5.79
topic 14	2.44	3.73	3.42	3.27	6.19	3.20
topic 15	0.00	0.77	-0.11	0.04	1.36	0.00
topic 16	2.73	3.26	2.15	1.60	3.79	1.92
topic 17	1.01	1.45	1.36	1.91	3.59	1.39
topic 18	5.54	6.36	6.47	6.23	9.38	6.18
topic 19	-5.06	-5.53	-11.54	-6.07	-6.37	-5.53
topic 20	0.50	0.84	0.62	0.00	1.72	0.53
topic 21	-1.50	-1.70	-2.42	-2.24	-1.98	-1.57
topic 22	3.68	4.05	3.36	3.36	4.91	3.88

topic 23	0.59	1.26	0.58	0.97	1.48	0.72
topic 24	-0.05	-2.24	0.36	0.50	0.00	0.00
topic 25	6.44	8.70	8.82	9.61	13.40	8.76
topic 26	-1.89	-1.87	-2.81	-2.88	-1.23	-2.61
topic 27	-0.28	-1.88	-1.39	-0.79	-0.17	-1.04
topic 28	-1.57	-1.37	-3.03	-0.57	-0.41	-0.92
topic 29	-0.34	0.00	-1.16	-1.85	-0.27	-0.66
topic 30	0.75	0.85	0.00	0.10	1.48	0.35
topic 31	-2.46	-4.29	-3.35	-3.56	-4.60	-2.87
topic 32	-2.59	-2.56	-6.16	-5.71	-4.36	-4.29
topic 33	-3.22	-4.41	-5.91	-7.45	-6.10	-4.95
topic 34	3.26	2.19	2.65	3.02	4.31	3.10
topic 35	0.48	0.80	1.17	1.36	2.18	1.21
topic 36	-2.98	-6.40	-2.91	-3.86	-5.21	-3.30
topic 37	-4.08	-4.04	-5.18	-5.31	-3.90	-5.18
topic 38	4.25	6.47	4.73	5.29	8.82	4.92
topic 39	-6.67	-2.11	-1.71	-1.55	-1.82	-4.09
topic 40	1.95	2.08	1.14	1.07	2.88	1.55

Table C8: 20-topic STM – Top predictors for lasso models where star ratings are the dependent variable.

	PHONE ACCESS EASE		APPOINTMENT EASE		GIVEN DIGNITY AND RESPECT		INVOLVED IN CARE DECISIONS		LIKELY TO RECOMMEND		UP-TO-DATE GP INFORMATION	
Topic	Model 1	rank	Model 2	rank	Model 3	rank	Model 4	rank	Model 5	rank	Model 6	rank
18. no appointments	-7.98	1	-7.78	2	-6.60	3	-7.81	2	-8.21	3	-7.70	1
15. great	5.22	3	8.42	1	6.37	4	7.07	3	10.4	1	6.21	3
14. discourage registration	-3.96	5	-4.54	5	-7.32	2	-7.86	1	-7.29	4	-6.80	2
4. poor management	-5.50	2	-7.14	3	-4.58	6	-5.45	5	-8.54	2	-5.51	4
20. rude reception	-4.99	4	-4.60	4	-10.6	1	-4.66	6	-5.56	6	-4.95	6
2. not enough time	3.69	6	3.60	6	5.29	5	5.56	4	7.04	5	5.32	5
6. comparisons	2.03	8	3.21	7	3.24	8	3.22	8	5.54	7	2.89	7
9. thanks	1.86	9	3.14	8	3.32	7	3.88	7	4.97	8	2.63	8
8. helpful	1.55	13	2.28	10	2.25	9	1.78	10	2.71	9	1.51	11
3. proper treatment	1.48	15	2.01	12	1.51	13	1.39	11	2.65	10	1.88	9
13. problem prescription	0.68	16	1.73	13	1.82	10	1.83	9	1.88	12	0.78	14
11. unwelcoming	1.86	10	1.50	15	1.59	12	0.74	13	0.84	15	1.74	10

17. hard appointments	-1.86	11	-2.65	9	0.70	15	0.78	12	-1.18	14	0.09	17
12. poor phone access	-2.63	7	-1.56	14	-0.77	14	-0.33	19	-0.77	16	-1.20	12
1. time expressions	0	17	-2.02	11	-0.34	16	-0.67	14	-1.89	11	-0.68	16
16. lack manners	-1.75	12	-1.22	16	-1.79	11	-0.42	16	-0.73	17	-0.72	15
19. late appointments	-1.52	14	-1.22	17	-0.13	17	-0.35	17	-1.35	13	-0.79	13
5. diagnosed and sorted	0	17	-0.05	19	-0.03	18	-0.67	15	-0.04	19	0	19
10. care is unprofessional	0	17	0.53	18	0	19	-0.34	18	0.12	18	-0.02	18
7. recommend	0	17	0	20	0	19	0	20	0.03	20	0	19

Notes: Predictors for each model are ranked by how different their coefficients are from 0. Magnitudes of topics from 0 correspond to how important each topic is for predicting the dependent variables. Topics with 0 as coefficient value are not statistically significant predictors

Overall, we identify that some valuable information is lost when a topic model is calculated with a smaller number of topics. This is particularly true for relatively less discussed subjects which nonetheless may be important to an understanding of service user satisfaction. There is no single best model with STM but definitely those with 5 and 10 models have much higher cross-validation errors than the rest. A model with more topics gives insight into more detail but, at the same time, some popular topics are represented multiple times which clouds interpretability of model outcomes.

Appendix D: Sentiment analysis of topics

Sentiment models have been computed to predict topics' sentiments. First, reviews were broken into sentence-length segments. For each sentence, the most likely topic was predicted and each topic was annotated with a star rating associated with the original review. 1* and 2* ratings were classed as negative sentiment labels of (31% of all sentences), 3* ratings were classed as neutral sentiment labels (15%). 4* and 5* ratings were classed as positive sentiment labels (54%). The sentences were tokenized using *spacy* v2.0.11 library in Python programming language. Multinomial Naïve Bayes model was trained on 51,855 tokens which occurred in at least 500 sentences to predict star ratings. Model's 50-fold cross-validation F1 score was about 0.96. Then, for each sentence, probabilities of each sentiment outcome were paired with the dominant topic. Sentiments probabilities were summed for all sentences. Then, a weighted sum of each sentiment corresponding to each topic was computed to compensate for unequal distribution of sentiments across the dataset. The highest weighted sentiment score was taken as the topic's sentiment. For example, if topic 1 was dominant in 10 sentences, for which the unweighted sentiments summed to 3 for neutral sentiment, 2 positive and 5 for negative sentiment, it's weighted score would be 3/0.15 for neutral, 2/0.31 for positive and 5/0.54 for negative. The highest score would indicate that topic 1 is first of all neutral. Table D1 lists the sentiment scores for each topic from the 20-topic STM model.

Table D1: Sentiment assignments to topics

Topic	Negative	Neutral	Positive
1	33868.67	42147.49	29440.38

2	114816.8	142182.8	119460.5
3	8548.746	36047.2	35972.07
4	15152.54	15870.35	8241.762
5	6167.633	21861.1	9433.851
6	23068.04	25613.02	46755.97
7	18835.67	27153.5	165996.1
8	42986.04	38066.94	212254.3
9	8968.752	17191.74	113527.7
10	19607.84	37060.49	17097.02
11	46704.71	72491.5	51452.69
12	224002.3	44920.36	30415.72
13	16409.33	59024.29	12284.58
14	18228.28	12152.66	5809.827
15	21417.1	26941.26	82742.74
16	51137.4	43681.13	80602.85
17	149250.9	53697.66	30584.71
18	173013.6	77729.13	63241.34
19	135309.9	73764.96	53041.53
20	107612.1	54861.58	42343.96

Notes: The most likely sentiment (highest weighted score) was used to determine whether a topic is positive, neutral or negative. Scores were weighted to compensate for unequal distribution of positive (4 or 5*), neutral (3*) and negative (1* or 2*) star ratings across dataset.*

Appendix E: Random Forest model quality

Calculating the average of averages that we use in the main paper: precision 0.39; recall 0.42; F1 0.36. The overall number of reviews is 208,282. At the disaggregate level, precision, recall and F1 scores for predicting the level of user satisfaction (number of review stars) is provided for each dimension of satisfaction (see Tables E1-E6 below).

Table E1: Precision, recall and F1 score of random forest model with ease of phone access star ratings as dependent variable

phone access ease			
	precision	recall	f1score
1 star	0.635	0.409	0.544
2 star	0.175	0.278	0.256
3 star	0.184	0.269	0.266
4 star	0.092	0.283	0.154
5 star	0.878	0.620	0.618

Table E2: Precision, recall and F1 score of random forest model with dignity and respect star ratings as dependent variable

given dignity & respect			
	precision	recall	f1score
1 star	0.758	0.472	0.685
2 star	0.031	0.216	0.059
3 star	0.243	0.291	0.349
4 star	0.102	0.300	0.176

5 star	0.927	0.809	0.746

Table E3: Precision, recall and F1 score of random forest model with likely to recommend star ratings as dependent variable

likely to recommend			
	precision	recall	f1score
1 star	0.939	0.723	0.846
2 star	0.002	0.333	0.003
3 star	0.003	0.436	0.005
4 star	0.004	0.412	0.009
5 star	0.938	0.816	0.845

Table E4: Precision, recall and F1 score of random forest model with appointment ease star ratings as dependent variable

appointment ease			
	precision	recall	f1score
1 star	0.919	0.581	0.678
2 star	0.043	0.269	0.080
3 star	0.028	0.260	0.053
4 star	0.134	0.351	0.214
5 star	0.834	0.540	0.654

Table E5: Precision, recall and F1 score of random forest model with involvement in care decisions star ratings as dependent variable

involved in care decisions			
	precision	recall	f1score
1 star	0.820	0.446	0.703
2 star	0.010	0.229	0.020
3 star	0.085	0.254	0.150
4 star	0.066	0.265	0.120
5 star	0.919	0.779	0.737

Table E6: Precision, recall and F1 score of random forest model with up-to-date GP information star ratings as dependent variable

up-to-date GP information			
	precision	recall	f1score
1 star	0.771	0.402	0.676
2 star	0.010	0.248	0.021
3 star	0.064	0.225	0.116
4 star	0.117	0.272	0.196
5 star	0.910	0.784	0.725

Random Forest model accuracies when predicting each of the dependent variable dimensions is reported in Table E7.

Table E7: Random Forest model accuracy for each of the dependent variable dimensions

	accuracy
phone access ease	0.476
appointment ease	0.537
given dignity & respect	0.624
involved in care decisions	0.616
likely to recommend	0.769
up-to-date GP information	0.602

Confusion matrices (rows - star predictions, columns - star values) for Random Forest models are also provided (see Tables E8-E13). Matrix diagonals contain counts of correct predictions.

Table E8: Random Forest confusion matrix for phone access ease

Note: Rows contain distribution of star ratings and columns contain distribution of star rating predictions. Matrix diagonal contains counts of correct predictions.

phone access ease					
	1	2	3	4	5
1	21763	4857	4191	1550	1903
2	14077	4920	4636	1904	2505
3	9888	4199	5533	2886	7500
4	5440	2675	4218	3779	24874
5	2055	1055	1968	3254	60080

Table E9: Random Forest confusion matrix for appointment ease

Note: Rows contain distribution of star ratings and columns contain distribution of star rating predictions. Matrix diagonal contains counts of correct predictions.

appointment ease					
	1	2	3	4	5
1	56111	1238	466	1042	2215
2	21177	1140	473	1103	2314
3	10371	866	613	2294	7689
4	5535	636	461	5141	26618
5	3286	363	343	5083	45634

Table E10: Random Forest confusion matrix for given dignity & respect

Note: Rows contain distribution of star ratings and columns contain distribution of star rating predictions. Matrix diagonal contains counts of correct predictions.

given dignity & respect					
	1	2	3	4	5
1	29462	916	4213	978	3278
2	13310	659	3812	1009	2521

3	11408	753	6028	2155	4503
4	5122	448	4428	2350	10618
5	3138	269	2216	1336	88501

Table E11: Random Forest confusion matrix for involved in care decisions

Note: Rows contain distribution of star ratings and columns contain distribution of star rating predictions. Matrix diagonal contains counts of correct predictions.

involved in care decisions					
	1	2	3	4	5
1	33141	223	1735	1173	4147
2	12980	183	1078	844	2501
3	13585	190	1837	1510	4464
4	9334	120	1663	1703	12879
5	5223	83	919	1186	84571

Table E12: Random Forest confusion matrix for likely to recommend

Note: Rows contain distribution of star ratings and columns contain distribution of star rating predictions. Matrix diagonal contains counts of correct predictions.

likely to recommend					
	1	2	3	4	5
1	72559	19	11	15	4646
2	11799	22	3	12	1404
3	6818	6	24	20	2218
4	3788	8	7	63	10185
5	5394	11	10	43	82105

Table E13: Random Forest confusion matrix for up-to-date GP information

Note: Rows contain distribution of star ratings and columns contain distribution of star rating predictions. Matrix diagonal contains counts of correct predictions.

up-to-date GP information					
	1	2	3	4	5
1	27068	164	1572	2540	3781
2	10465	163	909	1818	2261
3	12927	134	1389	2820	4404
4	11290	128	1523	3503	13455
5	5540	68	790	2194	86964