Driving responsive and timely improvements in patient experience feedback using Artificial Intelligence

Mustafa Khanbhai, Catalina Carenzo, Sarindi Aryasinghe, Dave Manton, Erik Mayer

Submitted to: Journal of Medical Internet Research
on: May 24, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Driving responsive and timely improvements in patient experience feedback using Artificial Intelligence

Mustafa Khanbhai[1] MD, PhD; Catalina Carenzo[1] MSc; Sarindi Aryasinghe[1] BSc; Dave Manton[2] BSc; Erik Mayer[1] MD, PhD

[1]Imperial Clinical Analytics, Research and Evaluation (iCARE), Imperial College London London GB
[2]NIHR Imperial Biomedical Research Centre Imperial College London London GB

Corresponding Author:
Mustafa Khanbhai MD, PhD
Imperial Clinical Analytics, Research and Evaluation (iCARE),
Imperial College London
The Bays, Entrance 2
South Wharf Road, St Mary's Hospital
London
GB

Abstract

Background: Understanding and improving patient care is pivotal for healthcare providers. With increasing volumes of the Friends and Family Test (FFT) data in England, manual analysis of this patient feedback poses challenges for many healthcare organisations. This underscores the importance of automated text analysis, particularly in predicting sentiments and themes in real-time.

Objective: Leveraging machine learning and natural language processing, this study explores the utility of a supervised algorithm to systematically test and refine the algorithm's cross-contextual performance in diverse healthcare settings, addressing variations in population characteristics, geographical locations, and care settings, ultimately driving improvements based on patient feedback.

Methods: The text analytics algorithm initially developed in a large acute Trust in London was further tested in nine healthcare organisations with diverse care settings across England. These Trusts varied in technical capacity and resource, population demographics, and FFT free text datasets. Testing and validation of the algorithm was performed including manual coding of subset of retrospective comments. Technical infrastructure was optimised including coding environments and packages for algorithm testing and deployment. The algorithm was iteratively trained using bag of words from anonymised data, tailored to accommodate contextual variations, and tested for change in algorithm performance whilst simultaneously rectifying issues identified.

Results: The algorithm demonstrated satisfactory overall accuracy (>75%) in predicting themes and sentiments embedded within free-text responses across a variety of care settings and population demographics. While the algorithm yielded strong and reusable models in relatively stable environments, such as adult inpatient care settings, the initial accuracy was notably lower in organizations providing services such as paediatrics and mental health. However, the accuracy of our algorithm significantly improved when individual Trust coding templates were applied. Thematic saturation was reached after the fifth organisation was recruited, and no further coding was required for the last four organisations. Subsequently, a framework and pipeline for deployment of the algorithm were developed to provide standardised approach for implementation and analysis of FFT free text, ensuring ease of use.

Conclusions: This study represents a significant step forward in leveraging free-text FFT data for valuable insights in diverse healthcare settings through the testing and development of a robust supervised learning text analytics algorithm. The disparity in some care settings was anticipated, given that the lexicon and phraseology used was inherently different from those prevalent in adult inpatient care (where the algorithm was developed). However, these challenges were addressed with further coding and testing. This approach enhanced the accuracy and reliability of the algorithm, encouraged inter- and intra-organisational collaboration, and shared learning.

(JMIR Preprints 24/05/2024:60900)

DOI: https://doi.org/10.2196/preprints.60900

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Driving responsive and timely improvements in patient experience feedback using Artificial Intelligence

Mustafa Khanbhai[1 2], Catalina Carenzo[1 2], Sarindi Aryasinghe[1 2], Dave Manton[2], Erik Mayer[1 2]

[1] Imperial Clinical Analytics, Research and Evaluation (iCARE), Department of Surgery & Cancer, Imperial College London, UK

[2] NIHR Imperial Biomedical Research Centre, UK

Abstract

Background

Understanding and improving patient care is pivotal for healthcare providers. With increasing volumes of the Friends and Family Test (FFT) data in England, manual analysis of this patient feedback poses challenges for many healthcare organisations. This underscores the importance of automated text analysis, particularly in predicting sentiments and themes in real-time. Leveraging machine learning and natural language processing, this study explores the utility of a supervised algorithm to systematically test and refine the algorithm's cross-contextual performance in diverse healthcare settings, addressing variations in population characteristics, geographical locations, and care settings, ultimately driving improvements based on patient feedback.

Methods

The text analytics algorithm initially developed in a large acute Trust in London was further tested in nine healthcare organisations with diverse care settings across England. These Trusts varied in technical capacity and resource, population demographics, and FFT free text datasets. Testing and validation of the algorithm was performed including manual coding of subset of retrospective comments. Technical infrastructure was optimised including coding environments and packages for algorithm testing and deployment. The algorithm was iteratively trained using bag of words from anonymised data, tailored to accommodate contextual variations, and tested for change in algorithm performance whilst simultaneously rectifying issues identified.

Results

The algorithm demonstrated satisfactory overall accuracy (>75%) in predicting themes and sentiments embedded within free-text responses across a variety of care settings and population demographics. While the algorithm yielded strong and reusable models in relatively stable environments, such as adult inpatient care settings, the initial accuracy was notably lower in organizations providing services such as paediatrics and mental health. However, the accuracy of our algorithm significantly improved when individual Trust coding templates were applied. Thematic saturation was reached after the fifth organisation was recruited, and no further coding was required for the last four organisations. Subsequently,

a framework and pipeline for deployment of the algorithm were developed to provide standardised approach for implementation and analysis of FFT free text, ensuring ease of use.

Conclusion

This study represents a significant step forward in leveraging free-text FFT data for valuable insights in diverse healthcare settings through the testing and development of a robust supervised learning text analytics algorithm. The disparity in some care settings was anticipated, given that the lexicon and phraseology used was inherently different from those prevalent in adult inpatient care (where the algorithm was developed). However, these challenges were addressed with further coding and testing. This approach enhanced the accuracy and reliability of the algorithm, encouraged inter- and intra-organisational collaboration, and shared learning.

Background

Insights into patients experience of their care provision is critical for healthcare providers delivering patient-centred care and recognised as integral to ensuring safe and high-quality care.(1) By the end of 2019, the national survey for patient experiences in the United Kingdom (UK), known as the Friends and Family Test (FFT), had accumulated 75 million pieces of feedback.(2) This number continues to grow by about 1.3 million every month, making it the largest source of patient opinions globally. However, manually analysing this vast amount of data requires a considerable resource, something that many healthcare organisations struggle to provide.(3, 4) In recent years there has been an ever-clearer desire for improvement to arise from patient feedback rather than just focusing on response rates, but most healthcare organisations are constrained from doing so due to lack of analytical expertise.(5)

Automated text-analysis by Artificial Intelligence (AI) and its component Natural Language Processing (NLP) predicts sentiments and themes within free-text responses in real-time and provides a useful opportunity for better integrating patient experience feedback into everyday healthcare delivery.(4) NLP algorithms, offer a practical solution to decode the 'why' behind patient narratives and through their ability to process large volumes of data accurately these patient-derived insights can inform ‘business as usual’ within healthcare organisations.(6, 7) We have previously demonstrated the benefit of applying NLP to free-text FFT data to deliver patient driven quality improvement in a single institution.(8)

Therefore, the objective of this study was to assess the acceptable accuracy threshold of the previously developed algorithm, and comprehensively appraise its performance within a real-world context, encompassing linguistic nuances, spelling errors, varying patient demographics and across diverse healthcare settings, including adult inpatient, primary care, community, paediatrics, and mental health care. The overarching aim was to provide near real-time interpretation of patient experience free-text comments, enabling healthcare providers to embed user insights into a culture of organisational patient-centred delivery.

Methods

The implementation of the FFT free-text algorithm demands thorough testing to ensure robustness, accuracy, and applicability. To iterate the algorithm for deployment in other healthcare services, we first selected healthcare organisations that would incorporate geographical variables, patient demographic variables, and different care settings, i.e., adult inpatient, primary care, community, paediatrics, and mental health.

The objectives of this study were to identify technical capabilities for algorithm deployment in the included healthcare settings, identify variation in user needs across different service settings and geographies, and iterate the algorithmic process and deploy a reproducible algorithm that can be implemented in different services.

Trust recruitment

A steering group consisting of lay representatives and patient experience experts was formed to guide the strategic direction of this study and help in recruitment through various patient experience networks including the National Insight Network set up by NHS England, the Q Community established by the Health Foundation, Advancing Quality Alliance, and the Heads of Patient Experience Network. A FFT review capability questionnaire (Appendix 1) was sent to interested organisations. This allowed the study team to identify organisations based on three key metrics, patient experience engagement, digital readiness/information technology infrastructure, and quality improvement involvement. This study received Health Research Authority and Health and Care Research Wales approval, 20/HRA/5924. To ensure that the selected Trusts possess the necessary infrastructural and organisational qualities to contribute effectively to the research objectives while adhering to data regulatory standards, a capacity and capability assessment was required. Only those Trusts that were able to confirm capacity and capability through their Research and Development department were finally recruited, as shown in Figure 1.

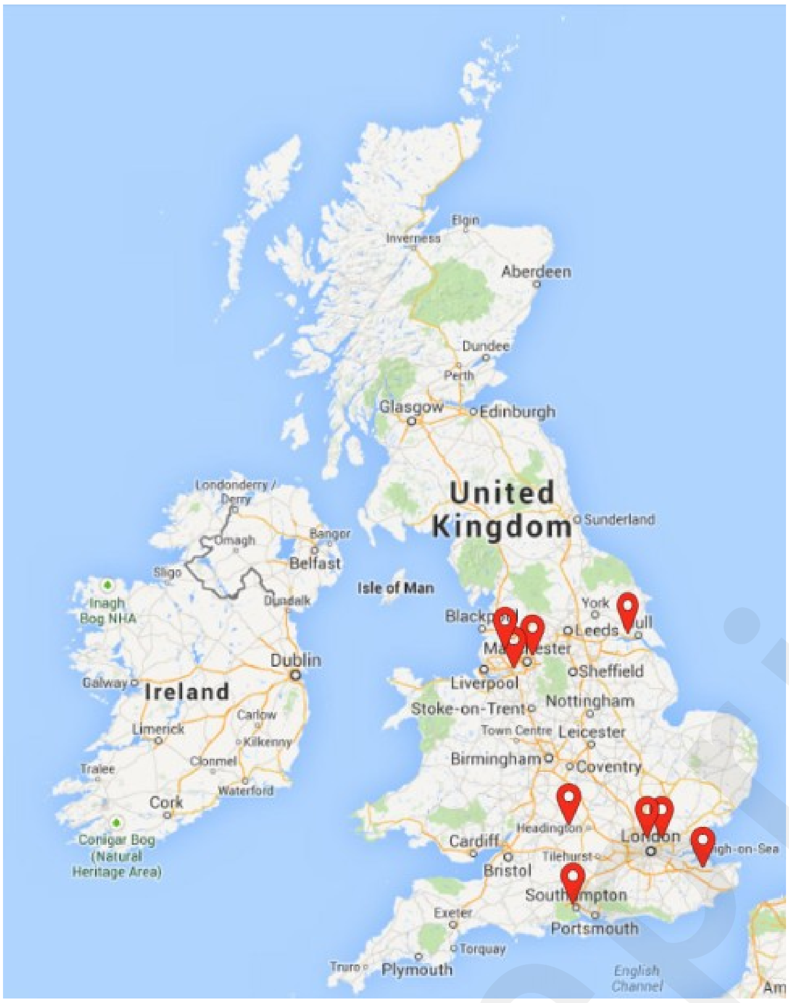


Figure 1. An overview of the geographical distribution of the nine recruited Trusts.

Pre-implementation

IT scoping tool

To work collaboratively, we identified the needs of all participating organisations from a patient experience digital readiness standpoint. This was facilitated through previous work by the study team funded by NHS England Insight and Feedback Team (9) where a scoping tool was developed to identify key contacts, FFT data collection, structured query language (SQL) Server database, handling and management, infrastructure capacity i.e., on-premise data centre or cloud computing, integrated development environment and coding language expertise, and data visualisation software (Appendix 2).

Data access

To ensure that participating organisations were fully supported by the project team a Data Protection Impact Assessment (DPIA) was completed with each organisation. DPIA is a systematic process designed to identify and minimize the associated data protection risks. The purpose of a DPIA in our study was to assess the impact of processing data and to ensure that the principles of data protection, such as those outlined in regulations like the General Data Protection Regulation were adhered to.

Adaptation of coding script

To achieve transferability of the support vector machine (SVM) supervised machine learning (ML) algorithm and test against different service settings and NHS provider organisations, the functionalities previously developed (6) were re-coded using Python (Python Software Foundation) given the popularity of using this coding platform in most NHS Trusts. Python is an interpreted high-level general-purpose programming language. The model was developed with Python 3.8.10 with the following free Python libraries; scikit-learn>=0.22.1 (24) (a machine learning library that provides relevant machine learning packages required for classification), Natural Language ToolKit (NLTK) >=3.4.4 (25) (a suite that contains libraries and programs to make machines understand human language), and Pandas >=0.25.3. (26) (a library for data manipulation and analysis). These requirements were then translated into a requirements.txt file for organisations to deploy the algorithm.

Updating training coding template

All FFT free-text comments were coded based on the NHS Patient Experience Framework (Appendix 3). This framework incorporates eight key themes that outline those elements which are critical to the patients' experience of NHS services.(10) In the previous work (7), three additional themes were added; ‘Unclassified’ (too many typographical errors to discern meaning), ‘General’ (e.g., NHS is great, everything), and ‘Staff’ (comments relating to staff). From the NHS England funded work (9) it was observed there was an overlap of comments between theme ‘Staff’ and other themes such as ‘Respect for patient-centred values, preferences, and expressed needs’, ‘Emotional support’, and ‘Physical comfort’. This is because often a staff member was involved in the delivery of the care as it related to the other themes. The theme ‘Staff’ was removed, and all comments coded as ‘Staff’ were recoded to develop a concise coding template. Therefore, 10 themes were selected for final analysis and a coding pack (Appendix 4) was developed with detailed guidance. Furthermore, we iterated the initial algorithm from our previous work which was developed using two follow-up free-text questions, i.e., ‘What did we do well?’ and ‘What could we do better?’, by combining and shuffling the training data from both questions to develop a streamlined SVM supervised machine learning algorithm for theme and sentiment prediction.

Implementation

A pipeline was constructed for the SVM model in Python for text classification and sentiment analysis, consisting of several steps, including Text Pre-processing, Feature Engineering, term frequency-inverse document frequency (TFIDF) transformation, and model training and testing.

FFT data retrieval

SQL Server was used to both extract the FFT data and supplementary free text for algorithm imputation, and to store the output of the algorithm. Pyodbc >= 5.0.1 (27) Python library was required to connect to the individual organisations SQL server. SQL’s combination of ease of use, flexibility and capabilities make it useful for storing, managing, querying, and analysing datasets and is widely used in healthcare organisations.

Establishing ground truth

In supervised learning algorithms, ground truth data is critical to training and updating an algorithm.(11) Therefore, more annotated data was required to improve the performance of the algorithm. The coding pack (Appendix 4) developed was provided to the participating organisations patient experience teams to make coding more feasible. The coders were a combination of patient experience team, clinical staff and Trust lay representatives. Two coders from the team completed manual coding individually using 500 stratified sample to calculate inter-annotator reliability and understand the interpretation of themes and sentiment. Disagreements in coding were reviewed, and theme and sentiment were finalised to create a master coding template in .csv file. This template was populated with data from all nine organisations.

Pre-processing

Pre-processing of textual data is the first essential step in the processing of text and has been proven to improve text classification models' performance by standardising the text before it is presented to the classification algorithm.(11) NLTK for Python was used for pre-processing and data analysis which includes tokenization of the comments and removal of stop words (e.g., the, was, is, etc.), punctuation, numbers, and special characters from the comments. The resulting words were represented as a bag of words (BoW) or corpus. Given that the average length of FFT free-text comments tends to be short, and the context is domain specific, i.e. patient experience, this bag or words approach was used rather than Word Embedding.

Feature engineering

Feature engineering is the process of transforming data into features/attributes that better represent the underlying structure of the data.(11) To identify the key features of the data, we used a statistical measure TFIDF (Term Frequency Inverse Document Frequency), used in the fields of information retrieval (IR) and machine learning, that can quantify the importance or relevance of words in a comment amongst a collection of comments.(12)

Determining accuracy

To understand if further coding was required, the accuracy for theme and sentiment was calculated using 500 stratified sample of manually coded comments from each of the individual organisations. The accuracy is defined as a metric used in classification to measure the percentage of accurate predictions.(11) It is also defined as the ratio of number of correct predictions by the algorithm to the total number of predictions. In other words, ratio of true positives (TP) and true negatives (TN) to all positive and negative observations (where FP and FN are false positives and false negatives respectively): Accuracy = (TP+TN)/(TP+TN+FP+FN). The acceptable threshold for overall accuracy for theme and sentiment for a supervised algorithm using patient feedback was determined at 75%.(11)

The accuracies yielded by the models on different ranges of hyperparameters were compared. For SVM, both L1 and L2 regularization, variance tolerance values for the stopping criteria, and different values for the inverse of regularization strength, C, were compared. As regularization has a known role in reducing overfitting, the hyperparameters were chosen for tuning to identify the amount and type of regularization and early stopping criteria that are suitable for the classification tasks. After this, the optimal hyperparameters found by GridSearchCV from the scikit-learn library. A 10-fold cross-validation was used to conduct a search to select the best hyperparameter combinations from specified ranges of hyperparameters, which were used to re-train the model on the training set and to predict on the test set.(11) This allowed us to establish the accuracy but also use the coded comments as a template to improve the accuracy by a further round of re-coding if needed until the accuracy was at or above 75%, i.e., reached thematic saturation.

Algorithm deployment

The master coded template data was then serialised using the Python pickle library (28). This module is used for serializing and de-serializing objects into Python format text strings, which are called pickles and can be created by Python code or sent from other Python programs. This Python pickle is easily carried over the network and can be validated without knowing the details of the received object. This enables the preservation and sharing of the ML models, allowing users to reload pre-trained models, significantly reducing the need for lengthy re-training. The files that were stored in the environment were as follows: Feature_sentiment.pkl, Feature_theme.pkl, Sentiment_classifier.pkl, Theme_classifier.pkl, Tfidftransformer_sentiment.pkl , Tfidftransformer_theme.pkl.

Establishment of a Community of Practice

In the pursuit of ensuring the correct sustainability of the algorithm, a strategic initiative was undertaken through the establishment of a Technical Community of Practice (CoP). This community served as a conduit for regular interaction and collaboration among the leaders responsible for overseeing the algorithmic maintenance across different Trusts. Recognising the complexity and potential challenges associated with deploying advanced algorithms, quarterly meetings were instituted as a core component of the CoP framework. These meetings provided a structured way for the leads of each site to convene, fostering an environment conducive to the exchange of critical insights, best practices, and lessons learned.

Results

Nine Trusts were recruited from a range of care settings, sociodemographic and geographical areas. Trusts were added to the study sequentially, one by one, as part of the recruitment process. None of these organisations had a pre-existing automated process for analysing their FFT free-text responses.

Identifying free-text comments with FFT survey for analysis

In the pre-implementation scoping exercise, it was observed that among the participating organizations, Trust A and D incorporated three follow-up free-text questions in their FFT survey. In contrast, all other Trusts included only one follow-up question (Table 1). The question with the highest number of comments in Trust A was analysed. The organisations provided a variety of services, including paediatrics, inpatient and outpatient, and community and mental health thereby facilitating testing and iteration of the algorithm for use in all care settings. Furthermore, most of the organisations (8/9) opted to house and deploy the algorithm on premise and the rest on Cloud (1/9).

Table 1. An overview of the variation in FFT free-text data, service setting and environment favoured to deploy the machine learning algorithm in nine NHS organisations in England. Version deployed refers to the version of the iterative addition of words to the BoW model, used to improve model performance.

* The free text from this question was used in the algorithm testing given the highest responses compare to the other two questions.

Trust	**Service coverage**	**Model deployment environment**	**Number of free-text columns**	***Free-text Question(s)***	***Version deployed***
A	**Community and Mental Health**	On Prem	3	*Please can you tell us why you gave your answer?**	*Version 0*
				Please tell us at least one thing that went well.	
				Please tell us at least one thing we could do better.	
B	**Acute and Inpatient**	Cloud	1	*Please can you tell us the main reason for the score that you have given?*	*Version 1*
C	**Acute and Inpatient**	On Prem	1	*Please can you tell us why you gave your answer and what we could have done better?*	*Version 2*
D	**Paediatrics**	On Prem	3	*FFT rating description**	*Version 3*
				Write what you think was good.	
				Write what you think was bad.	
E	**GP/Community**	On Prem	1	*Please can you tell us why you gave your answer?*	*Version 3*
F	**Acute and Inpatient**	On Prem	1	*Can you tell us why you gave that response?*	*Version 3*
G	**Acute and Inpatient**	On Prem	1	*Please can you tell us why you gave your answer and what we could have done better?*	*Version 3*
H	**Acute and Inpatient**	On Prem	1	*What was good about your care, and what could be improved?*	*Version 3*
I	**Acute and Inpatient**	On Prem	1	*Please can you tell us why you gave your answer.*	*Version 3*

Variation in patient experience themes

To understand the variation in themes, if any, of FFT free-text responses in the nine organisations, we looked at the count of comments from the FFT free-text responses based on the NHS patient experience framework. This was extracted from the output after initial model deployment for the period of January 2022 to January 2023 as demonstrated in Table 2. The highest count of themes was 'respect for patient centred values' followed by 'general'. The themes with the lowest count were 'transition and continuity' followed by 'welcoming the involvement of friends and family'. Trusts with care setting such as community, mental health and paediatrics did not have demonstrable variation in count of themes compared to adult inpatient. There were no demonstratable trends in the other patient experience framework themes.

Table 2. Distribution of count of themes based on the NHS Patient Experience Framework from nine NHS organisations in England, extracted from the output of the algorithm.

Trust	**% of FFT comments provided per NHS Patient Experience Framework themes**									
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
A	9	26	4	10	6	6	0	2	24	12
B	8	35	2	14	6	5	2	2	11	16
C	6	23	3	13	8	10	1	2	8	26
D	15	17	9	10	6	5	0	2	16	20
E	22	25	2	11	5	5	0	1	18	12
F	2	43	1	12	4	7	1	1	8	21
G	6	23	3	13	8	10	1	2	8	25
H	7	31	3	7	13	5	2	1	7	25
I	1	25	6	8	9	11	2	3	11	22

0 Unclassified, 1 Respect for patient-centred values, 2 Coordination & integration of care, 3 Information and communication, 4 Physical comfort, 5 Emotional support, 6 Involvement of family and friends, 7 Transition and continuity, 8 Access to care, 9 General.

Establishing ground truth

Inter-rater agreement was determined on the manual coding, and the coding template was updated to check accuracy and improve the model by acquiring an updated corpus from the BoW and feature extraction. The inter-rater agreement was better for sentiment compared to theme. There was almost perfect agreement (0.81–1.00) seen in 7 organisations. Trust A was the only organisations with a substantial (0.61–0.80) agreement with a Cohen's kappa of 0.79 (Table 3).

Table 3. Assessing the consistency of theme and sentiment ratings, and final accuracy via updated coding template.

Trust		***Category***	***Interrater reliability / Cohen's Kappa***	***Accuracy before re-coding (%)***	***Final accuracy (%)***
A	**Community and Mental Health**	Theme	0.79	57	76
		Sentiment	0.82	76	82
B	**Acute and Inpatient**	Theme	0.84	69	78
		Sentiment	1	87	91
C	**Acute and Inpatient**	Theme	0.85	84	90
		Sentiment	0.91	87	94
D	**Paediatrics**	Theme	0.89	67	82
		Sentiment	1	72	82
F	**GP/Community**	Theme	0.80	*64*	*77*
		Sentiment	0.82	*69*	*80*
E	**Acute and Inpatient**	Theme	0.96	*N/A*	*>75*
		Sentiment	1	*N/A*	*>75*
G	**Acute and Inpatient**	Theme	*0.83*	*N/A*	*>75*
		Sentiment	*0.85*	*N/A*	*>75*
H	**Acute and Inpatient**	Theme	*N/A*	*N/A*	*>75*
		Sentiment	*N/A*	*N/A*	*>75*
I	**Acute and Inpatient**	Theme	*N/A*	*N/A*	*>75*
		Sentiment	*N/A*	*N/A*	*>75*

Three Trusts noted that sentiment prediction related to certain words were misclassified. The study team reviewed the performance and identified words that were culpable for this misclassification, e.g., "Nothing", "Everything", and "Ok". These Trusts were advised to manually code a subset of 500 comments with equally distributed sentiments, especially focusing on the misclassified words, along with comments that represented an equal or almost equal number of comments per themes and/or sentiments to obtain a balanced dataset. To determine the accuracy of the algorithm, a pipeline was developed and refined to standardise the entire process within all the organisations as depicted in Figure 2.

The overall accuracy for theme and sentiment significantly improved after using a revised coded template. As Trusts were added to the study sequentially, one by one, as part of the recruitment process, the BoW were also applied sequentially to the algorithm. Thematic saturation was achieved after analysing and collating comments from a total of five Trusts.

This was observed when the algorithm was tested in four of the remaining Trusts without the need to implement their respective coding output and achieving an overall accuracy greater than or equal to 75%.

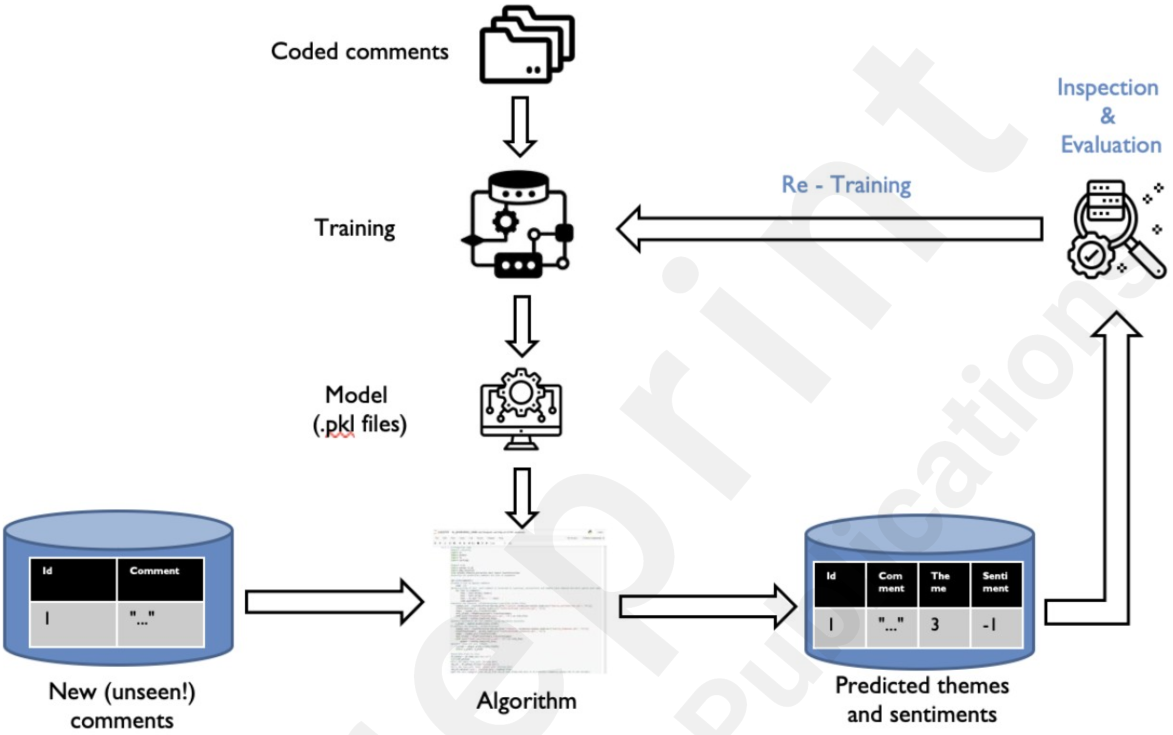


Figure 2. A pipeline created to develop and standardise overall accuracy calculation before final deployment.

The final four organisations achieved an overall accuracy rate of 75% or higher, having reached thematic saturation following the initial coding efforts of the initial five organisations. As a result, as seen in Table 1 four versions of the model were deployed as new, labelled comments were collated from the different Trusts that joined the programme and deployed the algorithm. This iterative learning was done by adding randomly scrambled new unique words that the algorithm encountered in the BoW within its code. Each time, this word repository was completely randomised, and each word was treated individually by removing any linkage. As a result, it was not possible to know from which Trust the word came from or if it was part of a positive or negative comment, with the study team holding the ground truth.

Adaptations and running without errors

The algorithm script was adapted according to agreements and requirements of the individual organisation. The solution tested in Jupyter Notebook for easier interaction with the code and checked for errors. The two most common errors encountered were “FileNotFoundError’ which was a result of not specifying the correct path to a .pkl file, “Invalid column name” as the code had not been adapted appropriately with the column names from the input table on the SQL server. Once errors were corrected the final deployment consisted of the output with predicted themes and sentiment exported to a SQL database for integrating into dashboard visualisation. This pipeline (Figure 3) was standardised after refinement with all participating organisations. Each Trust automated their dataflow based on the requirements defined by the frequency and volume of comments that the pipeline should be updated with.

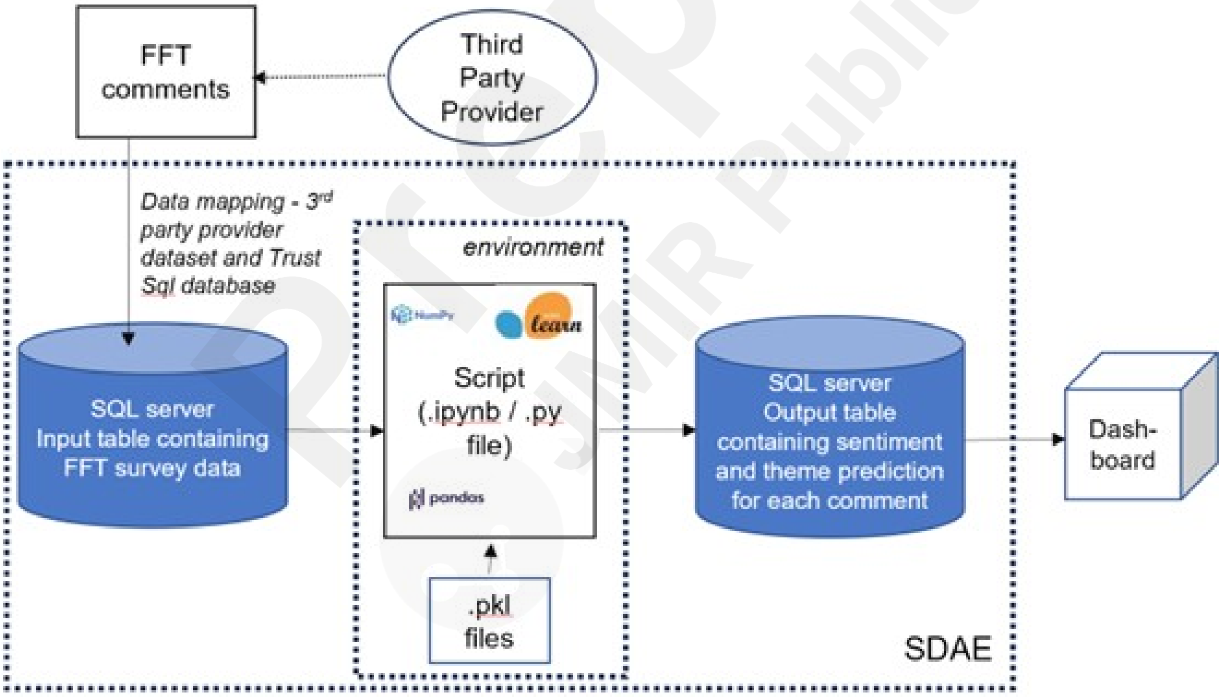

FFT comments
Third Party Provider
Data mapping - 3rd party provider dataset and Trust Sql database
environment
NumPy
learn
Script (.ipynb / .py file)
pandas
SQL server Input table containing FFT survey data
SQL server Output table containing sentiment and theme prediction for each comment
Dash-board
.pkl files
SDAE

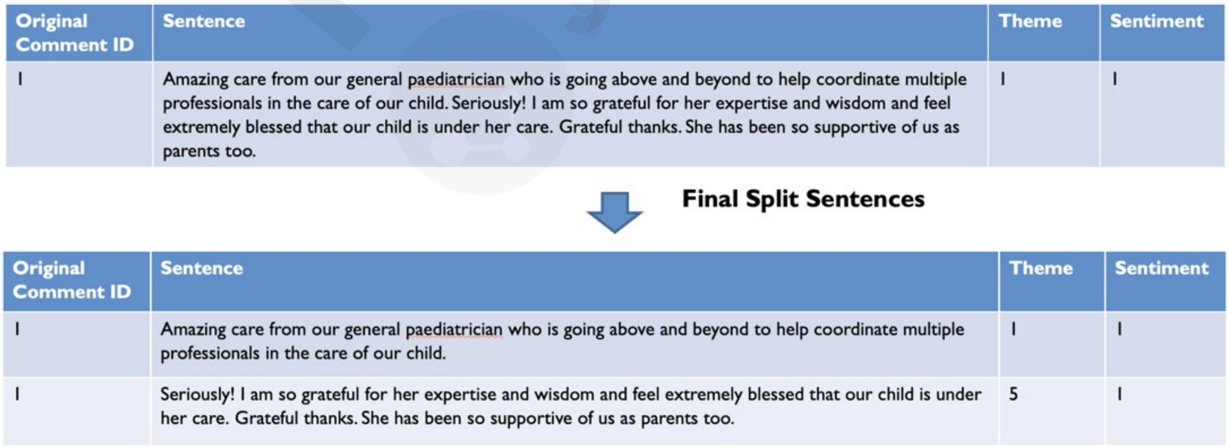
Figure 3. Final pipeline demonstrating steps required to deploy the final algorithm with FFT comments as input, prediction of themes and sentiment using Pickle files generated from summative coding outputted to SQL database for use in dashboard visualisation.

For this purpose, user friendly interfaces were designed and developed using Flask Python library (29) for the backed and HTML, Css and Bootstrap for the frontend. Functionalities of these interfaces included analysing the accuracy of the algorithm by uploading a labelled dataset and re-training the algorithm with more coded data.

Algorithm refinement

Multi-tagging

Potential refinements to the algorithm, particularly regarding the need to break long comments into sentences for Trusts encountering lengthier narratives, such as those in mental health and community settings were highlighted and a multi-tagging method was devised. The multi-tagging approach involved assigning multiple themes from the NHS Patient Experience Framework to a single text document based on its content. Unlike traditional single-label classification, where the FFT free-text is assigned to only one label, multi-tagging recognises that comments may encompass various themes. This approach enables a more nuanced representation of the content’s complexity and captures multiple dimensions of meaning within the same document. The punctuation marks, words and number of characters used to split the original sentence, can be customised based on different requirements using Spacy (>=3.7.2) and its English language model. As a default, sentences are split after a maximum of 75 characters, and each of these is analysed individually (Figure 4).



Original Comment ID	Sentence	Theme	Sentiment
1	Amazing care from our general paediatrician who is going above and beyond to help coordinate multiple professionals in the care of our child. Seriously! I am so grateful for her expertise and wisdom and feel extremely blessed that our child is under her care. Grateful thanks. She has been so supportive of us as parents too.	1	1

Final Split Sentences

Original Comment ID	Sentence	Theme	Sentiment
1	Amazing care from our general paediatrician who is going above and beyond to help coordinate multiple professionals in the care of our child.	1	1
1	Seriously! I am so grateful for her expertise and wisdom and feel extremely blessed that our child is under her care. Grateful thanks. She has been so supportive of us as parents too.	5	1

Figure 4: Example of how the sentencing approach works. The original comment is split into two sub-comments, each classified according to a theme and a sentiment. The original

comment identifier is retained to preserve the metadata of the comment.

Redaction of personal and identifiable information

Considerable variation was also observed in how Trusts approached the redaction process, ranging from automated processes with subsequent manual checks to entirely manual methods. To address this inconsistency, a redaction algorithm capable of automatically detecting and masking personally identifiable information was devised. A redaction algorithm was developed using NLTK, Regex and Presidio Python libraries to automatically detect and mask the personally identifiable information.

Community of Practice

During the quarterly meetings, there was a high level of engagement from at least 80% of staff members representing the nine Trusts. This level of participation indicated strong interest and commitment to the CoP's objectives. Additionally, feedback from participants revealed a positive perception of the CoP, with at least 90% reporting a sense of belonging and support within the community. This sense of camaraderie and mutual support contributed to the effectiveness of the CoP in addressing algorithm deployment challenges and driving improvements based on patient feedback.

Discussion

Using a test-and-iterate approach, this study deployed a text analytics algorithm to predict theme and sentiment from FFT free text data in nine NHS organisations in England with different care settings, geographical location and demographics. Furthermore, the study highlighted the variation in technical capacity and resources among Trusts, introduced and developed a framework and standardised data processing pipeline capable of supporting FFT text analytics bespoke to the organisation's needs. The key contributions of this study are: (1) the construction of technical infrastructure to support and run the algorithm effectively and being able to adapt the script for individual organisations; (2) the creation of master coding template with free-text data from different healthcare organisations to support algorithm iteration and improve overall accuracy for theme and sentiment; (3) the creation of pipeline to standardise the method of incorporating the algorithm for processing free-text data; and (4) foster collaborative non-siloed working that encourages local ownership and share-learning.

Addressing data management and standardisation through information technology infrastructure

Data collection and analysis is central to this project. Achieving this required a robust data infrastructure with different components that collaboratively created an effective framework for research data management. This framework enhanced data accessibility by presenting information in a consistent, standardised, predictable and accessible format while also strengthening data reliability through the automation of manual processes.(13) However, ensuring high-quality data demands a diverse set of information management skills, including data acquisition, storage, organisation, retrieval, and long-term maintenance.(14) This project offered a valuable opportunity for Trusts to gain expertise in this area by providing guidelines to streamline processes, define roles, and establish a resilient information technology infrastructure.

Creating a template with additional organisational specific free-text data

Each healthcare setting possesses unique attributes, patient demographics, and linguistic nuances as demonstrated above. Algorithm testing was required to allow for customization and tailoring of the algorithm to accommodate these contextual variations. As part of a supervised learning approach, re-coding was required to re-establish ground truth and new words were incorporated into an expanded BoW to improve performance across various care settings. Literature suggests that the larger the training sets used, the higher the accuracy of the algorithms at identifying similar comments within the broader dataset, but trade-offs with time and human coding are necessary to ensure the method is resource-efficient.(11) However, the nature of patient experience vocabulary, due to the domain of patient feedback from free-text is fixed in its nature, making it attractive data for supervised learning.(15) Therefore, it is possible to anticipate the meaning of various phrases and automatically classify the comments.(16) Thematic saturation was reached with five organisations and no further coding was necessary. Just as the domain is fixed, the perspective of a patient feedback document is also fixed; there is limited vocabulary that is useful for commenting about health services.(15) Rastegar-Mojarad et al. (17) also observed that a small (25%) vocabulary set covered a majority (92%) of the content of their patients' comments, consistent with a study exploring consumer health vocabulary used by consumers(18). This suggests that patients use specific vocabulary when expressing their experience within free-text comments.

In settings that were relatively fixed, i.e. adult inpatient, the classifier trained strong reusable models that was evident from satisfactory accuracy. However, the accuracy of the model was lower in organisations delivering different services such as paediatrics and mental health. This was to be expected given the language and type of words used in different care settings would vary from those found in an adult inpatient setting. The customization ensured that the algorithm aligns with the specific language and themes prevalent in the patient feedback of a particular healthcare organisation. The accuracy for theme improved in all organisations demonstrating that whilst the algorithm may not initially be transferable, with some retraining and adaptations it can easily be reused in different service settings.

Furthermore, 'respect for patient centred values' had the highest count in all organisations. This finding is similar to the literature around patient experience reporting in healthcare.(19-23) The reasons for this are twofold, firstly, primarily due to the large number of interchangeable words found within that theme, i.e., 'Respect for patient-centred values, preferences, and expressed needs (cultural issues, dignity, privacy and independence, awareness of quality-of-life issues and shared decision making)'. Secondly, the overall domain of patient feedback is the healthcare system, and this study revealed that the content of reviews tends to focus on a small collection of aspects associated with this, as demonstrated by the themes used for text classification in the studies.

Through CoP participation, multi-tagging was used to assign multiple relevant tags or categories to individual sentences within a single FFT free-text comment. Instead of assigning a single label to the entire document, this technique involved breaking down text into sentences and assigning appropriate themes and sentiments to each based on its content, thereby enhancing customisability within mental health and community care settings. Furthermore, the CoP assisted in developing a redaction solution using Python libraries which can be then customised according to the requirements of the Trust to define how and what data needs to be redacted. These solutions can subsequently be adopted by other Trusts as an improvement to their existing processes.

Creating a pipeline to standardise the algorithm deployment approach

As part of the model testing and to make the model more widely re-usable, we created a data pipeline, which infuses and integrates BoW from different Trusts to predict theme and sentiment at an accuracy greater than 75% when used in different Trusts. This pipeline serves as a blueprint to stakeholders and facilitates understanding of all pieces of the pipeline within the team members and standardisation when deploying the model. Compliance with regulations and governance is supported, and integration with DevOps practices enhances machine learning model integration into the software development lifecycle.

Generating cultural change to prioritise patient experience centred care

The establishment of CoP can play a pivotal role in establishing a multi-disciplinary team of stakeholders from IT, business intelligence, patient experience, and QI within each Trust. Furthermore, the CoP can provide the necessary guidance to use the algorithm and assist in developing a dashboard for quality improvement projects,(8) and sustained enthusiasm in the use of data and data analytics to improve the quality of care for patients and carers.(24) This cultural transformation will result in increased organisation agility and responsiveness to patient needs, driving positive changes in care delivery practices.(25)

Limitations

The Pickle files used in this study are specific to the version of Python and the libraries used to create them. Attempting to load a pickle file created with an older version of Python or with different library versions, may encounter compatibility issues. Newer versions of Python may have introduced changes to the pickle module, or the libraries being pickled, which can result in errors when loading the file. To mitigate challenges with loading an outdated pickle file, updating the code that creates the pickle file or alternative serialization formats that are more version-independent can be considered. The model can be containerized using Docker with a requirements file containing all the required libraries and Docker specifying the environment setup and start-up operations. The Docker Container can then be deployed to application hosting services, whether on cloud or on-premises servers.

The limited vocabulary range is a significant issue faced by the BoW model. For example, if the model encounters an unfamiliar or rare but informative word that it has not yet be seen during training, BoW model will tend to ignore it. To mitigate this limitation, we trained the model on five out of nine Trusts to ensure that the model has gathered a substantial BoW representation, further augmented by thematic saturation. Significant change to patient demographics or significant changes such as COVID-19 pandemic could reduce the performance due to unfamiliar words. Therefore, regular updates in the model would be required to keep it up to date. For this purpose, user friendly interfaces can be designed and developed using Flask Python library for the backend and HTML, Css and Bootstrap for the frontend. Functionalities of these interfaces included analysing the accuracy of the algorithm by uploading a labelled dataset and re-training the algorithm with more coded data. The interfaces allow organisations to periodically check accuracy as part of a responsible artificial intelligence strategy.

Responsible Artificial Intelligence

As Artificial Intelligence (AI) continues to advance, becoming increasingly integrated into healthcare systems, ethical considerations and technical challenges become increasingly intricate. For this reason, a key aspect of the project included providing tools and best practices to ensure that AI is being implemented responsibly. Firstly, the developed user-friendly interfaces, promote the ongoing monitoring of the accuracy of the algorithm, ensuring fairness and accountability of the version of the model that was deployed. Secondly, the establishment of a CoP, not only facilitates the dissemination of knowledge while contributing to a culture of continuous improvement, but also cultivates an environment where professionals can delve into the nuanced aspects of responsible AI. This contributes to the development and refinement of ethical guidelines, playing a crucial role in shaping the responsible AI landscape.

Conclusion

This study marks an advancement in harnessing free-text FFT data to gain valuable insights in healthcare settings through the creation of a robust supervised learning text analytics algorithm. The observed disparities in certain care settings were expected, considering the inherent differences in lexicon and terminology compared to the adult inpatient care setting where the algorithm was initially developed. Addressing these challenges involved additional coding and thorough testing under diverse scenarios. Through this iterative process, the accuracy and reliability of the algorithm were established to be robust and easy to use, fostering inter- and intra-organisational collaboration, and promoting shared learning within the healthcare domain.

Funding

This work was funded by the Health Foundation and supported by the National Institute for Health Research (NIHR) Imperial Biomedical Research Centre and the NIHR Imperial Patient Safety Research Centre. The research was enabled by the Imperial Clinical Analytics Research and Evaluation (iCARE) environment and used the iCARE team and data resources. Development of the IT scoping tool and coding pack was funded by NHS England Insight and Feedback Team. The study funder(s) did not play a role in study design; in the collection, analysis, and interpretation of data; in writing of the report; and in the decision to submit the article for publication. In addition, researchers were independent from funders, and all authors had full access to all the data included in this study and can take responsibility for the integrity of the data and accuracy of the data analysis.

Acknowledgments

The authors would like also to like to acknowledge the contributions of all individuals from the NHS Trusts involved in the Scale, Spread and Embed Programme:

Victor Banda, William Bryant, Suzanne Collin, Zeina Dahan, Paul Gough, Mark

Harris, Taraben Kapadia, Pavithra Rajendran, Jonny Sheldon, Caitlin McGovern, Marcus Beck, Ahmed Sharif, Mandy Dawley, Catherine Hunter, Brad Walker, Mark Ward, Luca Mercuri, Rohini Subramaniam, Ben Glampson, Amy Bond, Rosie Frost, Paul Morris, Emmanuel Okafor, Andrew Walsh, Dean Wild,Lesley Whittaker, Quinta Ashcroft, Vishnu Chandrabalan, Stuart Clough, Rebecca Hargreaves, Joan Ravencroft, Stephanie Iaconianni, John Howles, Karen Greenwood, Ben Morris, Lisa Bonifacio, Sarah Eastwood, Nathan Harding, Paul Hooker, Jacek Pawlowski, Julie Cheney, Ana Maria Aldana Gelves, Prajwal Khairnar, Trudy Taylor, Sarah Curtis, Simon Godley, Chris Lewis, Amy Marchant, Mary Mason, Oluwatoni Ojelade, Stephen Onamade, Vijay Rabadia, Chloe Roebuck, Francesca Roffey, Saad Saleem, Neil Springham, Jon Blacker, Edmund Gall, Matthew Stammer, Michael George, Laura White

In addition, we are very grateful to members of the Programme Steering Board for their oversight and constructive feedback during the Programme roll-out.

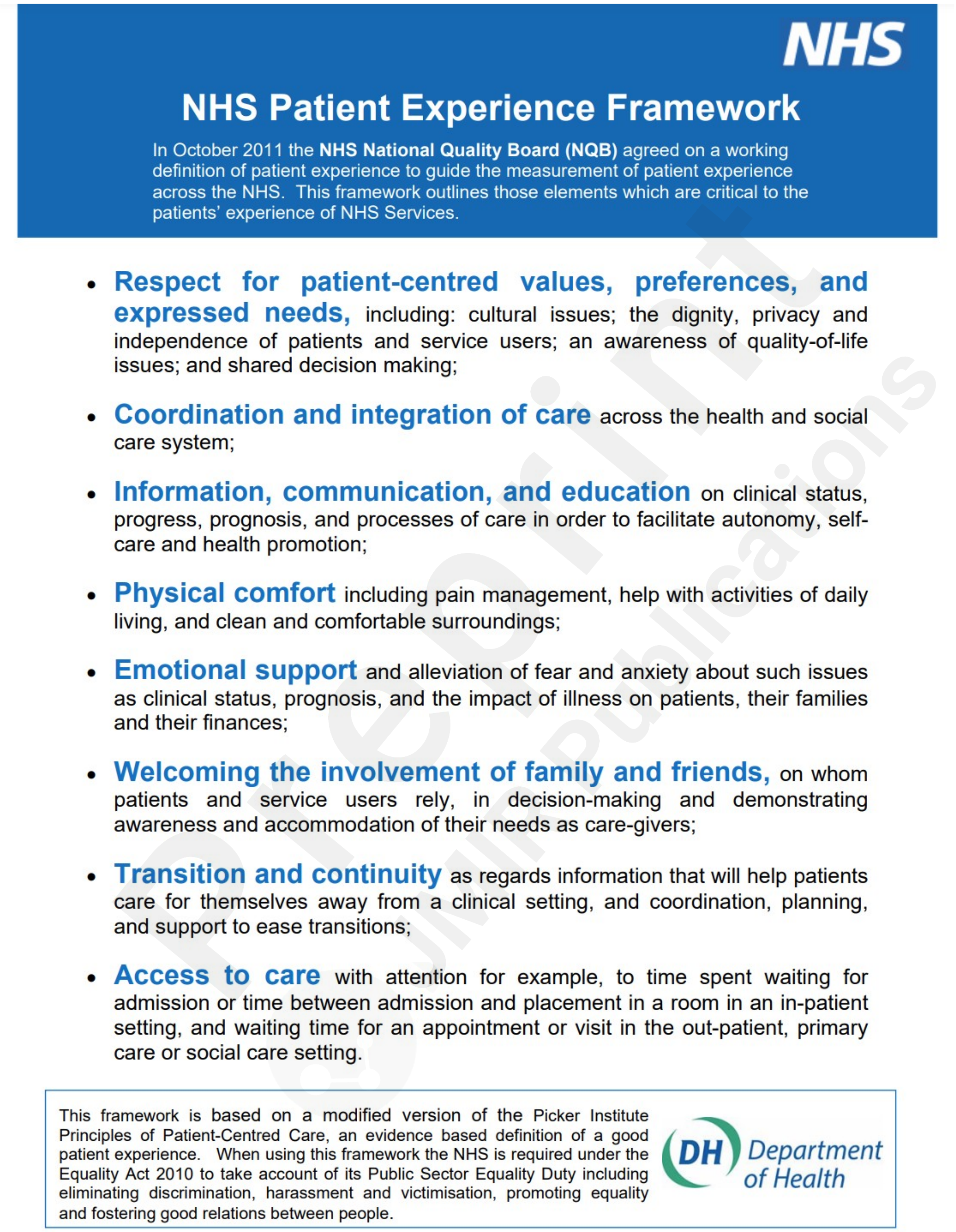
References

1. Flott KM, Graham C, Darzi A, Mayer E. Can we use patient-reported feedback to drive change? The challenges of using patient-reported feedback and how they might be addressed. BMJ Qual Saf. 2017;26(6):502-7.
2. England N. The Friends and Family Test. NHS England; 2014. Contract No.: Publication Gateway Ref No. 01787.
3. Gleeson H, Calderon A, Swami V, Deighton J, Wolpert M, Edbrooke-Childs J. Systematic review of approaches to using patient experience data for quality improvement in healthcare settings. BMJ Open. 2016;6(8):e011907.
4. Usman M, Mujahid M, Rustam F, Flores E, Vidal Mazon JL, Diez IT, Ashraf I. Analyzing patients satisfaction level for medical services using twitter data. PeerJ Comput Sci. 2024;10:e1697.
5. Using the Friends and Family Test to improve patient experience. NHS England and NHS Improvement; 2019.
6. Greaves F, Ramirez-Cano D, Millett C, Darzi A, Donaldson L. Harnessing the cloud of patient experience: using social media to detect poor quality healthcare. BMJ Qual Saf. 2013;22(3):251-5.
7. Khanbhai M, Warren L, Symons J, Flott K, Harrison-White S, Manton D, et al. Using natural language processing to understand, facilitate and maintain continuity in patient experience across transitions of care. Int J Med Inform. 2022;157:104642.
8. Khanbhai M, Symons J, Flott K, Harrison-White S, Spofforth J, Klaber R, et al. Enriching the Value of Patient Experience Feedback: Web-Based Dashboard Development Using Co-design and Heuristic Evaluation. JMIR Hum Factors. 2022;9(1):e27887.
9. Enston C. The power of the written word2021. Available from: https://www.england.nhs.uk/blog/the-power-of-the-written-word/.
10. Board NNQ. NHS Patient Experience Framework. Department of Health; 2011.
11. Khanbhai M, Anyadi P, Symons J, Flott K, Darzi A, Mayer E. Applying natural language processing and machine learning techniques to patient experience feedback: a systematic review. BMJ Health Care Inform. 2021;28(1).
12. Pang B, Lee, L. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval. 2008;2:1-135.
13. Towbin AJ. Collecting Data to Facilitate Change. J Am Coll Radiol. 2019;16(9 Pt B):1248-53.
14. Neumann J. FAIR Data Infrastructure. Adv Biochem Eng Biotechnol. 2022;182:195-207.
15. LaVela S, Gallan, AS. Evaluation and measurement of patient experience. Patient Experience Journal. 2014;1(1):28-36.
16. Garrisson M WJ. The role of volunteer in improving patient experience. Dallas TX: The Beryl Institute2016 [Available from: http://www.theberylinstitute.org/news/284517/The-Role-of-the-Volunteer-in-improving-Patient-Experience-Explored-by-The-Beryl-Institute-.htm.
17. Rastegar-Mojarad M, Ye Z, Wall D, Murali N, Lin S. Collecting and Analyzing Patient Experiences of Health Care From Social Media. JMIR Res Protoc. 2015;4(3):e78.
18. Indovina K, Keniston A, Reid M, Sachs K, Zheng C, Tong A, et al. Real-time patient experience surveys of hospitalized medical patients. J Hosp Med. 2016;11(4):251-6.
19. Hawkins JB, Brownstein JS, Tuli G, Runels T, Broecker K, Nsoesie EO, et al. Measuring patient-perceived quality of care in US hospitals using Twitter. BMJ Qual Saf. 2016;25(6):404-13.
20. Greaves F, Laverty AA, Cano DR, Moilanen K, Pulman S, Darzi A, Millett C. Tweets about hospital quality: a mixed methods study. BMJ Qual Saf. 2014;23(10):838-46.
21. Alemi F, Torii M, Clementz L, Aron DC. Feasibility of real-time satisfaction surveys through automated analysis of patients' unstructured comments and sentiments. Qual Manag Health Care.

2012;21(1):9-19.
22. Nawab K, Ramsey G, Schreiber R. Natural Language Processing to Extract Meaningful Information from Patient Experience Feedback. Appl Clin Inform. 2020;11(2):242-52.
23. Wagland R, Recio-Saucedo A, Simon M, Bracher M, Hunt K, Foster C, et al. Development and testing of a text-mining approach to analyse patients' comments on their experiences of colorectal cancer care. BMJ Qual Saf. 2016;25(8):604-14.
24. Dowding D, Randell R, Gardner P, Fitzpatrick G, Dykes P, Favela J, et al. Dashboards for improving patient care: review of the literature. Int J Med Inform. 2015;84(2):87-100.
25. P P. What gets measured gets done. Or does it?2014. Available from: https://www.health.org.uk/blogs/what-gets-measured-gets-done-or-does-it.
1. Flott KM, Graham C, Darzi A, Mayer E. Can we use patient-reported feedback to drive change? The challenges of using patient-reported feedback and how they might be addressed. BMJ Qual Saf. 2017;26(6):502-7.
2. England N. The Friends and Family Test. NHS England; 2014. Contract No.: Publication Gateway Ref No. 01787.
3. Gleeson H, Calderon A, Swami V, Deighton J, Wolpert M, Edbrooke-Childs J. Systematic review of approaches to using patient experience data for quality improvement in healthcare settings. BMJ Open. 2016;6(8):e011907.
4. Usman M, Mujahid M, Rustam F, Flores E, Vidal Mazon JL, Diez IT, Ashraf I. Analyzing patients satisfaction level for medical services using twitter data. PeerJ Comput Sci. 2024;10:e1697.
5. Using the Friends and Family Test to improve patient experience. NHS England and NHS Improvement; 2019.
6. Greaves F, Ramirez-Cano D, Millett C, Darzi A, Donaldson L. Harnessing the cloud of patient experience: using social media to detect poor quality healthcare. BMJ Qual Saf. 2013;22(3):251-5.
7. Khanbhai M, Warren L, Symons J, Flott K, Harrison-White S, Manton D, et al. Using natural language processing to understand, facilitate and maintain continuity in patient experience across transitions of care. Int J Med Inform. 2022;157:104642.
8. Khanbhai M, Symons J, Flott K, Harrison-White S, Spofforth J, Klaber R, et al. Enriching the Value of Patient Experience Feedback: Web-Based Dashboard Development Using Co-design and Heuristic Evaluation. JMIR Hum Factors. 2022;9(1):e27887.
9. Enston C. The power of the written word2021. Available from: https://www.england.nhs.uk/blog/the-power-of-the-written-word/.
10. Board NNQ. NHS Patient Experience Framework. Department of Health; 2011.
11. Khanbhai M, Anyadi P, Symons J, Flott K, Darzi A, Mayer E. Applying natural language processing and machine learning techniques to patient experience feedback: a systematic review. BMJ Health Care Inform. 2021;28(1).
12. Pang B, Lee, L. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval. 2008;2:1-135.
13. Towbin AJ. Collecting Data to Facilitate Change. J Am Coll Radiol. 2019;16(9 Pt B):1248-53.
14. Neumann J. FAIR Data Infrastructure. Adv Biochem Eng Biotechnol. 2022;182:195-207.
15. LaVela S, Gallan, AS. Evaluation and measurement of patient experience. Patient Experience Journal. 2014;1(1):28-36.
16. Garrisson M WJ. The role of volunteer in improving patient experience. Dallas TX: The Beryl Institute2016 [Available from: http://www.theberylinstitute.org/news/284517/The-Role-of-the-Volunteer-in-improving-Patient-Experience-Explored-by-The-Beryl-Institute-.htm.
17. Rastegar-Mojarad M, Ye Z, Wall D, Murali N, Lin S. Collecting and Analyzing Patient Experiences of Health Care From Social Media. JMIR Res Protoc. 2015;4(3):e78.

18. Indovina K, Keniston A, Reid M, Sachs K, Zheng C, Tong A, et al. Real-time patient experience surveys of hospitalized medical patients. J Hosp Med. 2016;11(4):251-6.
19. Hawkins JB, Brownstein JS, Tuli G, Runels T, Broecker K, Nsoesie EO, et al. Measuring patient-perceived quality of care in US hospitals using Twitter. BMJ Qual Saf. 2016;25(6):404-13.
20. Greaves F, Laverty AA, Cano DR, Moilanen K, Pulman S, Darzi A, Millett C. Tweets about hospital quality: a mixed methods study. BMJ Qual Saf. 2014;23(10):838-46.
21. Alemi F, Torii M, Clementz L, Aron DC. Feasibility of real-time satisfaction surveys through automated analysis of patients' unstructured comments and sentiments. Qual Manag Health Care. 2012;21(1):9-19.
22. Nawab K, Ramsey G, Schreiber R. Natural Language Processing to Extract Meaningful Information from Patient Experience Feedback. Appl Clin Inform. 2020;11(2):242-52.
23. Wagland R, Recio-Saucedo A, Simon M, Bracher M, Hunt K, Foster C, et al. Development and testing of a text-mining approach to analyse patients' comments on their experiences of colorectal cancer care. BMJ Qual Saf. 2016;25(8):604-14.
24. Dowding D, Randell R, Gardner P, Fitzpatrick G, Dykes P, Favela J, et al. Dashboards for improving patient care: review of the literature. Int J Med Inform. 2015;84(2):87-100.
25. P P. What gets measured gets done. Or does it?2014. Available from: https://www.health.org.uk/blogs/what-gets-measured-gets-done-or-does-it.

NHS Patient Experience Framework



NHS

NHS Patient Experience Framework

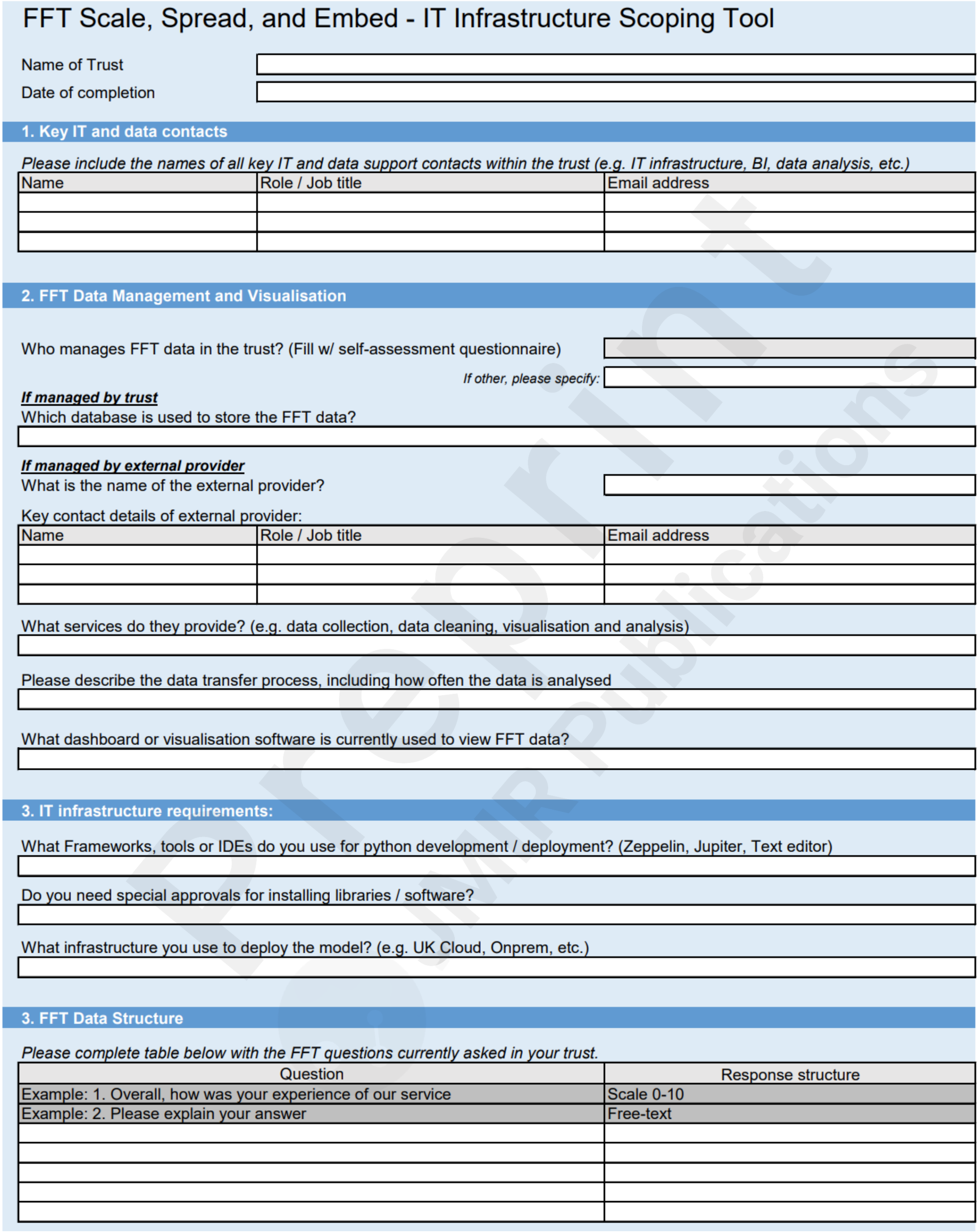
In October 2011 the **NHS National Quality Board (NQB)** agreed on a working definition of patient experience to guide the measurement of patient experience across the NHS. This framework outlines those elements which are critical to the patients' experience of NHS Services.

- **Respect for patient-centred values, preferences, and expressed needs,** including: cultural issues; the dignity, privacy and independence of patients and service users; an awareness of quality-of-life issues; and shared decision making;
- **Coordination and integration of care** across the health and social care system;
- **Information, communication, and education** on clinical status, progress, prognosis, and processes of care in order to facilitate autonomy, self-care and health promotion;
- **Physical comfort** including pain management, help with activities of daily living, and clean and comfortable surroundings;
- **Emotional support** and alleviation of fear and anxiety about such issues as clinical status, prognosis, and the impact of illness on patients, their families and their finances;
- **Welcoming the involvement of family and friends,** on whom patients and service users rely, in decision-making and demonstrating awareness and accommodation of their needs as care-givers;
- **Transition and continuity** as regards information that will help patients care for themselves away from a clinical setting, and coordination, planning, and support to ease transitions;
- **Access to care** with attention for example, to time spent waiting for admission or time between admission and placement in a room in an in-patient setting, and waiting time for an appointment or visit in the out-patient, primary care or social care setting.

This framework is based on a modified version of the Picker Institute Principles of Patient-Centred Care, an evidence based definition of a good patient experience. When using this framework the NHS is required under the Equality Act 2010 to take account of its Public Sector Equality Duty including eliminating discrimination, harassment and victimisation, promoting equality and fostering good relations between people.

DH Department of Health

IT scoping tool



FFT Scale, Spread, and Embed - IT Infrastructure Scoping Tool

Name of Trust	
Date of completion	

1. Key IT and data contacts

Please include the names of all key IT and data support contacts within the trust (e.g. IT infrastructure, BI, data analysis, etc.)

Name	Role / Job title	Email address

2. FFT Data Management and Visualisation

Who manages FFT data in the trust? (Fill w/ self-assessment questionnaire)

If other, please specify:

If managed by trust

Which database is used to store the FFT data?

If managed by external provider

What is the name of the external provider?

Key contact details of external provider:

Name	Role / Job title	Email address

What services do they provide? (e.g. data collection, data cleaning, visualisation and analysis)

Please describe the data transfer process, including how often the data is analysed

What dashboard or visualisation software is currently used to view FFT data?

3. IT infrastructure requirements:

What Frameworks, tools or IDEs do you use for python development / deployment? (Zeppelin, Jupiter, Text editor)

Do you need special approvals for installing libraries / software?

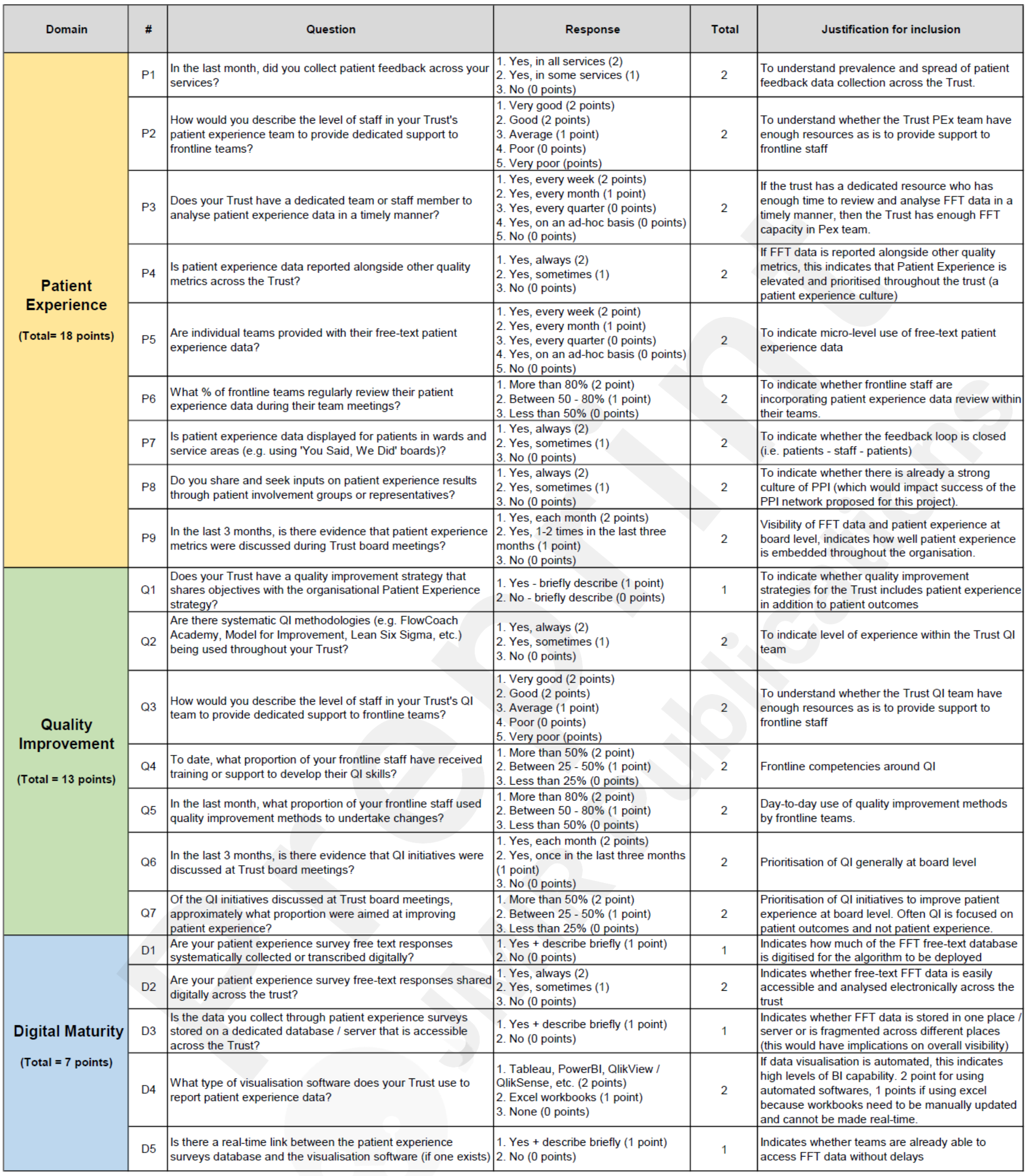
What infrastructure you use to deploy the model? (e.g. UK Cloud, Onprem, etc.)

3. FFT Data Structure

Please complete table below with the FFT questions currently asked in your trust.

Question	Response structure
Example: 1. Overall, how was your experience of our service	Scale 0-10
Example: 2. Please explain your answer	Free-text

FFT capability review

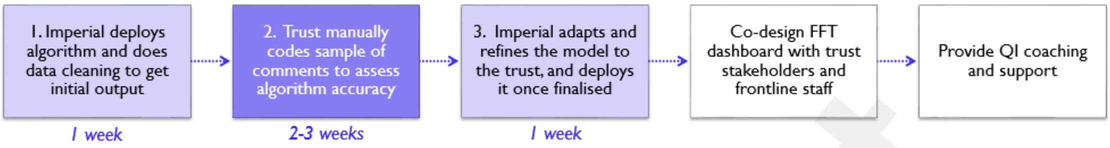


Domain	#	Question	Response	Total	Justification for inclusion
Patient Experience **(Total= 18 points)**	P1	In the last month, did you collect patient feedback across your services?	1. Yes, in all services (2) 2. Yes, in some services (1) 3. No (0 points)	2	To understand prevalence and spread of patient feedback data collection across the Trust.
	P2	How would you describe the level of staff in your Trust's patient experience team to provide dedicated support to frontline teams?	1. Very good (2 points) 2. Good (2 points) 3. Average (1 point) 4. Poor (0 points) 5. Very poor (points)	2	To understand whether the Trust PEx team have enough resources as is to provide support to frontline staff
	P3	Does your Trust have a dedicated team or staff member to analyse patient experience data in a timely manner?	1. Yes, every week (2 points) 2. Yes, every month (1 point) 3. Yes, every quarter (0 points) 4. Yes, on an ad-hoc basis (0 points) 5. No (0 points)	2	If the trust has a dedicated resource who has enough time to review and analyse FFT data in a timely manner, then the Trust has enough FFT capacity in Pex team.
	P4	Is patient experience data reported alongside other quality metrics across the Trust?	1. Yes, always (2) 2. Yes, sometimes (1) 3. No (0 points)	2	If FFT data is reported alongside other quality metrics, this indicates that Patient Experience is elevated and prioritised throughout the trust (a patient experience culture)
	P5	Are individual teams provided with their free-text patient experience data?	1. Yes, every week (2 point) 2. Yes, every month (1 point) 3. Yes, every quarter (0 points) 4. Yes, on an ad-hoc basis (0 points) 5. No (0 points)	2	To indicate micro-level use of free-text patient experience data
	P6	What % of frontline teams regularly review their patient experience data during their team meetings?	1. More than 80% (2 point) 2. Between 50 - 80% (1 point) 3. Less than 50% (0 points)	2	To indicate whether frontline staff are incorporating patient experience data review within their teams.
	P7	Is patient experience data displayed for patients in wards and service areas (e.g. using 'You Said, We Did' boards)?	1. Yes, always (2) 2. Yes, sometimes (1) 3. No (0 points)	2	To indicate whether the feedback loop is closed (i.e. patients - staff - patients)
	P8	Do you share and seek inputs on patient experience results through patient involvement groups or representatives?	1. Yes, always (2) 2. Yes, sometimes (1) 3. No (0 points)	2	To indicate whether there is already a strong culture of PPI (which would impact success of the PPI network proposed for this project).
	P9	In the last 3 months, is there evidence that patient experience metrics were discussed during Trust board meetings?	1. Yes, each month (2 points) 2. Yes, 1-2 times in the last three months (1 point) 3. No (0 points)	2	Visibility of FFT data and patient experience at board level, indicates how well patient experience is embedded throughout the organisation.
Quality Improvement **(Total = 13 points)**	Q1	Does your Trust have a quality improvement strategy that shares objectives with the organisational Patient Experience strategy?	1. Yes - briefly describe (1 point) 2. No - briefly describe (0 points)	1	To indicate whether quality improvement strategies for the Trust includes patient experience in addition to patient outcomes
	Q2	Are there systematic QI methodologies (e.g. FlowCoach Academy, Model for Improvement, Lean Six Sigma, etc.) being used throughout your Trust?	1. Yes, always (2) 2. Yes, sometimes (1) 3. No (0 points)	2	To indicate level of experience within the Trust QI team
	Q3	How would you describe the level of staff in your Trust's QI team to provide dedicated support to frontline teams?	1. Very good (2 points) 2. Good (2 points) 3. Average (1 point) 4. Poor (0 points) 5. Very poor (points)	2	To understand whether the Trust QI team have enough resources as is to provide support to frontline staff
	Q4	To date, what proportion of your frontline staff have received training or support to develop their QI skills?	1. More than 50% (2 point) 2. Between 25 - 50% (1 point) 3. Less than 25% (0 points)	2	Frontline competencies around QI
	Q5	In the last month, what proportion of your frontline staff used quality improvement methods to undertake changes?	1. More than 80% (2 point) 2. Between 50 - 80% (1 point) 3. Less than 50% (0 points)	2	Day-to-day use of quality improvement methods by frontline teams.
	Q6	In the last 3 months, is there evidence that QI initiatives were discussed at Trust board meetings?	1. Yes, each month (2 points) 2. Yes, once in the last three months (1 point) 3. No (0 points)	2	Prioritisation of QI generally at board level
	Q7	Of the QI initiatives discussed at Trust board meetings, approximately what proportion were aimed at improving patient experience?	1. More than 50% (2 point) 2. Between 25 - 50% (1 point) 3. Less than 25% (0 points)	2	Prioritisation of QI initiatives to improve patient experience at board level. Often QI is focused on patient outcomes and not patient experience.
Digital Maturity **(Total = 7 points)**	D1	Are your patient experience survey free text responses systematically collected or transcribed digitally?	1. Yes + describe briefly (1 point) 2. No (0 points)	1	Indicates how much of the FFT free-text database is digitised for the algorithm to be deployed
	D2	Are your patient experience survey free-text responses shared digitally across the trust?	1. Yes, always (2) 2. Yes, sometimes (1) 3. No (0 points)	2	Indicates whether free-text FFT data is easily accessible and analysed electronically across the trust
	D3	Is the data you collect through patient experience surveys stored on a dedicated database / server that is accessible across the Trust?	1. Yes + describe briefly (1 point) 2. No (0 points)	1	Indicates whether FFT data is stored in one place / server or is fragmented across different places (this would have implications on overall visibility)
	D4	What type of visualisation software does your Trust use to report patient experience data?	1. Tableau, PowerBI, QlikView / QlikSense, etc. (2 points) 2. Excel workbooks (1 point) 3. None (0 points)	2	If data visualisation is automated, this indicates high levels of BI capability. 2 point for using automated softwares, 1 points if using excel because workbooks need to be manually updated and cannot be made real-time.
	D5	Is there a real-time link between the patient experience surveys database and the visualisation software (if one exists)	1. Yes + describe briefly (1 point) 2. No (0 points)	1	Indicates whether teams are already able to access FFT data without delays

Coding pack

Implementing the NLP algorithm in your trust

Three key steps need to be completed before deploying the NLP algorithm so that it can accurately analyse your FFT data:



1. Imperial deploys algorithm and does data cleaning to get initial output

1 week

2. Trust manually codes sample of comments to assess algorithm accuracy

2-3 weeks

3. Imperial adapts and refines the model to the trust, and deploys it once finalised

1 week

Co-design FFT dashboard with trust stakeholders and frontline staff

Provide QI coaching and support

The project's natural language processing (NLP) algorithm is developed iteratively based on deployment within other trusts to ensure it is as robust as possible.

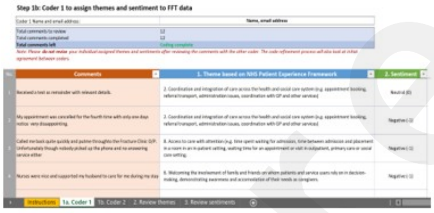
Each time the algorithm is deployed, we need to assess its accuracy by comparing it to a source of 'truth' – **the manual coding of a random sample of 500 FFT comments by two coders within the trust.**

This is a crucial and time-sensitive step as it ensures that the final algorithm can accurately analyze the FFT data, and we cannot continue with the project until this step is complete. **The coding process should take approximately 2-3 weeks to complete.**

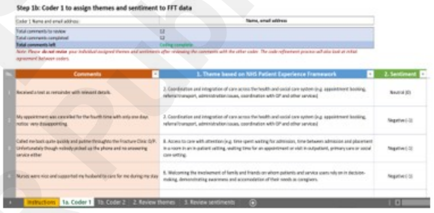
Excel files in the coding pack

You will have received three excel files for the coding process. All file have been prepared with a 500 sub-sample of FFT free-text comments from your trust.

1. **Master coding file** – main coding file that will be used by Coder 1 to individually review comments and to review any coding differences


Step 1b: Coder 1 to assign themes and sentiment to FFT data
Comments
1. Theme based on NHS Patient Experience Framework
2. Sentiment
Instructions
1a. Coder 1
1b. Coder 2
2. Review themes
3. Review sentiments

2. **Coder 2 coding file** – File that coder 2 can use to individually code comments. Once finished, the selected themes and sentiments should be copied over to the master file.

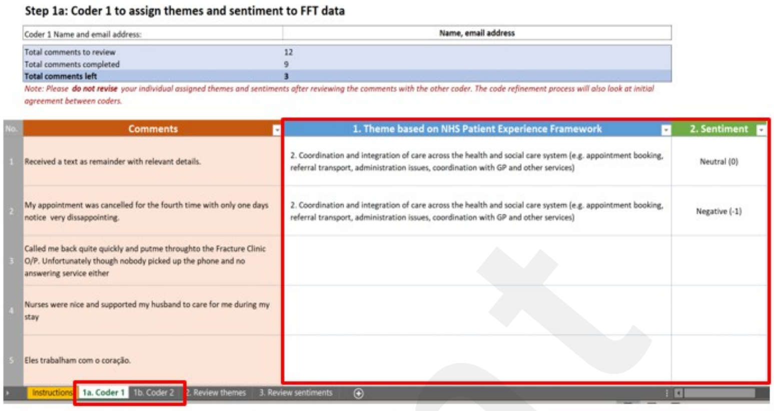

Step 1b: Coder 1 to assign themes and sentiment to FFT data
Comments
1. Theme based on NHS Patient Experience Framework
2. Sentiment
Instructions
1a. Coder 1
1b. Coder 2
2. Review themes
3. Review sentiments

Steps to code FFT free-text comments (1 of 3)

Agree on which two individuals within your trust will be doing the coding. We recommend selecting members of the Patient Experience team as they are most familiar with analysing FFT data for your trust.

Step 1: Over a week, code the first 100 comments together to ensure that both coders are analysing the FFT data in the same way. For each free-text comment, you will need to assign a **theme** based on the NHS Patient Experience Framework and a **Sentiment** (see slide 8).

Coder 1 should use the tab 1a. Coder 1, and coder 2 should use the tab 1b. Coder 2.



Step 1a: Coder 1 to assign themes and sentiment to FFT data

Coder 1 Name and email address:	**Name, email address**

Total comments to review	12
Total comments completed	9
Total comments left	**3**

*Note: Please **do not revise** your individual assigned themes and sentiments after reviewing the comments with the other coder. The code refinement process will also look at initial agreement between coders.*

No.	Comments	1. Theme based on NHS Patient Experience Framework	2. Sentiment
1	Received a text as remainder with relevant details.	2. Coordination and integration of care across the health and social care system (e.g. appointment booking, referral transport, administration issues, coordination with GP and other services)	Neutral (0)
2	My appointment was cancelled for the fourth time with only one days notice very dissappointing.	2. Coordination and integration of care across the health and social care system (e.g. appointment booking, referral transport, administration issues, coordination with GP and other services)	Negative (-1)
3	Called me back quite quickly and putme throughto the Fracture Clinic O/P. Unfortunately though nobody picked up the phone and no answering service either		
4	Nurses were nice and supported my husband to care for me during my stay		
5	Eles trabalham com o coração.		

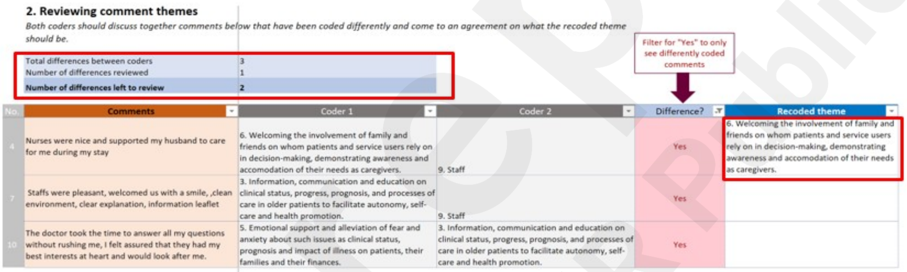
Steps to code FFT free-text comments (2 of 3)

Step 2: After coding the first 100 comments together, code the remaining 400 comments individually over a 1-2 week period.

Step 3: Copy Coder 2's individual coded comments into the master coding file.

Step 4: In the **"2. Review Themes" tab,** review any differences in coding of the themes between the two individuals. Filter for "yes" in the Differences column to only view comments that were coded differently. After discussion, recode the final agreed theme per comment.

Important: *Please do not revise your individual assigned themes after reviewing the comments with the other coder in the individual coding tabs. The code refinement process will also look at initial agreement between coders.*



2. Reviewing comment themes

Both coders should discuss together comments below that have been coded differently and come to an agreement on what the recoded theme should be.

Total differences between coders	3
Number of differences reviewed	1
Number of differences left to review	**2**

Filter for "Yes" to only see differently coded comments

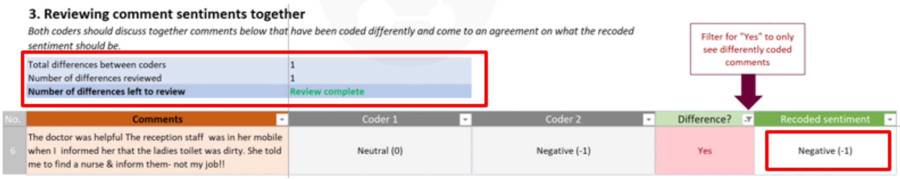
No.	Comments	Coder 1	Coder 2	Difference?	Recoded theme
4	Nurses were nice and supported my husband to care for me during my stay	6. Welcoming the involvement of family and friends on whom patients and service users rely on in decision-making, demonstrating awareness and accomodation of their needs as caregivers.	9. Staff	Yes	6. Welcoming the involvement of family and friends on whom patients and service users rely on in decision-making, demonstrating awareness and accomodation of their needs as caregivers.
7	Staffs were pleasant, welcomed us with a smile, ,clean environment, clear explanation, information leaflet	3. Information, communication and education on clinical status, progress, prognosis, and processes of care in older patients to facilitate autonomy, self-care and health promotion.	9. Staff	Yes	
10	The doctor took the time to answer all my questions without rushing me, I felt assured that they had my best interests at heart and would look after me.	5. Emotional support and alleviation of fear and anxiety about such issues as clinical status, prognosis and impact of illness on patients, their families and their finances.	3. Information, communication and education on clinical status, progress, prognosis, and processes of care in older patients to facilitate autonomy, self-care and health promotion.	Yes	

In this example, there are 3 comments with differences (comments 4,7,10), and the coders have recoded Comment #4 with Theme 6.

Steps to code FFT free-text comments (3 of 3)

Step 5: In the **"Review sentiments" tab,** review any differences in coding of the sentiments between the two individuals. Filter for "yes" in the Differences column to only view comments that were coded differently. After discussion, recode the final agreed sentiment per comment.

Important: *Please do not revise your individual assigned themes after reviewing the comments with the other coder in the individual coding tabs. The code refinement process will also look at initial agreement between coders.*



3. Reviewing comment sentiments together

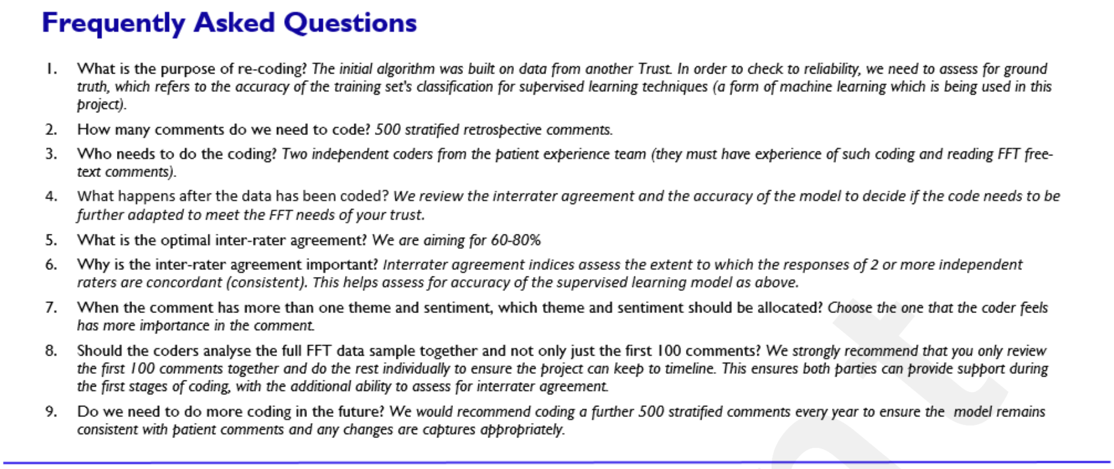
Both coders should discuss together comments below that have been coded differently and come to an agreement on what the recoded sentiment should be.

Total differences between coders	1
Number of differences reviewed	1
Number of differences left to review	Review complete

Filter for "Yes" to only see differently coded comments

No.	Comments	Coder 1	Coder 2	Difference?	Recoded sentiment
6	The doctor was helpful The reception staff was in her mobile when I informed her that the ladies toilet was dirty. She told me to find a nurse & inform them- not my job!!	Neutral (0)	Negative (-1)	Yes	Negative (-1)

In this example, there was only 1 difference in sentiment coding (Comment 6) and it was recoded to be a negative sentiment. The sentiment review is complete.



Frequently Asked Questions

1. What is the purpose of re-coding? *The initial algorithm was built on data from another Trust. In order to check to reliability, we need to assess for ground truth, which refers to the accuracy of the training set's classification for supervised learning techniques (a form of machine learning which is being used in this project).*
2. How many comments do we need to code? *500 stratified retrospective comments.*
3. Who needs to do the coding? *Two independent coders from the patient experience team (they must have experience of such coding and reading FFT free-text comments).*
4. What happens after the data has been coded? *We review the interrater agreement and the accuracy of the model to decide if the code needs to be further adapted to meet the FFT needs of your trust.*
5. What is the optimal inter-rater agreement? We *are aiming for 60-80%*
6. Why is the inter-rater agreement important? *Interrater agreement indices assess the extent to which the responses of 2 or more independent raters are concordant (consistent). This helps assess for accuracy of the supervised learning model as above.*
7. When the comment has more than one theme and sentiment, which theme and sentiment should be allocated? *Choose the one that the coder feels has more importance in the comment.*
8. Should the coders analyse the full FFT data sample together and not only just the first 100 comments? We *strongly recommend that you only review the first 100 comments together and do the rest individually to ensure the project can keep to timeline. This ensures both parties can provide support during the first stages of coding, with the additional ability to assess for interrater agreement.*
9. Do we need to do more coding in the future? We *would recommend coding a further 500 stratified comments every year to ensure the model remains consistent with patient comments and any changes are captures appropriately.*