Determining patient outcomes from patient letters: A comparison of text analysis approaches

Abstract

This paper presents a case study comparing text analysis approaches used to classify the current status of a patient to inform scheduling. It aims to help one of the UKs largest healthcare providers systematically capture patient outcome information following a clinic attendance, ensuring records are closed when a patient is discharged and follow-up appointments can be scheduled to occur within the time-scale required for safe, effective care. Analysing patient letters allows systematic extraction of discharge or follow-up information to automatically update a patient record. This clarifies the demand placed on the system, and whether current capacity is a barrier to timely access. Three approaches for systematic information capture are compared: phrase identification (using lexicons), word frequency analysis and supervised text mining. Approaches are evaluated according to their precision and stakeholder acceptability. Methodological lessons are presented to encourage project objectives to be considered alongside text classification methods for decision support tools.

Keywords

Decision support systems, Health service, Text mining, Information systems.

1 Introduction

The quantity of data being collected and generated in healthcare is ever growing (Huang et al. 2015). Numeric data is consistently used to support patient management and report performance, but unstructured textual data is not often utilised. Automated analysis of this textual data has moved forward significantly in the last few years with the development of processes and algorithms to retrieve, extract or classify information. A number of methods exist and analysts must choose methods according to the context and objectives of the problem.

There are many terms used to describe the processes undertaken to systematically and automatically analyse textual data: text analysis, text mining, natural language processing, natural language processing, information retrieval, information extraction. For consistency, we will refer to all text-to-data processes as *text analysis* (as is used by Byrd et al. (2014)).

This work is driven by issues faced when planning outpatient service capacity and scheduling patients in real life. Srinivas and Ravindran's (2017, 2018) work on the classification and prediction of patients failing to show for appointments highlights the overall opportunity of, and later the potential to optimise outpatient appointment systems, using machine learning algorithms. Planning services requires careful management of the waiting list, and a key data item is whether and when a patients requires an appointment. Existing work often assumes that the requirement for a patient to attend an appointment is robustly known, whereas that decision may in practice be informed by several case notes reviews by a clinician, with a degree of watchful waiting. Therefore the key motivations for this work were:

1. Address an issue of immediate patient safety.
2. The need for a tool to support clinical decision making and administrative validation.
3. The need to understand follow-up demand to optimise the scheduling of patients.

4. Recognition that over 80% of healthcare data is likely unstructured (Murdoch & Detsky 2013).

The aim of this paper is to explore what text analysis methods can be successfully used and how, to reduce the amount of manual administrative data entry within a healthcare organisation to support automatic appointment scheduling. Rich textual information from clinicians is already being captured regularly and holds the potential as a source for populating databases. The aim is to summarise and classify this rich, free-text information to convert it to data which is digestible by the patient management system (PMS) in order to use it to dynamically schedule patient appointments. The remainder of the paper is organized as follows:

- The following subsections summarise the use of data mining to support Operational Research (OR) projects and related works on text mining applied within healthcare systems.
- Section 2 describes the context of this study, highlighting the importance of accuracy within this project and the challenges faced in the text documents.
- Section 3 presents the methodology of the project and present the four methods used for classification of the corpus.
- Evaluation results are reported in Section 4 and Section 5 concludes this paper.

1.1 Scheduling and Data Mining

Scheduling methods are used in healthcare for the management of outpatients to automate appointment allocation and optimise resource utility. Methods for reducing outpatient waiting times range from simple evaluation of booking policies using queuing theory (O'Keefe 1985) or simulation modelling (such as Harper & Gamlin 2003), to optimisation methods to allocate individual appointments under difficult conditions (such as Cayirli & Gunes 2014) and data mining to predict no-shows when clinic scheduling (Srinivas & Ravindran 2018; Glowacka et al. 2009). This wait may be the time spent waiting to receive an appointment (wait for treatment) and wait at clinic on the day of the appointment (in clinic waiting time). In order to equitably schedule appointments by the wait for treatment, such as in outpatient scheduling, it is imperative to know the intended date of attendance.

OR projects have been undertaken in healthcare to explore scheduling solutions to minimise both of these waits. Several have utilised data mining methods alongside traditional OR methods to improve coherency and classification of data for planning and improvement (e.g. Ceglowski et al 2007). Although a 2009 review of 50 years of data mining and OR (Baesens et al.) did not identify healthcare as a key area of growth, they note the increasing popularity of data and text mining in OR and the challenge poor data quality presents when applying these methods. Mironczuk and Protasiewicz's 2018 review of the state-of-the art of text mining highlights the key elements of text mining, and future directions of integrating with other methods, but fails to note the opportunities within healthcare. However, the utility of classification algorithms for informing both clinical and operational decisions (Harper 2005), and emerging areas of data mining for healthcare resource management, administration and hospital management (Sharma & Mansotra 2014) have been previously identified. Therefore, this work sought to extend the knowledgebase of the utility of data mining as a core precursor to scheduling and other OR planning solutions, such as simulation.

1.2 Text Analysis

Luhn (1958) is the first documented use of text analysis: word frequency analysis was used to summarise abstracts. It was identified as a future direction (of data analysis) but was overshadowed

by emphasis on numerical data due to the challenges of working textual[1] data and the value being found in non-textual / numeric data. Automated analytical methods, such as text mining, are particularly well suited to identifying hidden connections not easily manually perceptible (Wu et al. 2014). Finding patterns and nuggets of information from within numeric and textual data is referred to by many names, spanning several fields. Overall, Hearst (1999) suggests that 'real' text mining is concerned with finding novel insights in textual data (illustrated in Table 1). Therefore, we use the term text analysis to cover a range of methods to systematically analyse non-categorical textual data (free-text).

Table 1. A classification of data mining and text data mining applications (Reproduced from Hearst, 1999, p.5)

	Finding Patterns	**Finding Nuggets**	
		Novel	**Non-Novel**
Non-textual data	Standard data mining	?	Database queries
Textual data	Computational linguistics	Real Text Data Mining	Information retrieval

Text analysis of a set of records (corpus) covers a range of 'tasks' (illustrated in Figure 1):

- Information retrieval (IR) refers to identifying appropriate documents or set of textual materials held online or in a file management system for analysis.
- Information extraction (IE) refers to identifying relevant information from the corpus. It frequently consists of processing (into simple structures), identification of key terms and entities, and interpretation of the relationships between them. Recent reviews on this highlight an emphasis on pattern matching methods (dictionary lookup) (Spasić et al. 2014).
- Categorisation seeks to devise rules for clustering the records within the corpus. Single or multiple categories may be assigned to each record.

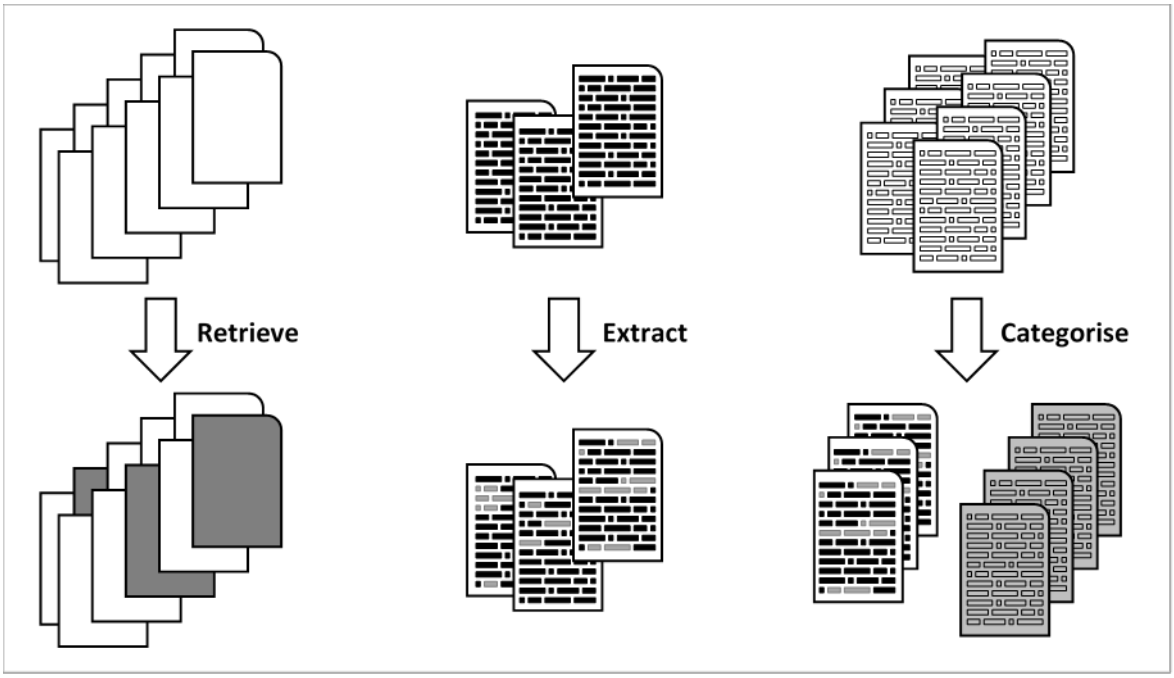

Figure 1. Illustration of text analysis tasks Information retrieval, information extraction and document classification

Text mining systems may require a composition of these tasks to give the most powerful knowledge discovery capabilities. More recently, the focus has shifted from algorithm development to

[1] Textual data is text that is perceived as unstructured by a numerically driven database.

application (Hearst 1999) due to the acknowledgement that context is a significant consideration in any text mining project.

The performance of classification tools in healthcare depends on the features of the dataset to be analysed: where a dataset may be regarded as lying on a continuum of requirements, shown in Table 2 (Schultz 2014). The challenges of information extraction, information retrieval and classification may be further exacerbated by each of these characteristics. Challenges in text analysis stem from the original purpose of the data collection, access and privacy which can be a significant issue for healthcare applications (Meystre et al. 2008; Chapman et al. 2011). This has led to a lack of coded datasets to test algorithms, and the wealth of possible applications within healthcare being inaccessible to researchers.

Text analysis has been used in practice in a number of areas of health, some more enthusiastically than others. The methods have been used extensively for purposes from literature review and knowledge discovery (Rebholz-Schuhmann et al. 2005); to security, biomedical, marketing, search/information access (Dalianis et al. 2011). For all the success of text mining, there remain a number of significant challenges to making the methods well-used on healthcare data and within healthcare:

- Ambiguity within the text remains problematic, whereby the tone of the author is not apparent and the text is open to multiple interpretations (Meystre et al. 2008).
- It has been previously noted that it is important to evaluate methodologies in the context of supporting real-life tasks (Korhonen et al. 2012). Algorithms are frequently developed on open-source datasets (news pages, twitter feeds, journal abstracts) meaning that applications outside of these areas could offer a range of new challenges.

Table 2. Continuum of dimensions of data (reproduced from Schultz 2014)

Well-structured data	Heterogeneous data sets
Automatically generated metadata	complex metadata issues
Static data	dynamically changing data
Data acquired under controlled conditions	crowd-sourced data
Centrally managed databases	Distributed data, no clear curation
Computationally simple	Data needing massive computing
Data that are used “raw”	Data that are understandable only after processing
Numerical data	Text data
Knowledgeable data communities	Communities scared of data
Communities with trust	Communities with no tradition of sharing, or even with distrust
Open data	Propriety/embargoed data, data with copyright issues
Impersonal data	Data with privacy issues
Privately generated data	Data with publicly funded stakeholders

1.3 Text analysis *in* Healthcare

There is interest in using automated processes to obtain richer insight into data being routinely collected within healthcare organisations (Chapman et al., 2011). The field presents a range of opportunities for text analysis, with particular interest in the automatic coding of patient diagnoses, procedures and events (Lin et al. 2017; Carroll et al. 2012; Miller & Velanovich, 2010); systematic identification of non-compliance, risk or harm; and to support medical research (identification of cohorts and retrospective coding). There is even interest in reversing the process – using data to produce summaries more readily read by clinicians to overview patient history (Scott et al. 2013).

Despite this breadth of research questions, it has led to a limited body of research. A 2008 review of text analysis of electronic health records noted that methods are rarely used "*outside of the laboratory they have been developed in*" (Meystre et al. 2008). Standardised medical dictionaries and look-ups exist which have been used to support text analysis, but algorithmic solutions are not commonplace. Rather, standardisation of text is favoured, but cannot be readily and systematically undertaken retrospectively.

Chapman et al. (2011) note the significant barriers to the assessment and utility of text analysis research in the clinical domain as:

- Lack of shared data available to researchers.
- Lack of common conventions and standard annotations mean high uniqueness within the data.
- The challenges of reproducibility - the nature of healthcare means that embedded solutions which can be adapted in-house as the system changes are required.
- Limited collaboration has led to unique methods tailored for each single context.
- Lack of user-focused development (usability) and scalability.

There is the potential for more general tools and algorithms to be evaluated on *real* datasets. However, in a recent review of text-mining of cancer related information it was noted that the lack of availability of clinical narratives is a barrier to the required shift from rule-based methods to machine learning due to the need for large training sets (Spasić et al. 2014).

This section has discussed the developments and remaining challenges of text analysis. Specifically, the wealth of opportunities, but barriers to research and implementation in healthcare highlight the relevance of this work. The specifics of the real life classification problem are discussed in the next section.

2 Classification problem

This section will discuss the context of the problem and the specifics of the data available to the researchers.

2.1 Patient outcomes

Cardiff and Vale University Health Board (CVUHB) is a large local health board that serves a population of 500,000 residents in Cardiff and the Vale of Glamorgan, Wales, UK and provides over 620,000 outpatient appointments per year. A single patient may have several appointments under one pathway of care (referred to as a patient pathway and illustrated in Figure 2). An outcome is required to be recorded against every attendance, which should convey the intended next steps for the patient: whether they are being discharged from the clinicians care (onwards to another clinician, to their GP or otherwise) or whether the patient will be reviewed again. Outcomes are recorded against every patient pathway on the PMS. A review of processes highlighted that a large cohort of patients were being assigned a 'query' outcome, whereby at the appointment the clinician has not conveyed their intention to discharge or follow-up the patient.

Clinicians use patient letters to record important and rich outcome information that is to be conveyed to the patient and/or their GP. These letters may be used to convey information already captured in the clinical outcome form or may represent the most up-to-date decision a clinician has made regarding the patient. For those patient pathways assigned a 'query' outcome, directorate staff will examine the relevant patient letters associated with the pathway to assign a sufficiently detailed outcome or to discharge the patient.

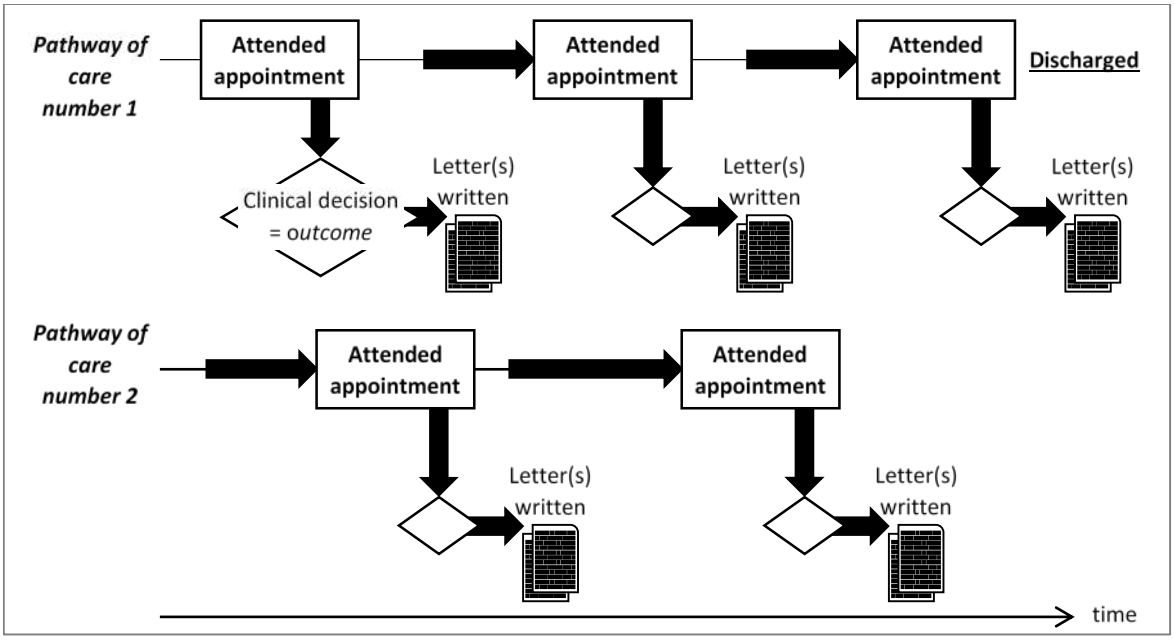

Figure 2. Illustration of one patient with two pathways of care (such as with different specialties) active in parallel

Four outcomes will be discussed in the remainder of this paper as they feature in the coded corpus that is used:

- (none) No outcome expressed clearly in the letter
- FU Patient requires a follow-up appointment
- WL Patient has been added to a waiting list (no follow-up required)
- DX Patient has been discharged by the clinician

2.2 Patient Letters

All electronically typed patient letters are stored in a database as plain text as well in their original (.doc, .pdf, .rtf) form. The letters are real correspondence sent from clinicians to the patient and/or their GP or other healthcare professional(s). Letters are used to update both parties on the status of the patient and may be used to indicate the next steps in the patient's care pathway. Letters may be generated from templates with the aim to convey specific information (such as 'discharge summaries', 'test results', or 'appointment') linked to the patient pathway or may be 'general letters' linked to the patient but not to the patient pathway. Specifically of interest to this project are letters which convey the outcome of a recent attendance and/or the intended next steps for the patient. Over 220,000 letters were identified by matching the clinician authoring the letter to the last event (attendance) on a patient pathway, or a letter that has been explicitly linked on the system to a patient pathway.

For this research, the focus is on identifying the decision made by the clinician as to whether the patient has been discharged from their care, or whether a follow-up appointment is required. These two are mutually exclusive - no overlap should occur. There may be ambiguity (such as a patient may be discharged following affirmative diagnostic tests), or no outcome may be expressed. In these cases a third category would be assigned: none.

Preliminary analysis of a sample of patient letters highlighted that they contain information being captured in the clinical outcome form. This is a doubling of information and therefore there is opportunity to use one to inform the other. Figure 3 shows a mocked up letter, based on a real

example, with pertinent information highlighted. In this example, it would be inappropriate to stem words to their simplest form as the past, present or future relevance would be lost.

Dear **[GP]**,

This pleasant [patient] was seen by **[the XYZ clinic]** **two months ago** during which review of [the patient] indicated **an onward referral to my care**.

They discharged her to me and I first saw [the patient] **in clinic three weeks ago**. I saw [this patient] again at **a follow-up appointment last week**. [The patients] condition remains stable but **another review is needed in a months' time before** [the patient] **can be discharged to your care**.

Follow-up appointment: 1 month

Yours sincerely,
[Dr Sample].

Key: **Past follow-up** **Future follow-up** **Other discharge (past/future/conditional)**

Figure 3. Example patient letter with mark-up of pertinent information

2.3 Specifics of the problem

2.3.1 Data type

To describe this data only as 'text' would underrepresent the understanding developed of the context of the problem. Drawing upon the dimensions of data detailed by Schultz (2014) and illustrated in Table 2, it is possible to specify characteristics which define the nature of the problem and influence the choice of methods.

- **Structure:** Letters are free-text, but there is an implied structure and formality in the language.
- **Static or dynamic:** This is a live system, and methods applied should not be a one-off endeavour. Letters are continuously generated and patients' outcomes change as they attend appointments.
- **Conditions of collection:** Data is collected in a controlled environment but is generated by a large number of individuals, meaning that data will likely display high variation in language and length.
- **Curation:** Letters are stored in a centrally managed database. However, some letters may be sent to patients but not uploaded to the system.
- **Raw or processing needed:** the data is readable in its stored format, no processing is required.
- **Numerical or text:** Text, with numerical meta and patient pathway data associated.
- **Community:** healthcare, with trusted partners.
- **Open or embargoed data:** Sensitive data, which cannot be extracted from healthcare site.
- **Privacy:** Data is subject to privacy rules

These characteristics indicate that care needs to be taken with the data (with appropriately qualified and trained staff assigning codes which indicate the clinical decision made or the lack thereof. The existing storage infrastructure is systematic, thereby providing rich, robust and timely text and numeric data available to apply automated text analysis approaches.

2.3.2 Logical and Lexical Semantics

Semantics, within this research, refers to the meaning in language (OED 2018). Textual data may be interpreted in different ways depending on the context of the data and the reason for analysis. Lexical semantics refers to the sub-word, affixes and compound word and phrase meanings. That is, lexical semantics may result in a single word having multiple meanings. Logical semantics refer to the sense of the word(s), the overall meaning. Logical semantics may result in a single sentence being ambiguous out of context from other sentences. Therefore, it may be important to take into account the full letter text rather than isolated words.

An example of lexical ambiguity is the word 'bank'. This may refer to the process of remembering information ("we will bank that suggestion") or a physical institution ("I will visit the bank").

An example of logical ambiguity is "old men and women". This may mean to apply 'old' only to 'men' or to both 'men and women'.

2.3.3 Tense: Past, present, future

There are no restrictions on the content of the letters and therefore they may contain any tense. A letter may: summarise past events to convey a patient history when referring a patient to another clinician; contain information about a recent event (such as an attendance, admission, or diagnostic test); and/or contain information about the next steps for the patient. This poses a challenge when analysing text as the presence of a single key term may not accurately indicate the current status of a patient pathway (such as the key term 'follow-up' which is used in reference to past and future intended activity in the example letter in Figure 3). As noted earlier, this structure would also imply that stemming words in the corpus may be misleading. This is because the variation in tense found in the letters indicates that whole words and word sequences are important, as well as overall position within the body of the text (such as at the start or end of the letter).

2.3.4 No outcome expressed

Letters are not guaranteed to include a clear statement of outcome or intended next steps from the clinician. In these cases it is preferable to not classify a letter or to classify as requiring follow-up. Therefore, the training dataset included ambiguous letters where no outcome could be assigned (category 'none').

This section has discussed the specifics of the problem sought to be addressed using text analysis. It has highlighted the challenges of validating the records by assigning one of up to four outcomes. The next section will describe the methodology employed for comparing four text analysis approaches. The characteristics of the coded corpus of letters, details of the four text analysis methods and the method for comparison are detailed.

3 Methodology

3.1 Data

The cohort of patient pathways requiring classification was identified as any patient pathway not yet discharged from outpatients, without a target date to indicate when they will be reviewed again.

This project involves a real operational issue in need of solving. Supervised training methods are preferred to provide confidence in outcomes. However, all pathways at the start of this project were unclassified. Creating a coded dataset was time and resource heavy. Therefore the methods detailed in the next section were progressively applied to the live data to assign outcomes to patient pathways (creating a coded dataset), suggest likely outcomes and ultimately implement systems

which adopt these methods to manage the problem going forward. Figure 4 illustrates the number of letters which form the coded corpus and which methods are applied to each.

The dataset consisted of 270,000 patient pathway records which contained a patient letter relevant to the latest clinical interaction; where each pathway contained one or more clinical interactions. The dataset contained nine numeric attributes for each pathway: Patient ID [1], Pathway ID [2], date and time of last clinical interaction [3&4], clinic code [5], speciality code [6], subspecialty [7], consultant [8], administrative outcome recorded [9][2]. Coded records contained an additional classification attribute. These nine numeric attributes were used for record selection and stratification of the coded samples, with only the free text content of the patient letter used for classification. Methods 1 and 2 are the application of manually constructed lexicons for classification. Methods 3 & 4 are statistical based methods: word frequency analysis and formal text analysis algorithms. These methods are detailed in the following sections.

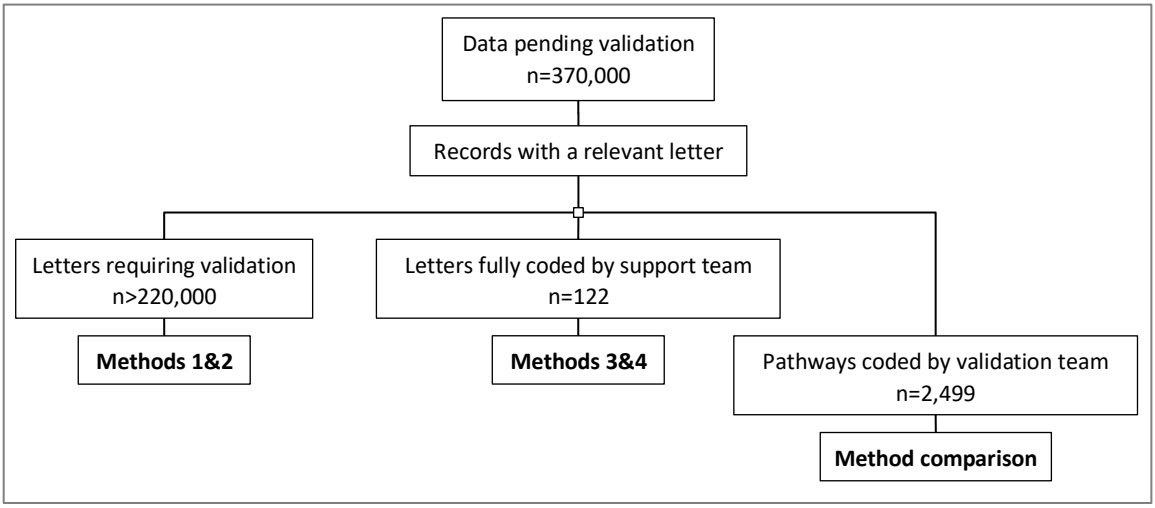

Figure 4. Proposed use of letter data for classification comparison

3.2 Methods for Text Analysis

This section will describe the four methods applied to the problem. Each section discussed the advantages and foreseeable limitations of the methods, and illustrates the application of the method using a flow diagram.

3.2.1 Method 1: Manual lexicon development

The aim of this method is to develop a set of phrases, without ambiguity, that can be used to robustly identify the outcome to be assigned to a patient pathway. The process, illustrated in Figure 5, consists of identifying a sample of letters on which to manually construct a set of phrases which determine the outcome for every letter in the sample. These phrases are extended by identifying appropriate synonyms and different phrase structures to create the lexicon to classify the remaining cohort. This is a simple method, which can be very time consuming if no patterns emerge in the letters. However, given that one specialty contains a finite number of clinicians who each may have their own style then it is reasonable to assume that the method will have a moderate success rate.

The advantages of this method is the transparency, which offers clinicians the reassurance that clinical decisions are not being made on their behalf, rather that the system is identifying a decision that has already been made and expressed in correspondence with the patient and/or their GP. The

[2] Note that the administrative outcome mainly consisted of 'query follow-up' – hence the need for validation.

size of the sample chosen and its representativeness of the cohort impacts the time for development. A small sample is quick to extract phrases from, but may require further iterations in order to improve the match rate. The success of this method is limited by the requirement for a letter to clearly state an outcome in a way which the clinical director and the medical board would agree is not ambiguous, or may carry an alternative meaning within another letter.

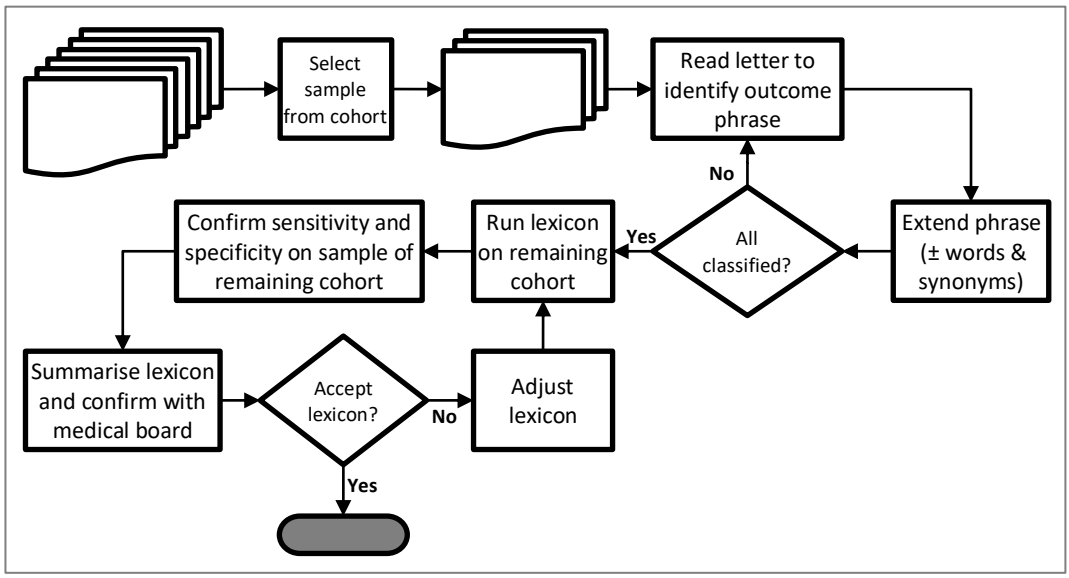

Figure 5. Method 1: Manual lexicon development process

3.2.2 Method 2: FastPhrase Lexicon

This method is designed as a speedy way to develop a lexicon (illustrated in Figure 6). Rather than approaching a sample of records and trying to manually code each in turn, this method requires the user to develop a set of 'seed' terms which may be reasonably associated with the context of the classification problem and the intended classification labels. In this case, seed terms associated with discharge or follow-up were used:

- Discharg (NB. This form enables the capture of 'discharge', 'discharging' and 'discharged')
- Follow-up
- Review
- Outcome
- Letter signoff - Yours sincerely, your faithfully, kind regards, best wishes. These seed phrases indicate the end of the body of the letter. Returning words prior to the letter signoff is indicative of the final statement to be conveyed to the patient and/or GP which may include the final outcome.

The advantage of this method is that a user is able to quickly identify letters which do not conform to their expectations as it is possible to observe the words searched for in the context of the letter. This method is used in conjunction with method 1, as a way to extend the lexicon by classifying pathways with relevant letters not being captured by the lexicon.

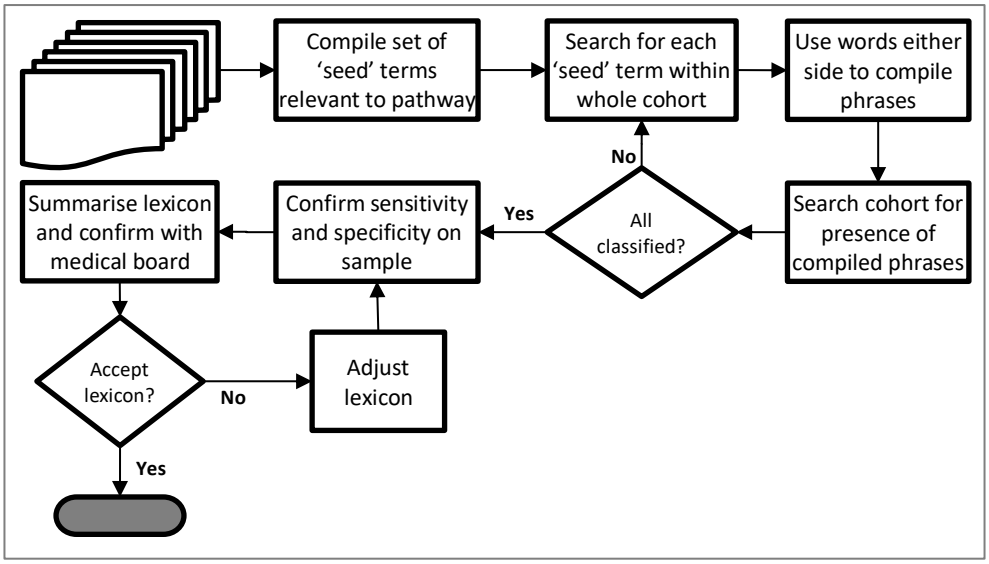

Figure 6. Method 2: FastPhrase lexicon development

3.2.3 Method 3: Word frequency analysis

Word frequency analysis has been found to be a simple yet useful form of text analysis when frequent terms not relevant to the context of the task are removed as it supports identification of a classification taxonomy (Pina, Chester et al. 2013). The word frequency analysis method is designed to try to capitalise on the content of the letters which have not been identified as key phrases for the lexicon. The rules are developed using a tree-based algorithm. This method (illustrated in Figure 7) is anticipated to be complementary to the development of the lexicon, rather than be applied in isolation by the Health Board.

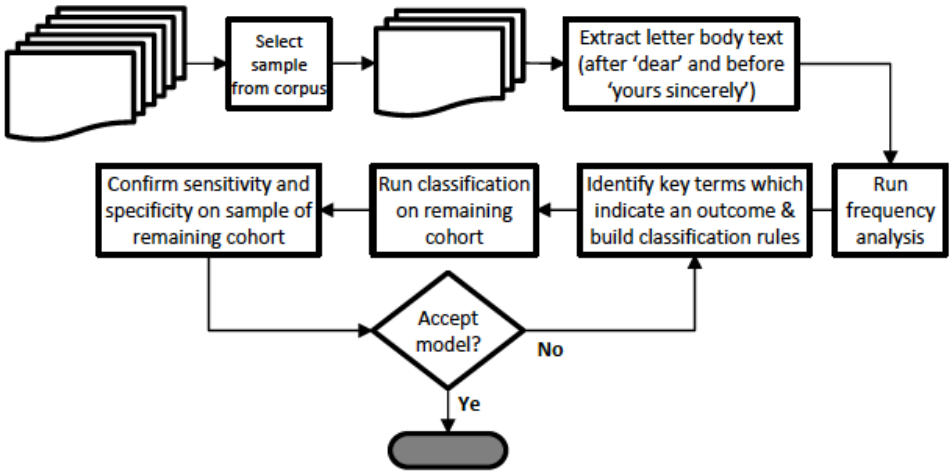

Figure 7. Method 3: Frequency of Words analysis to construct rules

3.2.4 Method 4: Text mining algorithms

Analysis was conducted in the statistical package R (R Core Team 2013). Seven text mining algorithms were applied using the package RTextTools (Jurka et al. 2012): Bagging, Logitboost, Maxent, Neural Networks (NNetwork), Random Forest, SLDA and Support Vector Machines (SVM). The process undertaken is illustrated in Figure 8. The training: testing split was selected following experimentation and tuning of the algorithms.

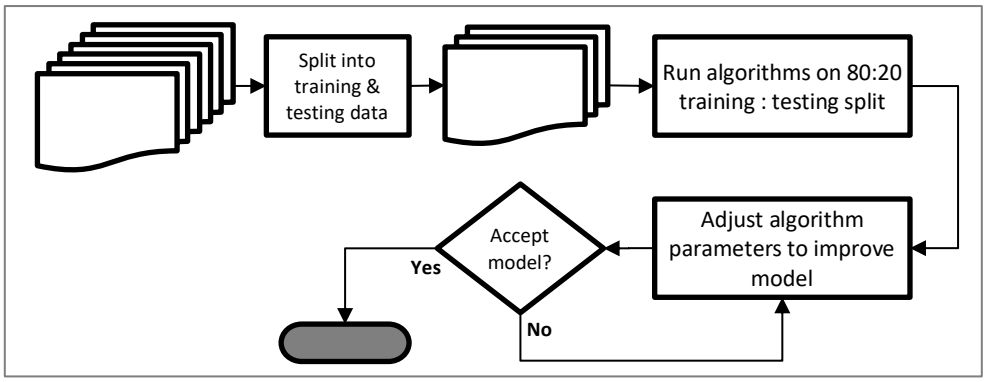

Figure 8. Method 4: Applying text mining algorithms through RTextTools

3.2.5 Comparison of methods

The tool developed for methods 1 & 2 has functionality for staff to manually validate pathways by assigning outcomes identified in the associated relevant letter(s). Letters classified in this way were not used to develop the training sets for methods 1-4 and represent an independent cohort of coded letter. This corpus of letters is used to compare the classification models developed using the above mentioned methods.

Comparison corpus: 2499 records manually classified by an independent team of validators seeking to assign outcomes to pathways. Outcomes were assigned through an in-house tool designed to tie patient letters to recent pathway activity. All outcomes assigned through the tool are assumed to be based upon the text in the most recent associated letter but some pathway information may be obtained from the numeric elements of the patient record.

Table 3. Overview of coded comparison corpus

Outcome	Count records	Length of record text (characters)			
		Average	StdDev	Minimum	Maximum
FU	2087	730.17	296.36	113	2567
DX	412	800.87	347.16	236	3253
Total	**2499**				

3.2.6 Definitions: measures of success

The primary measure of success is the number of records each method is able to classify correctly: the accuracy. Additionally, given the context of the problem, it is important to penalise misclassification more than failing to classify a letter. Therefore, increasing precision (reducing type 1 errors) is favoured over increasing recall (reducing type 2 errors).

$$Accuracy = \frac{(TruePositives + TrueNegatives)}{(TruePositives\ + FalsePositives + TrueNegatives + FalseNegatives)}$$

$$Precision = \frac{TruePositives}{TruePositives\ + FalsePositives}$$

Such measures are frequently used for binary classification (Sokolova & Lapalme 2009); in this context we will refer to 'positives' as 'discharges'. Therefore, the measure 'precision' refers to the number of correctly assigned discharges out of all assigned discharges and 'accuracy' is the percentage of all records correctly classified. These measures are used alongside the confusion matrix to evaluate the success of each method. A confusion matrix (otherwise known as a contingency table or error matrix) is a visualisation of the performance of a classification model by

representing predicated classification versus actual classification, and can be used to identify changes in performance which not observed in these measures (Chawla 2005).

This section has given an overview of the four methods being compared, with illustrations of the process undertaken with stakeholders at the health board. The following section provides the results of applying the methods to the cohort of letters requiring validation (method 1 & 2), and to the small coded corpus (method 3 & 4). The four methods are then applied to the coded corpus created by the validation team to establish their relative success.

4 Results and Analysis

4.1 Method 1: Manual lexicon development

This method created a lexicon of almost 15,000 phrases (constructed from 300 base phrases which were extended for generalizability using tags). Applied to the entire cohort requiring validation (approximately 270,000 records) it was possible to classify letters associated with 35,000 records (13%). Spots checks did not identify any misclassification, and feedback in the two months since implementation has revealed only one misclassification, indicating an accuracy of 99.997%.

The key advantage of this method is the level of confidence the health board is able to place in the classification. The lexicon was signed off by specialties and the patient safety committee. The method has made a significant inroad into the real-life classification problem. The initial outlay of analyst time for construction was large and the impact of adding new phrases diminishes over time (as less common phrases are identified). Phrases can be easily added and the process is simple to undertake regularly but this means that the method requires manual maintenance. Finally, given the size of the lexicon, exhaustively searching the corpus of letters requiring validation for lexicon phrases is time consuming.

4.2 Method 2: FastPhrase lexicon

This method identified an additional 52 base phrases (48 follow-up, 4 discharge), resulting in 3000 additional phrases in the lexicon. It offered users a systematic way to identify classification phrases without the need to extensively read letters and was therefore time efficient to implement. The process highlighted that discharge phrases were well represented within the lexicon (only 4 additions made). A comparable success to the manual lexicon is observed but with a far shorter development time (3 days). This is a method clinicians may adopt to validate their own cohort of patients.

4.3 Method 3: Word frequency analysis

This method was expected to give fresh insights into the corpus not yet observed through the development of the lexicon. However, when applied to both the initial fully coded dataset of 122 records and the comparison corpus (refer to Figure 4), no new synonyms for follow-up or discharge were identified. Therefore, it is expected that this particular problem context requires methods which use groups of words rather than single words (and their frequency within the text) to determine outcome.

Three classification algorithms were implemented through Weka (Hall et al. 2009) on frequent words which were agreed by CVUHB to be suitable to the context of the problem. Tree based methods were chosen to allow a set of rules based on word occurrence and co-occurrence to be devised. Ten-fold validation was used to make the most of the small corpus to train the best possible models. The confusion matrices below have the fourth column highlighted as discharge is the least

preferable misclassification (false positive - whereby 'discharge' is the assigned category but a follow-up or other treatment is required).

Simple CART: 99 of the 122 fully coded records were correctly classified.

Table 4. Method 3 simple CART confusion matrix - coded corpus

(none)	FU	WL	DX	-- classified as	
31	3	0	0	**a = (none)**	**Actual**
5	68	0	0	**b = FU**	
0	1	0	0	**c = WL**	
2	12	0	0	**d = DX**	

Random tree: 34 of the 122 fully coded records were correctly classified. Allowing no class to be assigned achieved the highest accuracy for this algorithm (83%: 34 of the 41 records classified).

Table 5. Method 3 random tree confusion matrix - coded corpus

(none)	FU	WL	DX	-- classified as	
17	0	0	0	**a = (none)**	**Actual**
0	17	0	0	**b = FU**	
0	0	0	0	**c = WL**	
1	6	0	0	**d = DX**	

Random forest: 98 of the 122 fully coded records were correctly classified. 14 instances were classified as none or follow-up when we wanted discharge. In fact, no discharge cases were successfully classified using this method. This makes this classification only robust in one direction (for FU) but is less informative than the lexicon (which can search for specific dates, times and clinics for follow-up).

Table 6. Method 3 random forest confusion matrix - coded corpus

(none)	FU	WL	DX	-- classified as	
32	2	0	0	**a = (none)**	**Actual**
6	66	0	**1**	**b = FU**	
0	1	0	0	**c = WL**	
2	12	0	0	**d = DX**	

Using carefully selected frequent words with tree based classification it is feasible to identify follow-up records, but fails to capture discharge. This method may be a useful starting point to investigate a corpus, but did not provide new insights following the development of the lexicon.

4.4 Method 4: Text mining algorithms

The package RTextTools (Jurka et al. 2012) was used to implement a selection of text analysis algorithms. Outputs from the best algorithms, according to the criteria stated in the previous section, are illustrated in Table 9.

Table 7. Method 4 overview of results - coded corpus

Algorithm	Number classified	Adjusted Accuracy percentage of those classified
Logitboost	17	94%
Nnetwork	19	84%
Bagging	19	84%
MaxEnt	19	84%
Random forest	17	82%
SLDA	13	92%
SVM	21	10%
Consensus of alg	13	93%

Examining the confusion matrices revealed that all methods were relatively poor at identifying discharged patients. This is expected due to the small number of coded cases within the corpus. Overall, results were improved by allowing the algorithms to assign an empty/null category rather than force each to assign discharge or follow-up.

4.5 Comparison of methods

Each method was applied with the parameters used above and the best algorithms to an independently coded dataset. A confusion matrix for each method is used to compare results (shown in Table 8). A confusion matrix shows the results of a classification process, comparing the actual classification versus the predicted classification.

Table 8. Confusion matrices of each method - validation corpus

Method 1: Lexicon

(none)	FU	WL	DX	-- classified as	
0	0	0	0	a = (none)	Actual
1279	**808**	0	0	b = FU	
0	0	0	0	c = WL	
352	39	0	**21**	d = DX	

Method 3: Frequent words

(none)	FU	WL	DX	-- classified as	
0	0	0	0	a = (none)	Actual
1287	**710**	0	90	b = FU	
0	0	0	0	c = WL	
31	351	0	**30**	d = DX	

Method 2: Lexicon modified with FastPhrase

(none)	FU	WL	DX	-- classified as	
0	0	0	0	a = (none)	Actual
1275	**812**	0	0	b = FU	
0	0	0	0	c = WL	
351	32	0	**29**	d = DX	

Method 4: Text mining algorithm - LogitBoost

(none)	FU	WL	DX	-- classified as	
0	0	0	0	a = (none)	Actual
0	**416**	0	13	b = FU	
0	0	0	0	c = WL	
0	33	0	**38**	d = DX	

Table 9. Accuracy of each method - validation corpus

Method(s)	Accuracy across all records (number correct)	Accuracy excluding unclassified	Accuracy assuming (none) = FU	Number classified as DX when should be FU (number incorrect)
1. Lexicon	33.2% (829)	95.5%	84.4%	0 (39)
2. FastPhrase	33.7% (841)	96.3%	84.7%	0 (32)
3. Frequent words	29.6% (740)	62.7%	81.1%	90 (461)
4. Text mining algorithms	90.8% (454)	90.8% *	==	13 (46)

The models created for methods 1 to 3 were applied directly to the validation team corpus (all 2499 records). The LogitBoost parameters used on the small coded corpus were used on a new training/testing split of the data to generate the outputs for method 4. A stratified sample was used for method 4 to reduce bias.

Table 9 shows the results of the four methods applied to the coded letters from the validation team. Although the lexicon method does not incorrectly classify any follow-up records as discharges (the hard constraint on this problem), the rate of classification of the records is low when compared to the text mining algorithms (33.2% versus 90.8%). The inclusion of phrases within the lexicon which were identified using the FastPhrase process improves the accuracy and reduces the number of misclassifications by identifying the records as having two contradictory categories (rather than a singular, incorrect category). The word frequency analysis had the poorest accuracy and the most misclassifications making it the least robust.

Given the context of the problem, an alternative measure of accuracy would be to assume all records assigned the 'none' category are 'follow-up' (column 4 of Table 9). This improves the accuracy of the word frequency analysis and improves the classification rate of methods 1 and 2 without strongly negatively impacting performance.

On this particular classification problem, we conclude that the LogitBoost classification algorithm is the least poor text mining method. This is in contrast to the conclusions of the review of classifiers by Fernandez-Delgado et al (2014) who conclude Random Forest, SVM and then tree-based methods perform with the highest accuracy across different datasets. The lexicon methods (FastPhrase and manual lexicon development) give higher accuracy when unclassified records are excluded. The text analysis literature suggests a shift is needed from rule-based methods to learning based methods. This paper has sought to illustrate the value of rule-based methods from an implementation perspective and how the rule-based methods have provided an acceptable platform to develop and implement learning-based methods.

In this project we have developed a clinical decision support tool which addresses an issue of immediate clinical safety. The precision and validity of this tool has been tested in comparison to other text analysis methods. This project demonstrated the value in using text analysis methods in healthcare on an operational issue. Provided a highly cost effect means of determining the clinical status, with a key focus on patient safety (by careful assessment of method accuracy). There is value to be gained by the standardisation of letters, indeed the Ophthalmology department have subsequently encouraged clinicians to include standard outcome phrases in their correspondence, and letter templates to be adjusted.

5 Conclusions

This paper has discussed a real life problem that is being solved using text analysis. Four methods for classifying patient letters have been compared. The iterative application of the methods in practice has provided insights into the perceived suitability of the methods to the context alongside their performance. The four key purposes which motivated the work are reflected on in this section.

Misclassification is a huge concern within this context and methods which are transparent, robust and exclude poor matching are preferable. A full clinical governance review process was undertaken when implementing this work with the view to addressing patient safety concerns and put into practice the decision making tool.

The scale of this problem meant that automated methods were vital. The cost and time of implementing a manual solution would have been huge. One person undertaking manual validation 200 days per year, 37.5 hours per day at a cost of £36,000 per year to the organisation. Therefore, the cost per pathway validated is estimated to be 90p requiring over £630,000 to validate the backlog. Limited resources available meant that undertaking and overseeing such a process was not possible. Plus the validation concern is an enduring concern requiring ongoing resource. The consequences for patients of a decision making tool are apparent, but existing methodologies (such as decision trees and pathway mapping) do not offer the automation desired and therefore alternative methods in the form of text analysis are evaluated and trialled.

A lack of certainty of the status of patients in the data has hampered understanding the follow-up demand and impeded dynamic scheduling implementation. Improved certainty is obtained through the phased application of methods: the alignment (consensus) of each method on a classification, evaluation of cases where incorrectness propagates and rejection where there is confusion (akin to the Swiss cheese model of risk). This will then feed into solutions that support scheduling under uncertainty.

Systematic methods are needed to convert data into information. This project has resulted in the development of a tool to start the HB on that journey.

The HB could try to provide review appointments for patients where there is uncertainty but the system is faced with no spare capacity within which to see them. Meaning more clinic slots need to be provided to prevent delaying pathways currently validly within the follow-up cycle. Therefore the impact of this work has been to:

- Prevent delays to patients incurred through the spill over of the un-validated pathways into the follow-up cycle,
- Avoid cost (for manual validation by non-clinical and clinical staff, and for additional clinical review clinics),
- Provide rich data for the continuous improvement of patient scheduling
- Present methods for systematically converting data to information for use throughout the patient management system.

Although it is preferable that methods can be applied quickly, the main priority is the transparency of the methods as this is pivotal to enabling confidence to be built among clinicians. A white box approach offers the most buy in. Methods which can be readily bought into are able to be agreed by the board and implemented swiftly, and reduce the amount of manual validation needed.

Using the formal text mining methods did not provide transparency. Throughout this project queries were raised by stakeholders as to whether the analysts were making a clinical decision to which the lexicon provided reassurance. Careful explanation of the processes employed and validation of the results further reassured stakeholders that the methods are trying to capture the clinical decision that has already been made (and expressed in the letter). In time, it may be possible to build sufficient support for statistical approaches which draw links between the patient type and the likelihood of follow-up

Methods were developed along the scale of usability, using the simplest methods as the entry point. Classification algorithms have been used to assign one of at most four outcomes to letters. However, additionally important is the assignment of a target date and clinic when follow-up is identified as the outcome. Therefore, more extensive training sets are needed or more advanced classification methods alongside information extraction methods are needed in further work.

6 Discussion

This work was done in conjunction with CVUHB information department, doctors, operational administrators and senior management. The validity of the methods and the overall approach was confirmed with clinical leads of the specialties and the head of information, alongside being used by administrators managing patient pathway records.

Positive feedback has been received throughout the process:

- "This work has significantly contributed to the extensive validation challenge the HB faced"
- "This project has been able to support the operational validation task [at the health board] in tandem with investigation into the viability and benefits of formal text analysis"
- "The transparency of the process and careful definitions of the lexicon has supported clinical and operational buy-in" to automating outcome validation.

The work has provided the basis of an ongoing system for linking outpatient clinic letters and the PMS on an intelligent, non-contradictory basis. It has led to CVUHB being involved in the Wales Clinical Record System being implemented by the NHS Wales Informatics Service (NWIS). The Wales Audit Office (2016) report into outpatient follow-up management noted the appropriateness of the methodology for automatic classification of patient outcomes: "*the health board is taking appropriate action to identify the volume of its outpatient follow-up*".

This project has overcome a number of the barriers to text analysis in healthcare noted by Chapman et al (2011):

- Access - working with the organisation, rather than on an open source dataset. This access enabled methods to be actively applied to help with an operational problem.
- Uniqueness – 'relevant' letters were identified according to the letter date and authorship. Letters were from all specialties providing outpatient appointments, composed by a large number of individuals. However, the large number of records counteracts this.
- Reproducibility - methods have been embedded on the live system, with users trained in the simpler methods, and all records flagged as categorised by the methods (for track-back).
- Collaboration across the health board (due to our position within the organisation) has been a key part of this work. The lexicon has been shared with another Health Board, the processes employed have been shared with the Wales Audit Office (with the automated classification methods recommended in their Outpatient follow-up report), and categorisation methods are being discussed with NWIS.

This work has focused on the precision and recall of each tool. This is due to the need to validate records versus developing a tool to optimise a system which have different measures of success. A clinical decision tool might seek total agreement which implies a better risk management tool.

The phased approach undertaken led to a coded dataset and a large lexicon. Next steps are to improve the text mining algorithms utilising both of these datasets, evaluate the impact of requiring total agreement between all methods, and apply the methodology to other areas of the organisation.

Acknowledgements

With thanks to Alex Poole and Leitchan Smith at Cardiff and Vale University Health Board for programming the in-house text search tool and provision of access and support to the dataset.

References

Baesens, B., Mues, C. , Martens, D., & Vanthienen, J. (2009). 50 years of data mining and OR: upcoming trends and challenges. *Journal of the Operational Research Society* 60(1): S16-S23.

Ben-Dov, M., & Feldman, R. (2005). Text Mining and Information Extraction: *The Data Mining and Knowledge Discovery Handbook*: 801-831. Springer.

Brailsford, S. C. & P. R. Harper (2007). Editorial. *Journal of the Operational Research Society* 58(2): 141-144.

Byrd RJ, Steinhubl SR, Sun J, Ebadollahi S, Stewart WF. (2014). Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records. *Int. Journal of Medical Informatics*, 83(12): 983-92.

Carroll, J., Koeling, R., & Puri, S. (2012). Lexical Acquisition for Clinical Text Mining using Distributional Similarity. *Computational Linguistics and Intelligent Text Processing, Springer Lecture Notes in Computer Science, 7182*: 232-246.

Ceglowski, R., Churilov, L. & Wasserthiel, J. (2007) Combining data mining and discrete event simulation for a value-added view of a hospital emergency department. *Journal of the Operational Research Society*, 58(2):246-254.

Chapman, W. W., Nadkarni, P. M., Hirschman, L., D'Avolio, L. W., Savova, G. K., & Uzuner, O. (2011). Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *Journal of the American Medical Informatics Association*, 18(5): 540-543.

Chawla, N. (2005) Data mining for imbalanced datasets: an overview. In O. Maimon & L. Rokach (Eds) *Data Mining and Knowledge Discovery Handbook.* Springer Science+Business Media, Inc. USA.

Cayirli, T. & E. D. Gunes (2014) Outpatient appointment scheduling in presence of seasonal walk-ins. *Journal of the Operational Research Society.* 65(4): 512-531. doi: 10.1057/jors.2013.56

Dalianis, H., Hassel, M., & Velupillai, S. (2011). Louhi 2010: Special issue on Text and Data Mining of Health Documents. *Journal of Biomedical Semantics, 2*(Suppl 3), I1-I1. doi: 10.1186/2041-1480-2-S3-I1

Fernandez-Delgado, M., Cernadas, E., Barro, S. & D. Amorim (2014) Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research* 15, 3133-3181.

Glowacka, K., Henry, R. & May, J. (2009). A hybrid data mining/simulation approach for modelling outpatient no-shows in clinic scheduling. *Journal of the Operational Research Society*, 60(8): 1056-1068.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The weka data mining software: An update. *sigkdd explorations, 11*(1).

Harper, P. R. (2005) A review and comparison of classification algorithms for medical decision making. *Health Policy*, 71: 315-331.

Harper, P. R. & Gamlin, H. M. 2003. Reduced outpatient waiting times with improved appointment scheduling: a simulation modelling approach. *OR Spectrum*, 25 (2): 207-222.

Hearst, M. A. (1999, June 20-26). Untangling Text Data Mining. *Paper presented at the Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics University of Maryland.*

Huang, T., L. Lan, X. Fang, P. An, J. Min and F. Wang (2015). Promises and Challenges of Big Data Computing in Health Sciences*. Big Data Research* 2(1): 2-11.

Jurka, T. P., Collingwood, L., Boydstun, A. E., Grossman, E., & van Atteveldt, W. (2012). RTextTools: Automatic Text Classification via Supervised Learning. *R package version 1.3.9.*

Korhonen, A., D. Ó Séaghdha, I. Silins, L. Sun, J. Högberg and U. Stenius (2012). Text Mining for Literature Review and Knowledge Discovery in Cancer Risk Assessment and Research*. PLoS ONE,* 7(4): e33427.

Lin, C., Hsu, C. J., Lou, Y-S., Yeh, S-J., Lee, C-C., Su, S-L., & H-C. Chen.(2017) Artificial Intelligence Learning Semantics via External Resources for Classifying Diagnosis Codes in Discharge Notes. *Journal of Medical Internet Research*, 19(11):e380. doi:10.2196/jmir.8344

Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. In I. Mani & M. T. Maybury (Eds.), *Advances in Automatic Text Summarization*, 15-22. Cambridge: The MIT Press.

Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., & Hurdle, J. F. (2008). Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of Medical Informatics*, 47(Suppl 1): 128-144.

Miller, J. M., & Velanovich, V. (2010). The natural language of the surgeon's clinical note in outcomes assessment: a qualitative analysis of the medical record. *The American Journal of Surgery, 199*(6): 817-822.

Mironczuk, M.M. & J. Protasiewicz (2018) A recent overview of the state-of-the-art elements of text classification. *Expert Systems With Applications* 106: 36-54

Mujtaba G, Shuib L, Raj R. G., Rajandram, R, Shaikh K, & M. A. Al-Garadi (2017) Automatic ICD-10 multi-class classification of cause of death from plaintext autopsy reports through expert-driven feature selection. *PLoS ONE* 12(2): e0170242. doi:10.1371/journal.pone.0170242

OED - Oxford English Dictionary Online. (2018). "semantics, n.". Oxford University Press. www.oed.com/view/Entry/345083 (May, 2018).

Pina, J., K. Chester, D. Danoff and M. Koyanagi (2013). Synonym-based Word Frequency Analysis to Support the Development and Presentation of a Public Health Quality Improvement Taxonomy in an Online Exchange. In C. U. Lehmann, E. Ammenwerth and C. Nøhr (Eds.), *Studies in Health Technology and Informatics*, MEDINFO. 192:1128-1128.

R Core Team (2013). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. http://www.R-project.org/.

Rebholz-Schuhmann, D., Kirsch, H., & Couto, F. (2005). Facts from text—Is text mining ready to deliver? *PLoS biology, 3*(2): 65. doi: 10.1371/journal.pbio.0030065

Scott, D., Hallett, C., & Fettiplace, R. (2013). Data-to-text summarisation of patient records: Using computer-generated summaries to access patient histories. *Patient Education and Counseling, 92*(2) 153-159.

Sharma, A. & Mansotra, V. (2014) Emerging applications of data mining for healthcare management - A critical review, *Computing for Sustainable Global Development (INDIACom)*, New Delhi: 377-382.

Sokolova, M. and G. Lapalme (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management,* 45(4): 427-437.

Spasić, I., Livsey, J., Keane, J. A., & G. Nenadić (2014). Text mining of cancer-related information: Review of current status and future directions. *International journal of medical informatics,* 83(9): 605-623.

Srinivas, S. and A. R. Ravindran. (2017) Systematic Review of Opportunities to Improve Outpatient Appointment Systems. Proceedings of the IIE Annual Conference; Norcross (2017): 1697-1702.

Srinivas, S. and A. R. Ravindran. (2018) Optimizing outpatient appointment system using machine learning algorithms and scheduling rules: A prescriptive analytics framework. *Expert Systems with Applications*, 102: 245-261.

Wales Audit Office (2016) Follow-up Outpatient Appointments – Summary of Local Audit Findings. *Briefing Paper for the NHS Wales Planned Care Programme Board*. Published by the Auditor General for Wales, May 2016. www.audit.wales/system/files/publications/outpatients-follow-up-briefing-paper-2016-english.pdf (May 14, 2018).

Wu, J.-S., Kao, E.-F., & Lee, C.-N. (2014). Discovering Hidden Connections among Diseases, Genes and Drugs Based on Microarray Expression Profiles with Negative-Term Filtering. *PLoS ONE,* 9(6).