Data Min Knowl Disc (2012) 24:195–217
DOI 10.1007/s10618-011-0225-y

Data mining from a patient safety database: the lessons learned

James Bentham · David J. Hand

Received: 21 December 2009 / Accepted: 23 May 2011 / Published online: 9 June 2011

Abstract The issue of patient safety is an extremely important one; each year in the UK, hundreds of thousands of people suffer due to some sort of incident that occurs whilst they are in National Health Service care. The National Patient Safety Agency (NPSA) works to try to reduce the scale of the problem. One of its major projects is to collect a very large dataset, the Reporting and Learning System (RLS), which describes several million of these incidents. The RLS is used as the basis for research by the NPSA. However, the NPSA has identified a gap in their work between high-level quantitative analysis and detailed, manual analysis of small samples. This paper describes the lessons learned from a knowledge discovery process that attempted to fill this gap. The RLS contains a free text description of each incident. A high dimensional model of the text is calculated, using the vector space model with term weighting applied. Dimensionality reduction techniques are used to produce the final models of the text. These models are examined using an anomaly detection tool to find groups of incidents that should be coherent in meaning, and that might be of interest to the NPSA. A three stage process is developed for assessing the results. The first stage uses a quantitative measure based on the use of planted groups of known interest, the second stage involves manual filtering by a non-expert, and the third stage is assessment by clinical experts.

Responsible editor: Eamonn Keogh.

J. Bentham (✉)
Department of Medical and Molecular Genetics, King's College, London, UK
e-mail: james.bentham@kcl.ac.uk

D. J. Hand
Department of Mathematics, Imperial College, London, UK

D. J. Hand
Institute for Mathematical Sciences, Imperial College, London, UK

Keywords Knowledge Discovery · Data Mining · Information Extraction · Patient Safety

1 Introduction

Every year in England and Wales, hundreds of thousands of National Health Service (NHS) patients experience an incident that causes them harm. These incidents include falls, problems during childbirth, mistakes made during operations and medication errors. Whilst most of these incidents do not cause permanent harm to patients, some can cause serious injury and even death. Given the human and financial cost of these incidents, the issue is clearly an important one for the British government. In the Year 2000 a working group was set up to examine the problem in depth; this group's recommendations (Department of Health 2000) led to the creation of the National Patient Safety Agency (NPSA) in 2001.

The NPSA works in various ways to try to reduce the numbers and severity of patient safety incidents. One major project that the NPSA carries out is to collect data describing the incidents, which is then used as the basis for research. The datasets are actually collected by individual NHS organisations, and then amalgamated to form the Reporting and Learning System (RLS), a database that contains several million entries. It has 73 variables, the majority of which are categorical and numerical, describing the incident type, severity, medication details (where appropriate) and so on. As well as these variables, each incident is described by a free text description, generally written by NHS staff. Although this is a potentially very rich source of information, the size of the dataset causes difficulties with the analysis, and the complexity of natural language means that any numerical analysis is quite difficult to carry out. This has meant that to date most of the NPSA's analysis work has focused on producing high level summary statistics describing the data, or on detailed, manual qualitative analysis of the data that is limited in breadth.

This paper describes an attempt to fill the gap between these two types of analysis, using a knowledge discovery process. The aim was to create a numerical representation of the incident descriptions, which would reflect their meanings to a sufficient extent that when an anomaly detection tool was used to find groups of incidents that were more tightly packed than expected, the groups identified would be coherent in meaning, and would comprise incident types that might be unknown to both the NPSA and the NHS as a whole.

In order to assess whether the groups were in fact novel and interesting to the NPSA, it was necessary to develop an assessment method that was appropriate for this particular knowledge discovery process. This method used the specific resources available in a way that meant that the NPSA's clinical experts, who were available only for a relatively short amount of time, could assess results of as high a quality as possible.

2 Knowledge discovery process

The past thirty years have seen a transformation in many aspects of everyday life due to developments in Information Technology. These developments have allowed

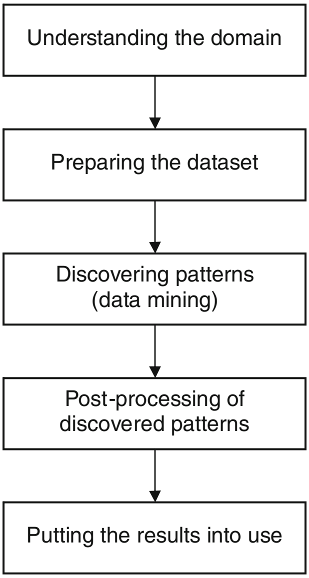

Understanding the domain
Preparing the dataset
Discovering patterns (data mining)
Post-processing of discovered patterns
Putting the results into use

Fig. 1 A characterisation of the KDD process

us to collect increasingly large amounts of data. However, our ability to analyse and understand these datasets lags behind our ability to gather the data (Fayyad 1996). In order to try to extract useful information from these datasets, the field of Knowledge Discovery in Databases (KDD) has emerged. We use the term KDD to describe the entire process of collecting a large dataset, preprocessing it, analysing it in some way, and finally extracting information from the results that is interesting and useful; the term *data mining* is used to describe the component of the KDD process that involves discovering patterns in the data.

Various characterisations of the KDD process have been made; we feel that the steps described by Mannila (1996) most closely resemble those used in our process. They are shown in Fig. 1.

3 Understanding the domain

In order for us to understand better the domain and then prepare the dataset, it was necessary that we should examine the data in detail.

3.1 RLS free text descriptions

Although the free text descriptions are a potentially extremely rich source of information, some of their characteristics make analysis more complicated. The examples shown in Table 1 illustrate some of these difficulties.

The descriptions are of varying length and quality. The median length of the descriptions is around 20 words, but the range of lengths is quite large. The minimum length is a single ‘word’, but this may be a full stop, or a series of ‘x’ characters.

Table 1 Examples of RLS free text descriptions

Patient was due to go to theatre, following an EPAU appointment for? ectopic pregnancy. Patient was supposed to be on emergency list. Contacted theatre 16.30 to see what time patient would be going. Patient not listed for theatre. Reg on call informed.
Pt. attended ANC following being seen on Ward 26 via A&E on 17.04.05 No record in notes today re being seen on Ward 26 or A&E. On investigation appropriate information had been filed in another pt notes Unit No : span class="Number "onmouseover="doHover(this);" onMouseOut="doUnHover(this);" 2208595N :b Number /b /span indicating that wrong pt notes had been taken to LWS.
PHENOBARBITAL DOSE CHECKED ON DRUG CHART NO CD IN PHARMACY BOX DRUG BOXES CHAECK POM GIVEN AS PRESCRIBED TO PATIENT IMMEDIATELY CHECKED IN BNF AS STILL UNSURE STATES CD BUT PATIENT HAD ALREADY TAKEN TABLETS. 3 OPENED BOXES IN BAG FROM KINGFISHER S/N HT INFORMED.
The patient suffers from a rare condition—Sjogren Syndrome, and is taking a cocktail of drugs to keep his symptoms under control. He was admitted on the 28/10/05 as an emergency, with an acute exacerbation and considerable pain. He had not brought his very specific medicines with him. They were prescribed on 29/10/05 and the 30/10/05. At 15.00 h on the 30/10/05 the Staff Nurse contacted the Lead Nurse, as the drugs had not all been dispensed. She contacted the on—call Pharmacist, who refused to come in and dispense the remaining drug required. The Senior House Officer was contacted, who also contacted the Pharmacist at home, to no avail. This resulted in a very ill man being without his essential medicines for two days!
Found on the floor.

The longest descriptions are several hundred words long, comprising several paragraphs. The sophistication of the language used varies: some descriptions use medical terminology to describe complex illnesses or procedures, whilst others are more colloquial, and may not in fact describe patient safety incidents at all.

The formatting of the descriptions varies between entries, which may be partly because the RLS is an amalgamation of individual NHS organisations' data. Some of the descriptions are all in upper case, and there are superfluous spaces before commas and full stops, both of which would make finding sentence boundaries more difficult. There are no pound (sterling) signs, and no consistent formatting of times or dates. At some point, either in the organisations' or NPSA's systems, some of the entries have become corrupted, and contain text such as '*onmouseover=doHover(this)*'.

Another issue is that the spelling is of variable quality. Spelling mistakes are of two main types: misconceptions of the correct spelling, and typographical errors. Problems may also caused by the systems that the NHS organisations use for collecting the data. In many cases, the staff write the incident descriptions on paper forms; the descriptions are then transferred to the database by clerical staff. This can lead to errors being made, particularly where complex medical terminology is used; the handwriting is presumably of varying legibility.

A further issue that makes the analysis of the free text descriptions more difficult is the use of initialisms, acronyms and terminology that are specific to the NHS: many of the descriptions contain acronyms, such as 'NICE' (National Institute for Clinical Excellence). There are also conventions such as writing 'RIP' for 'dead' or 'death'. Terminology that is specific to the NHS includes initialisms describing organisations, departments and staff grades. It will be seen below that for a model that treats different words as orthogonal dimensions in a Euclidean space, these characteristics of the free text descriptions can potentially cause a reduction in the quality of the model. General features of natural language, such as homonyms and synonyms cause similar problems.

3.2 RLS categorical and numerical variables

Whilst the RLS free text descriptions are the focus of our work, the database also contains categorical and numerical variables describing the incidents, which are a further important source of information. These variables describe characteristics of the incidents such as their type, location and severity, details about the patients and staff involved, times and dates, and, where appropriate, details of the medication or devices involved in the incidents.

The way in which the RLS was created, as an amalgamation of individual NHS organisations' data, has affected both the proportion of incidents for which entries are made for certain variables, and the structure of the database. The organisations use various different systems for collecting patient safety data, which meant that the NPSA had to create a complex mapping between the organisations' variables and the RLS variables. Given that the organisations do not all collect the same information, making an entry for some of the RLS variables is voluntary; this means that some variables have very few incidents with entries. Other variables, such as the free text descriptions, incident type and severity are mandatory.

Most of the categorical variables are nominal, but one of the most important variables is ordinal. This is the degree of harm variable. The NPSA is most interested in incidents involving deaths and severe harm; the other categories are moderate, low and no harm. This distinction has changed over the lifetime of the RLS: the NPSA

Table 2 RLS categorical variables used for supervised dimensionality reduction

Var code	Variable	Description	Examples	No. of classes
1	IN03 Location, Level 1	In which location did the incident occur?	General/acute hospital Mental health unit	12
2	IN03 Location Level 2		Dental surgery Outpatient department	26
3	IN05 Incident Cat, Level 1	Please categorise the patient safety incident	Medical device Patient accident	16
4	IN05 Incident Cat, Level 2		Diagnosis—wrong Slips, trips, falls	87
5	PD05 Specialty, Level 1	Please indicate the specialty or service	Surgical specialty Mental health	16
6	PD05 Specialty, Level 2		Gastroenterology Haematology	83

was initially most interested in incidents involving deaths, severe harm *and* moderate harm.

Six of the RLS categorical variables were of particular interest during this project. They are described in Table 2, and involve incident type, location and the specialty (i.e., area) of the NHS, where the incident took place. It will be seen below that these variables were of use as response variables when carrying out supervised dimensionality reduction of the vector space model.

We found that using the categorical variables as the basis for rotations of model spaces representing the text was a better way of incorporating that information than trying to use the categorical variables directly to model the incidents. The free text descriptions are a rich source of information, but have the disadvantage that our numerical model maps to their meanings in a complicated manner. The categorical variables provide a coarser differentiation between incidents than the free text, but are related to meaning in a more straightforward manner. Rotating the free text model space based on the categorical variables therefore provides a way to use the characteristics of the two types of data in a complementary manner.

4 Preparing the dataset

In order to be able to analyse the RLS free text descriptions, it was necessary to model them numerically. However, the preparation of the dataset was constrained by the data mining method to be used later in the KDD process, i.e., anomaly detection in a low dimensional Euclidean space. This meant that any Natural Language Processing methods that were used had to be appropriate for this type of analysis, and that the model space needed to be of sufficiently low dimensionality that problems caused by both the curse of dimensionality and limitations on computing resources were avoided.

4.1 Vector space model

There was no obvious way to place the free text descriptions into a low dimensional space directly such that their positions reflected their meanings; the approach that was adopted was to produce a high dimensional model of the text, and then reduce its dimensionality in such a way that the most interesting information was retained.

The basic model that was used was the vector space model (Salton et al. 1975), where each dimension in the vector space corresponds to a particular word, and the position of an entry in a particular dimension is a function of the number of times the corresponding word appears in that piece of text.

The vector space model is a simplification; only a small amount of the complex information carried by natural language is captured by the model. No account is taken of word ordering, any information carried by grammatical constructions is lost, and relationships in meaning between words are not captured.

However, we had to strike a balance between a model that is sufficiently complex that it captures enough information to produce a meaningful representation of the text, and one that is simple enough to be practicable.

One other defect of the model is that it treats each word as being of equal importance to the meaning of the incident description. A more realistic model of meaning would stretch the vector space, to attach more importance to more meaningful words, such as ‘patient’ and ‘fell’. This stretching of the vector space can be achieved using term weighting.

The term weighting method that was used in this analysis was the well known TF-IDF method (Salton and McGill 1983). This method relies on the intrinsic properties of the text, and is applicable to any sample of text that can be divided into separate units, such as incident descriptions.

The words are weighted in two ways. The first is the *TF*, or *term frequency* weighting, which is related to the frequency of the appearance of a word in a particular incident description. The second is the *IDF* or *inverse document frequency* weighting, which is related to the number of descriptions in which a word appears.

As with much of the work described in this paper, technically more sophisticated methods are available for term weighting. For example, Zhang et al. 2008 describe term re-weighting based on ontologies. However, for a practical KDD process such as this, we had to balance the possible (but not certain) benefits of using more cutting-edge methods against the costs in terms of time and effort of adding them to the analysis.

4.2 Dimensionality reduction

The data samples analysed each contained 25,000 incident descriptions. For samples of this size, the vocabulary was around 20,000 words, which meant that the corresponding vector space was ca. 20,000 dimensional. This was far too large to be analysed using an anomaly detection algorithm.

One reason why this is not feasible is that the behaviour of high dimensional spaces is rather different to the behaviour of more familiar two or three dimensional spaces.

As the dimensionality of these spaces increases, the data become increasingly sparse, and it can be difficult to discriminate between nearest and furthest neighbours. This can make concepts of density and anomalies difficult to define, and means that in high dimensional spaces the number of data points required to find a meaningful measure of density increases exponentially. The behaviour of high dimensional spaces is examined in depth in Aggarwal et al. (2000) and Hinneburg et al. (2000) using both theoretical and empirical results. We avoided this issue by reducing the models' dimensionality sufficiently that no problems were encountered.

A practical problem with high dimensional datasets is that they contain an enormous amount of information, much of which may not be particularly useful or interesting. A further potential advantage of dimensionality reduction is that it can allow uninteresting data to be discarded. For the RLS data, it appeared that the meanings of the descriptions were more closely related to position in the low dimensional spaces: the rotation revealed the more meaningful dimensions in the vector space model of the text, and non-meaningful dimensions were discarded.

We used both unsupervised and supervised dimensionality reduction methods. For unsupervised methods, the rotation is carried out based on the intrinsic properties of the data. Supervised methods rotate the model space based on the relationship between the data and information such as a categorical, or response, variable.

The unsupervised and supervised methods that we used were the extremely well-known principal component analysis (PCA) and linear discriminant analysis (LDA). PCA rotates the space so that the first dimension of the rotated space explains the maximum possible amount of variance in the data. Having fixed this dimension, the remaining subspace is rotated in a similar manner, and the process continues. LDA uses information that divides the data points into various groups, such as the incident types taken from the RLS categorical variables. The model space is rotated such that the ratio of the between groups sum of squares to the within groups sum of squares is maximised. Once the first dimension is fixed, the remaining subspace is rotated in a similar manner.

Whilst it is not necessary to carry out unsupervised dimensionality reduction before calculating linear discriminants, we found that it was beneficial. This is also reported in some of the literature (for example, Manning et al. (2008)) describing analysis of internet search engines. We make the conjecture that earlier principal components represent combinations of words that are meaningful in natural language, whereas later dimensions are combinations of words that appear together by chance. A model using a fairly small proportion of the principal components is therefore a better representation of natural language than the full vector space model.

An advantage of PCA compared with other methods such as factor analysis or projection pursuit is that it is *not* scale invariant, allowing the term weights to be included in the model. This property of PCA is not always considered beneficial, but is so in this case.

Again, PCA is by no means a complex method, and it would have been possible to use other, more sophisticated techniques: for example, Independent Component Analysis (Hyvärinen et al. 2001), taken from the Signal Processing literature, or Semantic Hashing (Salakhutdinov and Hinton 2009) from Natural Language Processing.

However, we continued to view the use of methods of low complexity as being of potential overall benefit to the KDD process.

Indeed, despite the simplicity of the basic model, the term weighting and the dimensionality reduction, a low dimensional model was produced that reflected meaning sufficiently well that coherent groups could be found. Nevertheless, it is clear that the true relationship between the vocabulary used and the meaning of natural language is extremely complex. This suggested that a more flexible supervised dimensionality reduction method, that produced non-linear combinations of the original variables, such as neural networks (Ripley 1996), might produce superior models of the text. However, analysis carried out using neural networks produced extremely poor results. We suggest that this is due to overfitting; whilst the mapping from vocabulary to meaning is more complex, it is less accurate, and the simple linear combinations produce a better approximation of the true relationship.

We also considered using other methods taken from the Natural Language Processing literature. One possible method that could have been used is parsing, where algorithms are used to analyse the syntax of text. A concern that we had was that the parsers would not be accurate when used with the RLS free text descriptions, given their relatively poor quality. However, work has been carried out that examines the performance of parsers using Wikipedia entries (Honnibal et al. 2009) and grammatically noisy text (Foster et al. 2008). A further concern was that the parser might find it difficult to identify the many domain-specific terms used in the RLS descriptions; again, however, work has been carried out in this area (Lease and Charniak 2005). Nevertheless, we did not use parsing in this analysis. This is because of the nature of our model, which is a Euclidean space. Any Natural Language Processing method would have to either stretch or rotate this space, and it is difficult to see how this could be done easily using syntactic information; the vector space model does not allow for this structure to be modelled.

The ca. 20,000 dimensional TF-IDF weighted vector space model was reduced to 2,000 principal components. This 2,000 dimensional dataset was used to calculate linear discriminants. Up to 15 linear discriminants were retained; the models produced using Response Variable 1 (see Table 2) were 11 dimensional. These 11 or 15 dimensional models were the final numerical representations of the text. Having prepared the dataset, an anomaly detection tool could be used to try to find incident descriptions of anomalously similar meanings, which are the patterns of interest for this KDD process.

5 Data mining

We decided following discussions with the NPSA about the type of results that they would find most useful that the most appropriate method for data mining of the RLS was anomaly detection. This type of analysis is similar to non-hierarchical clustering methods, such as the k-means algorithm (MacQueen 1967), where groups of data points that are particularly similar to one another are assigned to a cluster. However, in anomaly detection the groups of data points that are being sought are small compared with the overall size of the dataset, and most of the data will not be labelled as being part of a potentially interesting group. Whilst methods for clustering medical documents have been developed (Saad and de la Iglesia 2006), we are not aware of any work using anomaly detection algorithms.

Several algorithms are available that may be used for anomaly detection, such as DBSCAN (Ester et al. 1996). We again chose an algorithm, PEAKER (Zhang and Hand 2005), that was as simple as possible. This lack of complexity reduced the amount of time needed to run the algorithm, which was imperative given the large number of parameter combinations (see below) to be examined. The parameters controlling the density estimates and the size of the anomalies were easy to interpret and control, and, crucially, easy to explain to the NPSA, who had to give feedback on the results. The potential difficulties involved in explaining statistical algorithms to non-statisticians during the KDD process should not be underestimated.

Other methods that might have been used include mixture models, where the data are assumed to be generated from a number of distributions. Inferences about the underlying distributions can be drawn from the data, and used to find interesting patterns. However, a method such as the Infinite Gaussian Mixture Model (Rasmussen 2000), uses Markov Chain Monte Carlo (MCMC) methods, which are extremely computationally expensive. This makes them inappropriate for a KDD process of this sort, where very large numbers of runs of the data mining algorithm are to be made.

PEAKER works in a straightforward way, starting by calculating an empirical estimate of the density at each point. The algorithm then looks for ‘peaks’, where a peak is a point that has an estimated density greater than surrounding points. The algorithm returns the peaks and their nearest neighbours as the points of potential interest.

Given a population of data points, P, where each point is labelled $\mathbf{x_i}$, $i \in P$, the estimate for the density at each point $\mathbf{x_i}$ is

$$\hat{f}(\mathbf{x_i}) = \hat{f}(\mathbf{x_i}; K) = \left[\frac{1}{K} \sum_{j \in \{N\}} d(\mathbf{x_i}, \mathbf{x_j}) \right]^{-1}$$

where

$$\{N\} = \{\mathbf{x_j} : \mathbf{x_j} \in N(\mathbf{x_i}; K)\}$$

i.e., N is the set of K nearest neighbours to $\mathbf{x_i}$. The smoothing of the density estimate is therefore varied by increasing or decreasing the value of K. It is possible to use any distance measure to calculate $d(\mathbf{x_i}, \mathbf{x_j})$: the Euclidean and Manhattan distance metrics are two well known measures.

For each point $\mathbf{x_i}$, the point is defined as a peak with

$$M(\mathbf{x_i}) = m$$

iff

$$\hat{f}(\mathbf{x_i}) > \hat{f}(\mathbf{x_j}), \quad \forall \mathbf{x_j} \in N_m(\mathbf{x_i})$$

and

$$\hat{f}(\mathbf{x_i}) \leq \hat{f}\left(x^{(m+1)}\right)$$

where N_m is the set of m nearest neighbours to $\mathbf{x_i}$, defined using some distance metric, and $\mathbf{x^{(k)}}$ is the kth nearest neighbour to $\mathbf{x_i}$. Using the same distance metric to calculate the density estimates and M values increases the computational efficiency of PEAKER.

The value m describes the size of the group of points associated with a peak; it is the number of neighbouring points that have lower density estimates than the peak in question. For a knowledge discovery process using real data, there might be practical limits to the minimum and maximum sizes of interesting groups ($M_{\min}$ and $M_{\max}$): for example, a very small group of incidents might not provide sufficient information for the inference to be drawn that those types of incidents were occurring systematically in the NHS. Conversely, very large groups might be difficult to assess qualitatively as there might be too much information for an expert to assess manually. Following discussions with the NPSA, the values of $M_{\min}$ and $M_{\max}$ were set at:

- $M_{\min} = 5$
- $M_{\max} = 99$

The groups identified, including the peaks, were therefore of between 6 and 100 incidents.

The desired result was that the groups should be coherent in meaning, describing similar patient safety incidents, and that they should use different vocabulary to avoid being trivial. It was also necessary that they should represent information that was useful, novel and interesting to the NPSA, i.e., small groups of incidents of novel types, which the NPSA would act on by producing advice or instructions to be distributed around the NHS.

An example of one of the groups found using data mining is shown in Table 3. This group is typical of many of the groups that were identified: it comprises incident descriptions that use varying vocabulary to describe what appear to be similar issues, which suggests that the group is coherent in meaning (note the spelling mistake '*surery*' in the fourth description). In this case the descriptions describe problems that occurred during surgical procedures associated with missing or incorrect notes.

We have therefore shown that it is possible to build a low dimensional model of the RLS free text descriptions that can reflect meaning, rather than simply vocabulary. This is despite (or possibly because of) the fact that the model, term weighting, classification techniques and anomaly detection algorithm are all straightforward and/or well-known methods.

However, no obvious method was available for assessing the results of the modelling process; this meant that a novel assessment, or post-processing, method had to be produced for this particular KDD process.

Table 3 Group of incidents identified in the surgical specialty sample

Patient coming for surgery 2/2/06. Notes missing. Rang 4987, last dept to have notes. Notes had been sent in post 31/1/06, didn't get to Eye Unit until late afternoon 2/2/06. Patient surgery cancelled for 2/2/06, no notes.
2 different case sheet numbers for patients with same name and same consultant. Wrong number on list for patient in theatre. Problem noted when getting blood.
Pt planned to go on trauma list, all prepared as per protocol. Pt notes had been requested but not available for anaesthetist. Admissions and medical records aware of urgency for notes. Theatre cancelled in view of medical history and no notes available.
Patient taken to theatre for surery on morning list. ODP noticed that another patient labels were enclosed in the notes.
Patient admitted for surgery 10/8/06, preassessed at . . ., notes were then put in for transfer and booked to ward 28, notes never arrived, unable to find them anywhere, patient surgery cancelled.
Above patient arrived on unit for an endoscopic procedure. When checking notes it was found that the patient had been sent to us with another patients notes. Ward informed and correct notes brought down.
Notes and theatre list for tomorrows list sent from SJUH. On checking list and notes the name and unit number on list did not match one of the sets of notes. On further investigation the name and procedure on the list were correct but the unit number and date of birth were for another patient. The notes sent were for this patient but were not for the named patient on the list and needed a different procedure.

6 Three stage post-processing process

6.1 Introduction

At the start of the KDD process it was not clear what the nature of the results produced by the data mining algorithm would be, or what resources would be available for assessing them. This is likely to be generally the case for KDD processes, particularly those of a longer duration, where the assessment process will develop gradually as the process progresses.

The resources that were ultimately available to assess the results were limited. The NPSA provided six clinical experts, who spent a day each analysing the results manually and qualitatively. Whilst this was a generous allocation of resources by the NPSA, it meant that it was necessary that the experts should be given results that were as coherent as possible, and that were likely to be of novel and interesting types.

A PhD student (the first author) was available, who spent much more time carrying out non-expert qualitative analysis of the results. The volume of results that could be analysed was therefore larger, but the expertise of the assessment was much more limited.

Finally, quantitative analysis of the results could be carried out. In this case, it was necessary to develop a quantitative measure that could act as a quasi-qualitative measure of the quality of the results. The quality of this part of the assessment was therefore dependent on finding an appropriate measure that could be used to assess very large volumes of results.

Our process therefore had three stages:

1. Use of a quantitative measure using planted groups
2. Manual filtering
3. Assessment by clinical experts

There was a large number of parameters that could be varied during the modelling process, including the number of principal components to be retained or the function used to transform the IDF term weighting, for example. The way that we proceeded was to find some sort of ‘optimum’ parameter combination using the quantitative measure, and then use it to calculate models for six datasets corresponding to areas of expertise of clinical experts at the NPSA. Around 200 to 300 groups per dataset were identified by PEAKER, which was too many to be assessed by the clinical experts in the time available; a non-expert qualitative filtering of the results was therefore carried out, to remove groups that were obviously incoherent or uninteresting. This also increased the quality of the results passed to the experts. Finally, the approximately 100 remaining groups per sample were each passed to a clinical expert for qualitative assessment.

6.2 Quantitative measure

The aim of this part of the process was to find a combination of parameters that produced a high quality model of the text, with respect to a particular measure of model quality. The NPSA provided seven groups that were similar to those that they were

Table 4 Known groups used with the quantitative measure

Incident type	Code	Size
Anaesthetics	An	100
Chest Drains	Ch	7
Latex	Lt	7
Methotrexate	Mt	5
Obstetrics	Ob	100
Self Harm	Sl	100
Sexual Safety	Sx	100

trying to find in the RLS; the groups had been found during previous manual analysis of the database. The groups of incidents are described in Table 4. These 419 incidents were combined with a random sample of 3,000 incidents. Models were calculated using each parameter combination of interest and PEAKER was run; the results were examined to see whether the known groups had been found.

The assumption that was made was that if the model was of sufficient quality that it reflected the meanings of the incident descriptions, incidents of similar meaning, such as those from the known groups, would be close together in the model space, and those groups would therefore be identified by PEAKER. The number of known groups found could therefore be taken as an *indication* of the quality of a model produced by a particular parameter combination.

The groups were defined as being 'found' if the following conditions were met:

– A group was found that contained at least six incidents
– At least 50% of the incidents in the group came from the known group of incidents
– There were more incidents in the group from that known group than any other known group

Clearly, we had to set these requirements arbitrarily.

An example of the type of results found using the quantitative method is shown in Table 5 (the response variables are described in Table 2). In this case, the quantitative measure has been used to compare the results for two different IDF normalisation functions. The results in Table 5 suggest that the square root normalisation function produces superior models to the logarithm, as generally more of the known groups were found. The quantitative measure appears to be quite stringent, at least for the vector space model; we have found that parameter combinations for which only one or two known groups can be found have produced models that contain large numbers of coherent, potentially interesting groups of incidents.

This method is dependent on small groups of this type being available, and will therefore only be useful for certain projects. There is also clearly scope for improving the method. For example, numerical measures of how tightly clustered the descriptions from the planted groups were could have been used instead of presence/absence indicators. It would also have been possible to examine the remaining unlabelled data to see how well it had been modelled, perhaps using a semi-supervised learning algorithm such as that described by Blum and Mitchell (1998).

Table 5 Comparison of normalisation functions using the quantitative measure

Normalisation function	Response variable	Known groups found						
Function	Variable	An	Ch	Lt	Mt	Ob	Sl	Sx
Square root	1	–	–	–	–	✓	–	✓
	2	–	–	✓	–	–	–	✓
	3	–	–	✓	–	✓	✓	✓
	4	–	–	–	–	–	–	✓
	5	–	–	✓	–	✓	–	✓
	6	–	–	✓	–	✓	✓	✓
Logarithm	1	–	–	–	–	–	–	–
	2	–	–	–	–	–	–	–
	3	–	–	–	–	–	✓	✓
	4	–	–	–	–	✓	✓	✓
	5	–	–	–	–	–	–	–
	6	–	–	–	–	–	–	–

However, using any quantitative measure is quite a crude way to evaluate the optimum parameter combination. This optimisation, which aims to produce the most meaningful possible model of the text, appears to be a very difficult problem. In the first place, there is a large number of parameters, which may well interact in a complicated manner. This means that a more sophisticated method for optimising the parameters would also increase the overall complexity of the KDD process.

6.3 Qualitative filtering method

Having found the ‘optimum’ parameter combination using the quantitative, planted groups method, models were calculated for six datasets corresponding to the areas of expertise of the NPSA’s clinical experts. Each of the six RLS response variables was used to calculate linear discriminants, to examine the relationship between the results produced by the quantitative measure and the number of coherent, potentially interesting groups identified using the filtering method.

The six samples were taken from the following areas:

- Medical Devices (Med)
- Surgical Specialty (Sur)
- Treatment Procedure (Tre)
- Diagnostic Services (Dia)
- Accident and Emergency (AE)
- Deaths and Severe Harm (DS)

The non-expert filtering of the results was then carried out. The groups were placed into four broad categories:

- A: Coherent, using varying vocabulary

- B: Coherent, potentially interesting, but using similar vocabulary
- C: Coherent, but already known and using similar vocabulary
- D: Incoherent

Coherence was assessed subjectively; in the case of any doubt, the groups were placed into the highest plausible group. This was to try to avoid any potentially interesting groups being removed at this stage. However, any doubt was relatively rare; it seems to be possible for a non-expert to understand most of the descriptions (although we may have missed some of the subtleties or underlying issues). It will be seen that the requirements for a group of incidents to be novel and interesting to the NPSA are so stringent that we are confident that no novel groups were removed at this stage.

The distribution of the groups across the four categories gives an indication of the quality of that model, allowing different models for the same sample to be compared. This makes it possible to select the 'best' response variable for each sample, and pass the groups from Category A for that response variable to the experts. For example, if one model produces no coherent groups and another produces fifty, it is assumed that the latter model is superior. The results of this analysis are shown in Table 6.

It can be seen that certain response variables generally produced more groups in Category A: as might be expected, carrying out dimensionality reduction using incident type as a response variable produced the most meaningful models, although obstetrics incidents appear to be better classified by the categorical variable related to specialty (Variable 5 in Table 2). There were also differences between the different samples: some produced far more groups in Category C, presumably because these datasets contained more incident descriptions that were written using NHS organisations' proformas. The deaths and severe harm dataset produced more groups in Category D; this dataset comprised descriptions that were generally longer, more complex and of more disparate types than the other samples, and were therefore more difficult to model numerically.

For each sample, the groups from Category A for the response variable that produced the highest proportion of coherent, potentially interesting groups were selected for assessment by the clinical experts.

6.4 Expert assessment

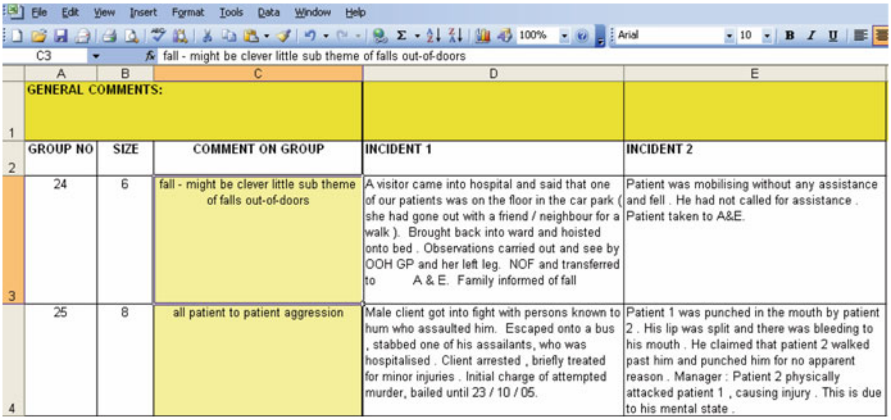
The third and final stage of the process was for the clinical experts at the NPSA to assess qualitatively the selected groups. The first step was to meet the experts to describe the project and the feedback that was required from them. They were each provided with a spreadsheet containing the groups; a screenshot from the Deaths and Severe Harm sample in shown in Fig. 2.

The experts were asked to provide a free text description of each of the groups, describing whether they were coherent, and if so, what type of incident they described and whether the NPSA might need to take further action regarding that incident type. They were also provided with an opportunity to make a general overall comment on the groups.

The experts found that, depending on the sample, between 70 and 96% of the groups were actually coherent. Follow-up meetings were then held with three of the experts,

Table 6 Distributions of groups using filtering method to examine subsets of data

Category	Variable	Datasets					
		Med	Sur	Tre	Dia	AE	DS
A	1	8	0	28	58	1	17
	2	52	24	14	35	7	36
	3	–	97	–	71	106	64
	4	48	109	70	139	152	89
	5	66	–	42	–	–	57
	6	70	39	83	81	–	65
B	1	0	0	0	0	0	0
	2	0	0	0	3	0	0
	3	–	0	–	6	0	0
	4	0	0	4	2	0	0
	5	0	–	0	–	–	0
	6	0	0	0	0	–	0
C	1	8	65	58	102	20	12
	2	14	85	76	77	31	18
	3	–	180	–	130	111	53
	4	13	162	83	181	75	66
	5	19	–	90	–	–	32
	6	7	113	83	111	–	29
D	1	134	204	141	125	171	121
	2	25	72	54	94	85	56
	3	–	13	–	15	12	28
	4	802	48	45	21	42	54
	5	36	–	49	–	–	52
	6	67	56	39	95	–	84



C3 fall - might be clever little sub theme of falls out-of-doors

	A	B	C	D	E
1	GENERAL COMMENTS:				
2	GROUP NO	SIZE	COMMENT ON GROUP	INCIDENT 1	INCIDENT 2
3	24	6	fall - might be clever little sub theme of falls out-of-doors	A visitor came into hospital and said that one of our patients was on the floor in the car park (she had gone out with a friend / neighbour for a walk). Brought back into ward and hoisted onto bed . Observations carried out and see by OOH GP and her left leg. NOF and transferred to A & E. Family informed of fall	Patient was mobilising without any assistance and fell . He had not called for assistance . Patient taken to A&E.
4	25	8	all patient to patient aggression	Male client got into fight with persons known to hum who assaulted him. Escaped onto a bus , stabbed one of his assailants, who was hospitalised . Client arrested , briefly treated for minor injuries . Initial charge of attempted murder, bailed until 23 / 10 / 05.	Patient 1 was punched in the mouth by patient 2 . His lip was split and there was bleeding to his mouth . He claimed that patient 2 walked past him and punched him for no apparent reason . Manager : Patient 2 physically attacked patient 1 , causing injury . This is due to his mental state .

Fig. 2 Screenshot from an expert's spreadsheet

following which final conclusions about the knowledge discovery process could be made.

7 Putting the results into use

For the KDD process to be completely successful, it was necessary that the results found to be coherent and potentially interesting by the clinical experts should lead to further action by the NPSA. We were told by the experts and other staff at the NPSA that had the results been available in the first two years of the KDD process, some of the groups identified would have been interesting to the NPSA, and would probably have triggered further action. However, by the time the results became available, none warranted further action by the NPSA, and so ultimately the results were not put into use.

There are various explanations for this. When the process started, the NPSA had only been in existence for two years; as it has matured, it has developed many different routes by which it finds out about novel types of patient safety incidents: e.g., coroners' reports, research by health professionals or reports by the Chief Medical Officer. It is therefore much more difficult for novel incident types to be found in the RLS than was envisaged at the start of the project.

The resources that the NPSA has for producing advice and instructions for dissemination to the NHS are much more limited than is anticipated in Department of Health (2000). A novel incident type must be extremely serious for the NPSA to take action. Again, this limits the number of potential groups to be found. The NPSA is now only likely to take action if incidents involve severe harm or death, rather than looking at incidents involving moderate harm (or, as suggested by Department of Health 2000, near misses); this makes it feasible for these incidents to be analysed manually by the clinical experts at the NPSA, particularly strictly in the case of incidents involving death. This means that the data from which potential groups might be drawn is both much smaller and much better understood than thought at the start of the process.

However, there is a possibility that the process could be put to use in a different manner. We were told by the experts that the output from our KDD process was similar to the results found from manual, thematic analysis of similar data samples, which is carried out by clinical experts. In this case, the experts read a large number of incident descriptions from a particular area, and use their knowledge to draw out the, say, 10 to 20 main types of incidents from the sample. It is encouraging that the two types of analysis reinforce each other's findings.

8 Discussion

Whilst the data mining process has not identified any novel groups of incidents that triggered action by the NPSA, we feel that the KDD process has been generally successful. We have developed meaningful models of the free text descriptions, have developed a novel post-processing method, and have learned various lessons about carrying out a KDD process.

8.1 Modelling free text

One of the main achievements of the work has been to show that it is possible to produce a numerical model of free text descriptions in a relatively low dimensional Euclidean space, which is of sufficient quality that the positions of the descriptions in the space reflect, at least partially, their meanings. This is despite the defects and idiosyncrasies of the descriptions, which mean that they are of lower quality than, for example, newspaper articles or clinical trial descriptions.

As far as we are aware, this precise type of modelling has not been carried out before, probably because in many applications there is no need either for the use of a Euclidean model space, or if there is, there is no need for it to be low dimensional. However, the fact that this modelling is possible may present an avenue for further research; the modelling described above is quite straightforward, and could be made much more sophisticated. The ability to create a 'cloud' of data, where position reflects meaning might be useful in various applications, most obviously in KDD projects that use data similar to the RLS free text descriptions.

8.2 Post-processing

It has been suggested to us that this is the most controversial part of our work, and we understand this comment. Much of the KDD literature emphasises the importance of being able to repeat results. Clearly, this is not the case with our method; different experts would be likely to come to different subjective judgments of the worth of a group of results, and there is no way of fully codifying the mental processes that they go through to come to these conclusions. However, we feel that in a practical KDD process, it is very likely that some sort of unrepeatable subjective judgment will need to be used; some sort of financial profit/loss metric might be used in certain applications, but isolating the effect of a KDD process might be difficult. In other applications, the final assessment might be entirely subjective. There is perhaps a distinction between assessing the results of an algorithm, where it is desirable that given the data, the output should be deterministic, and assessing a one-off KDD process, where the work will not be repeated, and the value of the results is determined by more intangible properties.

A further question regarding the post-processing is whether it is sufficiently rigorous. We would certainly concede that the quantitative measure that we used is rather ad hoc, and it would be possible to create a measure that used more mathematic rigour. However, we would question whether this measure would necessarily produce better results, given that the ultimate aim of the analysis is to produce groups of incidents that are subjectively meaningful. It does not seem possible to produce a rigorous measure of subjective similarity between incident descriptions. We focused on the quality of the overall process, and having found that our method produced groups of incidents that were coherent, and that the NPSA was enthusiastic about, we decided to focus on other parts of the process. Of course, had this method failed to work as well as it did, we would have put more effort into improving it.

8.3 Using simple methods

All of the methods that we have used are very well established or straightforward; it is combining them that creates a potentially interesting KDD process.

This KDD process has suggested to us that in many cases simple methods may produce superior results. For example, PCA and LDA both produce only linear combinations of the variables in the vector space; however, the models produced using these methods were vastly superior to those produced using neural networks. This appeared to be caused by overfitting, which is not surprising given the volume and dimensionality of the data.

Simplicity may also be a necessity; this project required some knowledge in the fields of patient safety, Natural Language Processing, multivariate statistics and KDD. Inevitably, the analysts carrying out the KDD process will only have a certain amount of experience and expertise, and using more straightforward, well-known methods may well be the only realistic option, or if not, may avoid mistakes being made or time wasted. It is also likely that there will be constraints on the time available for a practical KDD process.

8.4 Communication and organisation

The NPSA was extremely supportive throughout the process, and without this support it would have been much less successful.

We suggest that the aims and objectives of a KDD process should be stated clearly at the start of the process, to try to ensure that the output will be useful, and that the analysis can be carried realistically. Contact between different interested parties is important, and should be maintained throughout the process. We found it useful to present the NPSA with preliminary results for various methods; the feedback allowed us to focus on particular components of the process.

Once the final results of the analysis are available, there must be some mechanism or process by which the results can be assessed efficiently. We suggest that in a process of this kind, the expert assessment is one of the key elements, and the project must be adapted to this stage. The assessment will be dependent on the expertise and commitment of the experts, so every effort must be made to engage them, and maintain their confidence in the KDD process.

Feedback on the process as a whole is likely to be informative, given that the lessons learned are likely to be applicable to other KDD processes; even where the particular methods used are only applicable to a single process, there may be principles that are transferrable.

This communication is likely to take place against a background where the client organisation will change staff, and possibly carry out major reorganisations; this is particularly pertinent for a project lasting several years. The analysts will therefore need to be flexible, and will have to accept that clients' priorities may change.

8.5 KDD pitfalls

It is important to be aware that a large dataset will not necessarily contain useful information. This is an inevitable risk for a practical KDD process: the client may not have a particularly good understanding of their data, and the analysts will not know what the limitations of their analysis are until the work has been carried out. However, good communication at the start of the analysis may avoid unsuccessful projects being carried out, and the risk of a negative result should always be borne in mind. It is also possible that although the primary objectives of the project might not be achieved, unexpected outputs of the algorithm may still be useful.

The breadth of the work in a KDD process means that the research in each part of the process may be relatively shallow. It also means that some of the work may be ad hoc. This limits the extent to which methods can be transferred to other processes, although this should be avoided where possible. It also means that more cutting-edge algorithms are not being tested using these data.

The final part of the KDD process, putting the data into use, proved to be a stumbling block for us, and this may be the case for many KDD processes. Whilst the results produced may look impressive, they need to be of use to the client organisation. There is a danger that if KDD processes consistently produce results that are academically interesting, but not practically useful, the people who own the databases that we are interested in analysing will lose faith in their data, and our ability to extract information from them.

9 Further work

There is scope for further work at each stage of the knowledge discovery process. We think it would be useful for general principles or advice for the initial discussion stages of a KDD process to be developed. Post-processing is another area where it might be useful to draft basic principles that should be followed, which could then be adapted to specific processes.

A potential weakness of the modelling process presented in this paper is that it relies on the existence of categorical variables for supervised dimensionality reduction. This limits the applicability of the method to datasets where these variables are available. We have developed an alternative technique for dimensionality reduction that does not use classification methods, and therefore does not rely on the existence of categorical variables. This method replaces the information contained in the categorical variables with information elicited from an expert about the meanings of words. A small sample of incident descriptions is selected, and the vocabulary used is identified. Information is elicited that allows a low dimensional representation of the relationships in meaning between the words in the vocabulary to be created. The high dimensional vector space model is also calculated for the small sample. An optimum scaling matrix between the high and low dimensional spaces is then calculated. The vector space model for a much larger sample of incidents can then be calculated; multiplying it by the scaling matrix will produce a low dimensional representation of the larger sample. The anomaly detection and assessment processes can then proceed as previously. This method has been found to produce coherent, meaningful groups of incidents that are similar in quality to the group shown in Table 3.

10 Conclusions

We have shown that it is possible to develop a knowledge discovery process using a practical dataset that comprises a mixture of free text, and numerical and categorical variables describing patient safety incidents, and produce results that are found to be coherent by clinical experts.

A model of the free text incident descriptions in a low dimensional Euclidean space has been created, which reflects the meaning of the incident descriptions sufficiently well that groups of incidents that are coherent in meaning can be identified using an anomaly detection algorithm. We have found that both term weighting and reduction of the dimensionality of the vector space model are useful in creating this model of the text, which suggests that methods that stretch or rotate the models should be used in this type of analysis.

The importance of developing methods for assessing the value of the results produced by a practical knowledge discovery process has also been shown, as have the difficulties involved in assessing subjectively the value of the results, particularly when resources are limited.

The individual methods used in the process are all quite straightforward; however, when they are used together, knowledge discovery can be carried out successfully, suggesting that it is necessary to carefully consider the process as a whole, rather than necessarily choosing the most sophisticated method for a particular component.

Acknowledgements James Bentham would like to thank the EPSRC and NPSA for funding his CASE award. The work of David Hand was partially supported by a Royal Society Wolfson Research Merit Award. We would also like to thank two anonymous reviewers and an editor for their helpful comments.

References

Aggarwal CC, Hinneburg A, Keim DA (2000) On the surprising behavior of distance metrics in high dimensional spaces. Lect Notes Comput Sci 1973:420–434

Blum A, Mitchell T (1998) Combining labeled and unlabeled data with co-training. In: COLT: proceedings of the workshop on Computational learning theory, pp 92–100

Department of Health Expert Group (2000) An organisation with a memory. Department of Health, London

Ester M, Kriegel H-P, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. Knowledge Discovery and Data Mining. AAAI Press, Menlo Park, pp 226–231

Fayyad UM, Piatetsky-Shapiro G, Smyth P (1996) The KDD process for extracting useful knowledge from volumes of data. Commun ACM 39(11):27–34

Foster J, Wagner J, van Genabith J (2008) Adapting a WSJ-trained parser to grammatically noisy text. In: Proceedings of the 46th annual meeting of the association for computational linguistics on Human lanauge technologies, pp 221–224, 16–17 June 2008

Hinneburg A, Aggarwal CC, Keim DA (2000) What is the nearest neighbor in high dimensional spaces. In: Proceedings of the 26th VLDB conference, Egypt, pp 506–515

Honnibal M, Nothman J, Curran JR (2009) Evaluating a statistical CCG parser on wikipedia. In: Proceedings of the 2009 workshop on the People's web meets NLP: collaboratively constructed semantic resources, pp 38–41, August

Hyvärinen A, Karhunen J, Oja E (2001) Independent component analysis. Wiley, New York

Lease M, Charniak E (2005) Parsing biomedical literature. In: Second international joint conference on Natural language processing, Jeju Island, pp 58–69

MacQueen JB (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of 5th Berkeley symposium on Mathematical statistics and probability. University of California Press, Berkeley, pp 281–297

Mannila H (1996) Data mining: machine learning, statistics, and databases. In: Proceedings of the eighth international conference on Scientific and statistical database management, 18–20 June 1996, pp 2–9

Manning C, Raghavan P, Schuetze H (2008) Introduction to information retrieval. Cambridge University Press, Cambridge

Rasmussen C (2000) The infinite Gaussian mixture model. Adv Neural Inf Process Syst 12:554–560

Ripley BD (1996) Pattern recognition and neural networks. Cambridge University Press, Cambridge

Saad FH, de la Iglesia B (2006) A comparison of two document clustering approaches for clustering medical documents. In: Proceedings of the 2006 international conference on Data Mining, Las Vegas, USA

Salakhutdinov R, Hinton G (2009) Semantic hashing. Int J Approx Reason 50(7):969–978

Salton G, McGill MJ (1983) Introduction to modern information retrieval. McGraw-Hill, New York

Salton G, Wong A, Yang CS (1975) A vector space model for automatic indexing. Commun ACM 18(11):613–620

Zhang Z, Hand DJ (2005) Detecting groups of anomalously similar objects in large data sets. In: Proceedings of LNCS. Springer, Heidelberg, pp 509–519

Zhang X, Jing L, Hu X, Ng MK, Jiangxi JX, Zhou X (2008) Medical document clustering using ontology-based term similarity measures. IJDWM 4(1):62–73