Submission to Health Informatics Journal

Automated classification of primary care patient safety incident report content and severity using supervised Machine Learning (ML) approaches

Introduction

Learning from patient safety incident reports is a vital part of improving healthcare. However, the volume of reports and their largely free-text nature poses a major analytic challenge.

Objectives

Test the capability of autonomous classifying of free text within patient safety incident reports to determine incident type and the severity of harm outcome.

Materials and Methods

Primary care patient safety incident reports (n=31333) previously expert-categorised by clinicians (training data) were processed using J48, SVM and Naïve Bayes.

Results

The SVM classifier was the highest scoring classifier for incident type (AUROC, 0.891) and severity of harm (AUROC, 0.708). Incident reports containing deaths were most easily classified, correctly identifying 72.82% of reports.

Conclusions

Supervised ML can be used to classify patient safety incident report categories. The severity classifier, whilst not accurate enough to replace manual processing, could provide a valuable screening tool for this critical aspect of patient safety.

Background and Significance

Harm associated with healthcare is the third leading cause of death in the United States [1]. It affects over 10% of patients in hospital [2,3] and 2-3% of those seen in primary care settings [4]. A patient safety incident is said to occur when a situation that could have resulted, or did result, in avoidable harm to a patient is observed during healthcare delivery [5]. Many of these incidents can involve life and death moments. ~~It should be beyond debate that health systems extract the maximum value from analyses to prevent similar occurrences in the future.~~

Healthcare has a poor record of creating actionable learning for quality improvement from patient safety incident reports [6]. One important reason for this is that the most important information ~~– the elements that throws light on causation –~~ is described in the free-text part of an incident report. Whilst every incident report is read and actioned locally, it is often not until they are aggregated that patterns become apparent. In order to aggregate this data though, it must be categorised in the same manner and to the same standard. ~~Unless read and pre-categorised, this information cannot be aggregated to establish frequency and nature of factors that may have contributed to the harm or potential harm to patients.~~ A traditional approach of establishing a classification framework, creating categories and rules for applying them, and then training coding clerks is invariably defeated by the logistics. For example, in England and Wales over 100,000 patient safety incident reports are submitted by frontline clinical staff every month [7]. On a national level, only a

small proportion of ~~the approximately one million~~ patient safety incidents a year is ever analysed for causation [8,9]. This is a remarkable and troubling failure to use data that have already been collected in order to protect patients from harm and inform health system improvements. Rather than focusing decisions on which small minority of incidents to prioritise for analysis [10], a potential solution ~~to the large-scale data loss~~ is Natural Language Processing (NLP) used in conjunction with machine learning (ML). Together, they can convert unstructured free text into structured information autonomously [11-15]. Automatically and accurately assigning incident categories to incident reports would remove a major manual component of our current patient safety strategy on a national level. ~~These computing methods were a key priority for future research into patient safety incident reporting systems in a recent government-funded evaluation [16].~~

A pre-determinant of success of a supervised NLP implementation is the availability of *large* quantities of suitable *training data* from which the machine can learn [11], and which have been categorised by a domain expert [13]. The recent *PISA study* [16] provided a unique corpus of primary care patient safety incident reports that had been read, categorised and coded by trained clinicians with expertise in patient safety and human factors.

Commented [HE1]: From methods as requested by reviewers

Formatted: Normal

Aim

This study aimed to test the capability of NLP/ML to classify unstructured free text within patient safety incident reports in two main themes: the incident

category and harm severity. Each incident had been previously classified manually by an expert clinical and human factors team applying a classification framework that had been developed and validated by the research group [16]. For each of these, the study sought to examine whether this could be achieved using just the unstructured free text description of an incident report alone, or whether the addition of structured categorical data (routinely collected as part of incident reports, such as specialty) improved the success of the autonomous classification.

2. Materials and Methods

(a) Classifiers

This study tested supervised machine learning classifiers, which use pre-existing categorised data to derive learning [17]. Machine learning classifiers and techniques which are able to classify text in documents, including within patient safety incident reports, were identified through literature review. For each research question, three different machine learning classifiers were trained and subsequently evaluated – Naïve Bayes, J48 and Support Vector Machine (SVM) with a polykernel. J48, Naïve Bayes and SVM were chosen since they have been successful in classifying medical incident reports in previous studies [18,19], and represent two distinct approaches to supervised machine learning, namely generative and discriminative models [20]:

- Naïve Bayes, a traditional generative classifier, has repeatedly demonstrated success in document classification tasks [15].
- J48's decision tree structure provides an output that can be intuitively checked by domain experts with limited ML/NLP experience, allowing validation of the core logic of the tree [21].

- SVMs are discriminative classifiers which can cope well with training data consisting of large numbers of irrelevant features, as is the case with our text data. For this reason they have consistently outperformed other classifiers in a number of text categorisation tasks. They are also less prone to class imbalance problems [22].

(bc) Data sources

Patient safety incident reports are principally a free text description [23] with additional categorical values such as location and time to add context. As part of the PISA study the incidents have been categorised against a framework which was iteratively developed, validated and is described in detail elsewhere (the PISA framework) [7]. ~~A pre-determinant of success of a supervised NLP implementation is the availability of *large* quantities of suitable *training data* from which the machine can learn [11], and which have been categorised by a domain expert [13]. Patient safety incident reports have been aggregated nationally in England and Wales since 2003 [7]. Cardiff University holds a complete anonymised copy of this for research purposes. Approximately 50,000 incident reports have been read, categorised and coded by trained clinicians with expertise in patient safety and human factors at Cardiff University. The incidents have been categorised against a framework which was iteratively developed, validated and is described in detail elsewhere (the PISA framework) [7].~~ Categories were applied using the *Recursive Model of Incident Report Analysis* which ensures a chronological listing of incidents culminating in the event that directly harmed the patient [7]. This leads to several *levels* of incident type – "primary" denoting the incident directly impacting the patient, then subsequent levels show the chain of factors that may have contributed to the incident. The PISA study [16] also ~~examined the originally submitted incident reports and~~ reclassified the severity rating. This was used in the present study.

Subset

Incidents that had been categorised as part of the PISA study and related studies were extracted from the database at Cardiff University. Those that had not been categorised by the main PISA incident and severity framework were removed, leaving 31333 incident reports. There were 16 categorical ~~columns~~ variables and four free-text ~~columns~~ variables of data extracted for all incident reports (see appendix 2). One free-text category was rarely completed and often with similar material, and therefore was treated as categorical. The data were then split into two subsets – one including just the free text "description of what happened" field, and another that included all the columns of data available to allow evaluation of whether the additional columns of categorical data assisted the classifier or not. The data were then converted into the Attribute Relation File Format (ARFF) ready for importing into the machine learning software, as per previous studies in the area [19,24].

Dataset processing – characteristics

Figure 1 shows the class imbalance inherent at a high level, with 12649/31333 (40.4%) incidents in the "*0 - Incorrect use of system" category*~~"~~, compared to only 501 (1.6%) in the "*10 – Other"* category. Therefore, the "*0 - Incorrect use of system"* category~~"~~ and *"6 – Medications"* categories were expanded to their second level categories to reduce the class imbalance. Figure 2 shows the incident categories after the expansion.

Formatted: Font: Not Italic

Figure 3 shows the incident severity categories. There were 19323 (61.7%) incidents that did not contain a severity category since they involved

categories that were excluded from severity assessment during the PISA study (e.g. “No harm from primary care” or “defensive reporting”).

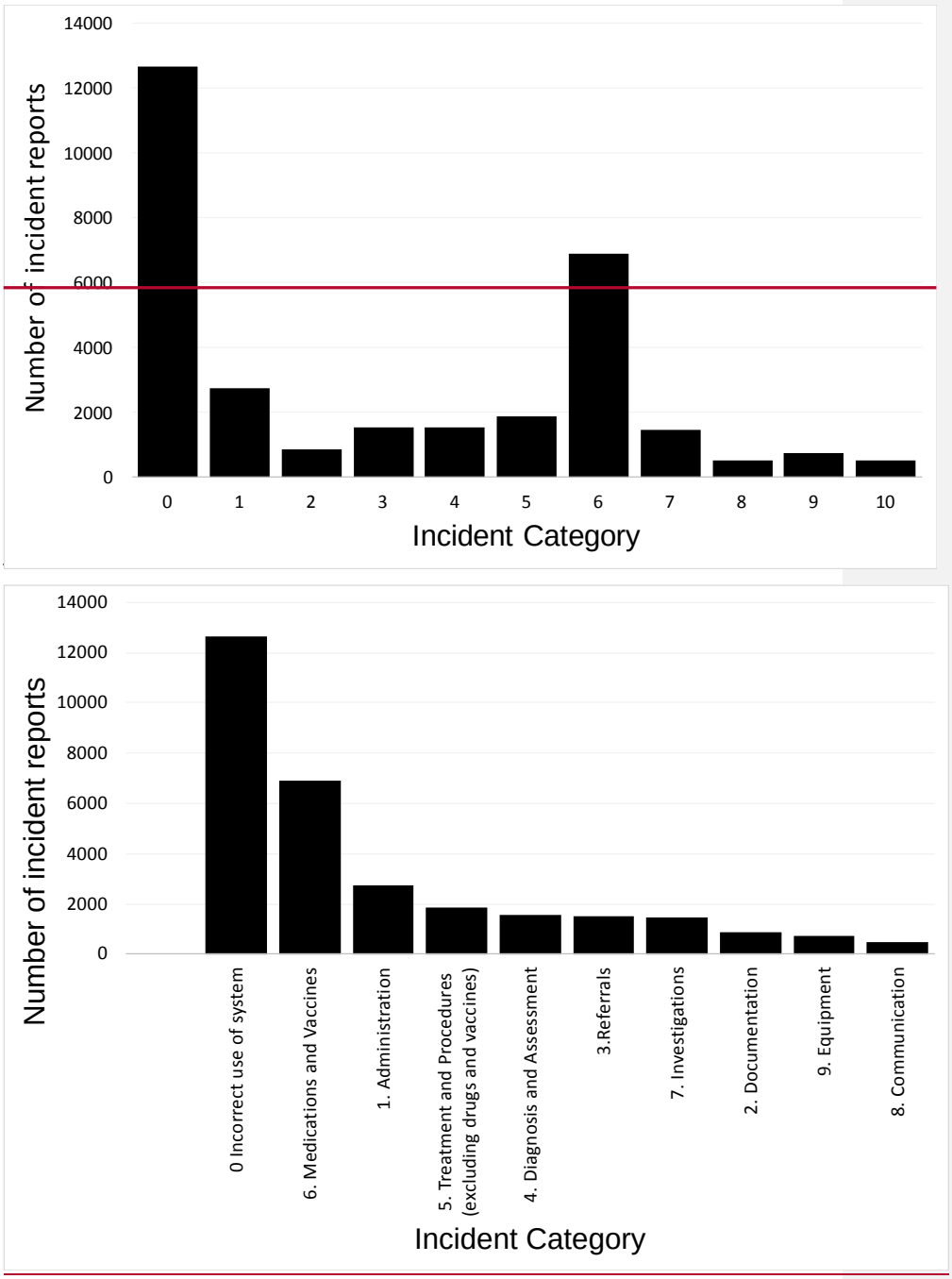
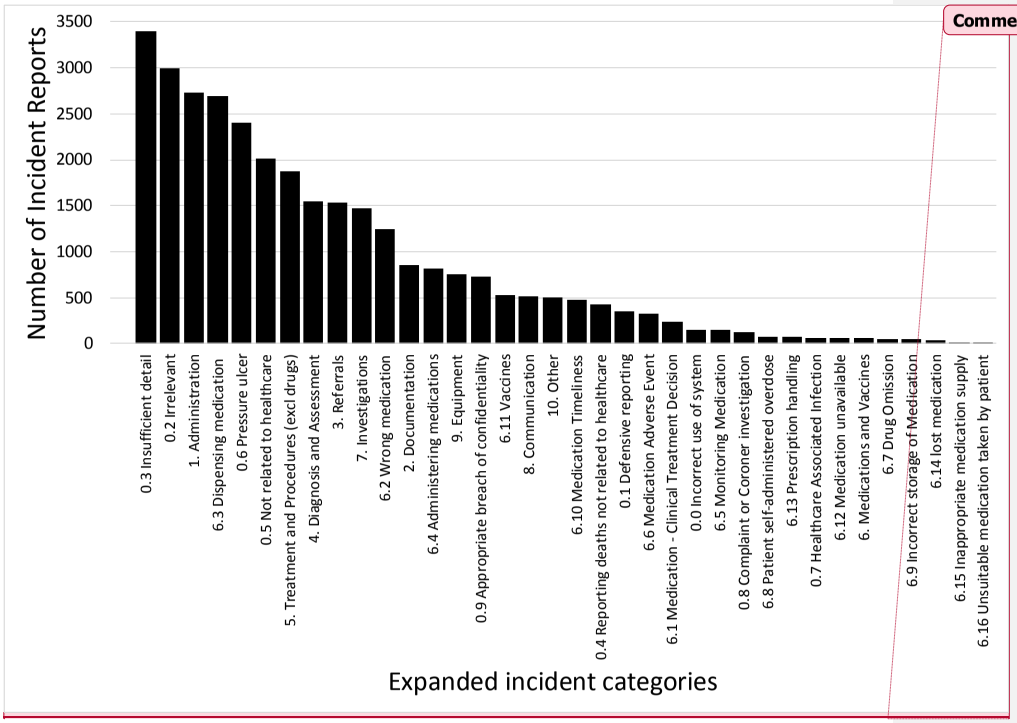


Figure 1 - Number of incident reports by highest level incident categories (0-10)

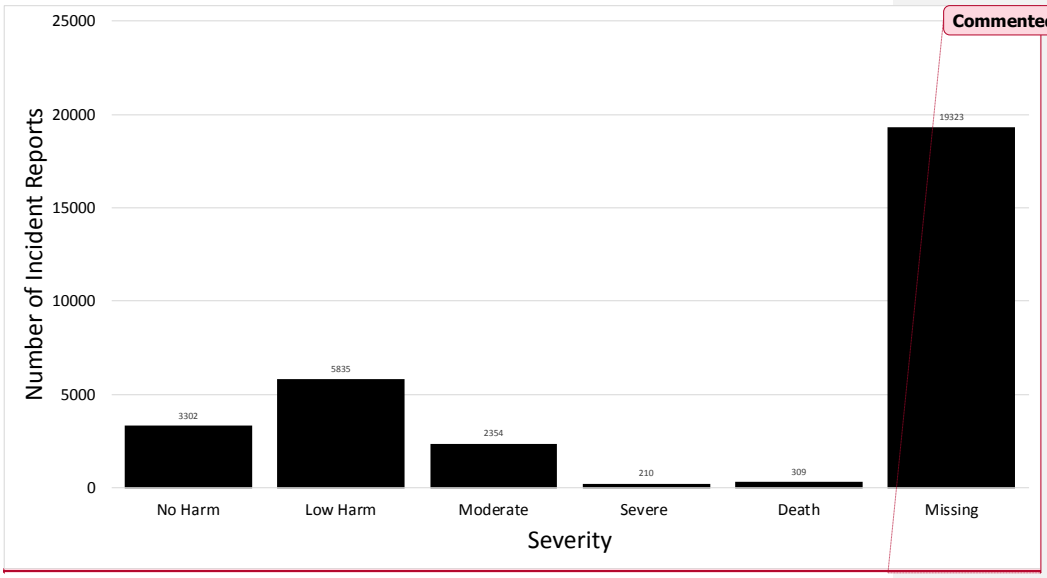


Commented [HE2]: Changed to Pareto chart

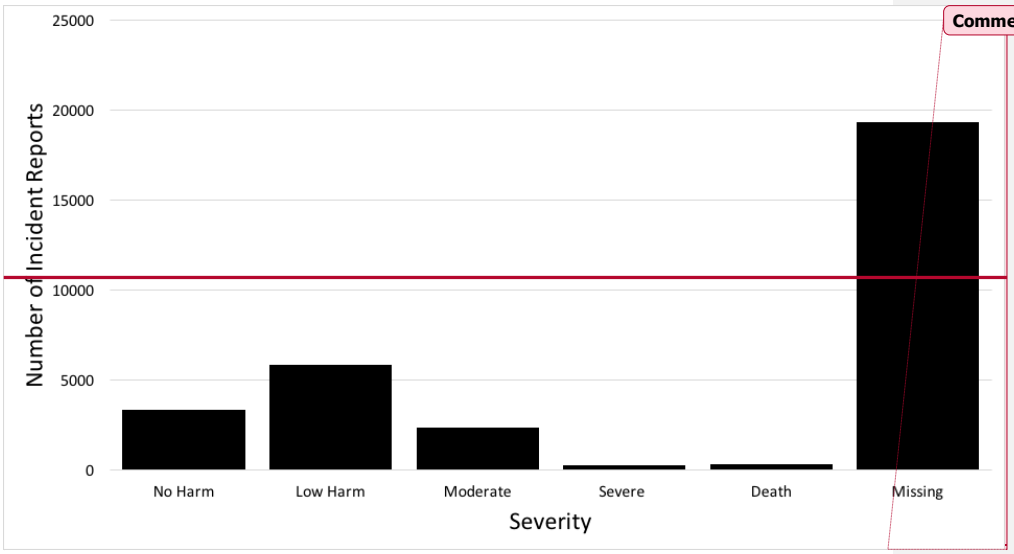
4000

Commented [HE3]: Label rather than numbers and turn into pareto chart

Figure 2 ~~– Number.~~ Number of incident reports by expanded incident categories (0.1 – 10)



Commented [HE4]: Data labels added



Commented [HE5]: Add numbers

Figure 3: Number of incident reports by severity

(cd) Software

Data were accessed and extracted through Microsoft SQL Server 2014, hosted on a secure Microsoft Windows Server 2012R2 instance at Cardiff University. Data were subsequently imported into the *Waikato Environment for Knowledge Analysis (Weka)* 3.8.0, an NLP and ML environment [25]. Weka is regularly used in healthcare document classification and has been used in previous studies into incident report classification [18,19].

(de) Pre-processing incident reports

- All free-text variables were firstly processed using the Weka's *StringToText* filter in order to create a uniform representation for the reports. The following procedures were applied: **NGram Tokenisation** to produce trigrams, bigrams and unigrams [13]. Unigrams represent individual terms (e.g. "patient", "wound", etc). Bigrams and trigrams are sequences of two or three terms (e.g. blood form, blood group, blood result, blood request, pressure ulcer), which were utilised to add an element of semantic processing as negation could also be added (e.g. not allergic), which is important for producing correct classifier rules.
- **Lower case normalisation** was used to ensure that all forms of the same word were classified together (e.g. Patient, patient, pAtient etc.) [26]
- **"Stopword" filtering** was used to exclude common words (such as he, she, it, why, we etc) which hold no classification value [13]. This technique is commonly used in Information Retrieval and NLP

document classification implementations [26]. The “Rainbow” stopwords list built into Weka was used [27]

- **Term frequency filtering**: previous studies have excluded words that appear infrequently in the corpus [17] and due to the large size of the corpus it was decided a minimum term frequency of 10 should be used.
- **Number of words in training set -** 3000 words were kept as a balance between accuracy and resource (CPU/Memory) use. Once the features to be represented were defined through the above procedures, uniform vectorial representations of each report were created where each feature was assigned a TFxIDF (Term Frequency, Inverse Document Frequency) score for that report. TFxIDF values are a function of the frequency of the term in the report, weighted according to the frequency of occurrence of the term in the data set. Intuitively, these scores encode ~~the intuition~~ that the more often a term appears in a report, the more representative of that report it is, while the more reports it occurs in the less discriminative it is. TFxIDF scores can highlight relevant words when categorising large numbers of text documents [26,28,29]. In order for the TF to be accurate, all documents were normalised so longer incident reports did not skew the results [30].

(e) Data Security

All data were stored and accessed on a designated patient safety research computing cluster at Cardiff University, which has been designed with full NHS Information Governance Toolkit assurance for secondary use of data (IG Toolkit ID: 8WG65-PISA-CAG-0182). All data were stored and accessed in accordance with a data sharing agreement between NHS England and Cardiff University.

All data was anonymised by NHS England, compliant with the highest standards of information governance regulations, before being received by Cardiff University. ~~There is no way for researchers to re-identify patients or healthcare organisations.~~

(g) Training and testing the individual classifiers

Each classifier (e.g. SVM, NB) was trained and evaluated using a stratified ten-fold cross validation technique built into Weka, ensuring the maximum amount of training material was available for the training whilst also ensuring rigour and reproducibility [15,19].

(h) Statistics and analyses

Some types of incident reports are naturally reported more frequently (such as those related to medications and vaccines [16]) leading to a “class imbalance”. ~~Due to this, success measures were required that were not susceptible to class imbalance. Therefore,~~ The Area Under the Receiver Operating Curve (AUROC) was chosen as our primary outcome measure since it provides a

single global measure of performance even in imbalanced data [31]. Previous studies in ML/NLP have shown an AUROC of approximately >0.8 as being satisfactory, and the closer to 1.0 the better [32,33]. However, to allow comparability with previous NLP and ML studies in this field, percentage correct and incorrect, precision, recall and F-measure are also reported. Weighted average values, as natively produced by Weka are reported.

(i) Ethical research considerations

The training data used for this current study were generated as part of the NIHR HS&DR study – "Characterising the nature of primary care patient safety incident reports in England and Wales: mixed methods study" - the *PISA study,* which analysed patient safety incident reports submitted to the National Reporting and Learning System from primary care in England and Wales between 2005 and 2013 [7]. The PISA study did not require Health Research Authority's REC approval and the Aneurin Bevan University Health Board research risk review committee waived the need for ethical approval (ABHB R&D Ref number: SA/410/13). Ethical approval for the current study was granted by the Swansea University REC (REF: 040816).

Results

(A) Incident type classification – highest level incident categories (0-10)

Table 1 shows the results of the ML categorisation for the highest level incident categories. SVM had the highest AUROC, improving from 0.839 to 0.854 with the additional columns of data available (see appendix 2).

Classifier	Correct (%)	Incorrect (%)	Cohen's Kappa	Precision	Recall	F-Measure	AUROC
With all ~~columns~~ variables of data available							
SVM	64.111	35.889	0.523	0.629	0.641	0.633	0.854
J48	58.437	41.563	0.4227	0.542	0.584	0.550	0.736
NB	16.092	83.908	0.106	0.540	0.161	0.168	0.564
With only "Description of Incident" available							
SVM	61.845	38.155	0.490	0.602	0.618	0.607	0.839
J48	56.643	43.357	0.421	0.539	0.566	0.550	0.717
NB	12.22	87.780	0.074	0.512	0.122	0.112	0.544

Table 1 - Results of incident type categorisation for the highest-level incident categories

(B) Incident type classification – expanded incident categories (0.1-10)

Table 2 shows the results of the ML categorisation for the expanded incident categories. SVM consistently had the highest AUROC and was improved by the addition of the additional columns of data from 0.870 to 0.891. Neither J48 classifiers completed, aborting after 15 hours (see discussion).

Classifier	Correct (%)	Incorrect (%)	Cohen's Kappa	Precision	Recall	F-Measure	AUROC
With all ~~columns~~ variables of data available							
SVM	52.558	47.442	0.493	0.515	0.526	0.516	0.891
J48	Did not complete						
NB	4.270	95.730	0.037	0.318	0.043	0.061	0.520
With only "Description of Incident" available							
SVM	46.855	53.145	0.4313	0.462	0.469	0.462	0.870
J48	Did not complete						
NB	3.20	96.799	0.0277	0.302	0.032	0.045	0.515

Table 2 - Results of incident type categorisation for the expanded incident categories

Table 3 shows the AUROC for each individual incident category, when using the SVM classifier and all ~~columns~~ variables of data. ~~It has been coloured as a heat map.~~ Classes that achieved AUROC >0.98 included 0.4, 0.6, 0.7, 0.9 and 6.11. Seventeen of the 18 medication categories achieved an AUROC of >0.8.

Table 3 also shows that the number of incident reports in a category is not necessarily proportional to AUROC. For example, category 6.3 has 2686 incidents, AUROC 0.977, but category 6.14 has only 40 incidents but an AUROC of 0.973. In addition, some categories had high numbers of incident reports but low AUROC such as category 0.3, which had 3392 incident reports but an AUROC of only 0.791.

Class	AUROC	Number of incidents	Precision	Recall	F-Measure
0 – Incorrect use of system	0.851	157	0.291	0.102	0.151
0.1 – Defensive Reporting	0.773	357	0.313	0.084	0.132
0.2 – Irrelevant	0.726	2991	0.302	0.280	0.290
0.3 – Insufficient detail	0.791	3392	0.460	0.479	0.469
0.4 – Reporting deaths	0.983	422	0.616	0.737	0.671
0.5 – Incident not related to healthcare	0.929	2015	0.608	0.594	0.601
0.6 – Pressure ulcer	0.981	2398	0.757	0.786	0.772
0.7 – Healthcare associated infection	0.993	64	0.593	0.547	0.569
0.8 – Complaints/Coroner investigation	0.857	129	0.318	0.109	0.162
0.9 – Appropriate breach of confidentiality	0.990	724	0.782	0.822	0.801
1 – Administration	0.884	2734	0.432	0.533	0.477
2 – Documentation	0.932	855	0.537	0.483	0.509
3 – Referral	0.878	1532	0.356	0.337	0.346
4 – Diagnosis and Assessment	0.895	1553	0.387	0.458	0.420
5 – Treatment and procedures	0.866	1876	0.399	0.418	0.408
6 – Medications and vaccines	0.807	58	0.000	0.000	0.000
6.1 - Clinical Treatment Decision Errors in the treatment decision-making process	0.806	238	0.291	0.067	0.109
6.2 – Wrong Medication prescribed	0.948	1243	0.541	0.648	0.590
6.3 – Dispensing medication orders error	0.977	2686	0.785	0.842	0.812
6.4 – Administering medication errors	0.931	819	0.504	0.591	0.544
6.5 – Monitoring medications	0.944	152	0.478	0.283	0.355
6.6 – Adverse event (inc allergies)	0.945	321	0.533	0.505	0.518
6.7 – Drug omission	0.902	54	0.000	0.000	0.000
6.8 – Patient self-administered overdose	0.911	81	0.333	0.086	0.137
6.9 – Incorrect storage	0.967	44	0.400	0.091	0.148
6.10 – Medication Timeliness	0.916	476	0.432	0.408	0.419
6.11 – Vaccines	0.988	534	0.806	0.801	0.804
6.12 – Medication unavailable	0.887	61	0.600	0.148	0.237
6.13 – Prescription handling	0.969	74	0.444	0.162	0.238
6.14 – Lost medication	0.973	40	0.450	0.225	0.300
6.15 - Inappropriate medication supply	0.911	14	0.000	0.000	0.000
6.16 - Unsuitable medication taken by patient	0.806	3	0.000	0.000	0.000
6.17 – OTC medication	0.500	1	0.000	0.000	0.000
7 – Investigations	0.977	1473	0.776	0.777	0.776
8 – Communication	0.840	510	0.198	0.159	0.176
9 – Equipment	0.899	751	0.461	0.379	0.416
10 - Other	0.870	501	0.342	0.188	0.242

Formatted Table

Table 3 - AUROC for Expanded incident categories, with all columns of data available using the SVM classifier

(C) Severity classification

Table 4 shows the results for severity classification. SVM achieved the highest AUROC at 0.708 with all columns of data, although this was not above our threshold for accuracy ~~for clinical use~~. Figure 4 shows the confusion matrix for the SVM classifier for the expanded incident categories and has been coloured to demonstrate where the classifier has classified correctly and where it has failed. In the death category it correctly identified 72.85% (225/309) of cases involving death compared to only 20.95% in the *severe harm* category.

Classifier	Correct (%)	Incorrect (%)	Cohen's Kappa	Precision	Recall	F-Measure	AUROC
With all ~~columns~~ variables of data available							
SVM	64.371	35.627	0.448	0.643	0.644	0.643	0.708
J48	64.355	35.645	0.420	0.644	0.644	0.629	0.694
NB	20.900	79.001	0.113	0.589	0.209	0.276	0.573
With only "Description of Incident" available							
SVM	61.091	38.909	0.392	0.609	0.611	0.609	0.683
J48	58.943	41.058	0.359	0.585	0.589	0.587	0.647
NB	16.728	83.272	0.088	0.595	0.167	0.226	0.561

Table 4 -Results of severity categorisation for the expanded incident categories

		Classified by NLP/ML Classifier					
	Classified as →	1 (No Harm)	2 (low Harm)	3 (Moderate)	4 (Severe)	5 (Death)	% Correct
Classified by humans	1 (No Harm)	2028	1183	87	1	3	61.42
	2 (Low Harm)	1285	3957	563	15	15	67.81
	3 (Moderate)	167	628	1477	57	25	62.74
	4 (Severe)	14	51	88	44	13	20.95
	5 (Death)	7	31	33	13	225	72.82

Figure 4 - Confusion matrix for severity classification for SVM with all columns of data available *(green = ~~perfect~~ correct classification, yellow = close to perfect orange and red = classification failures)*

Discussion

This study has shown great promise for automatically analysing patient safety incidents, and has achieved this in several incident categories. It has succeeded in accurately ~~identifying~~ classifying the content of incident reports particularly in medication incidents (17/18 categories achieving an AUROC of >0.8) and in pressure ulcers (AUROC 0.981). We have also succeeded in identifying patients who have died, from the content of incident reports, correctly 72.82% of the time which will provide a valuable safety net ~~for national analyses~~.

However, we have also shown that this method does not perform well when classifying the severity of harm of patient safety incident reports. Whilst the so-called “bag of words” approach yields limited success, this may be sufficient to serve as a safety net to ensure that important cases are not missed during review. This study has also highlighted the categories that need both further refining of their definitions, and where additional categorised incident reports are needed to most efficiently improve and refine the classifier. For example, vaccine errors achieved an almost perfect AUROC of 0.988 – thus further human classification would not improve this value considerably. In contrast, further training material for the category “8 – Communication” (with an AUROC of 0.84 and only 510 reports) may improve its accuracy considerably.

We found that the number of incident reports is not proportional to the overall success of the categorisation. This is consistent with Ong et al. (2010). Potentially, once the classifier has ascertained the best words to identify an incident category, further reports do not add to its accuracy.

Certain categories were harder to classify autonomously than others. This is also true of incidents studied in the aviation industry [34]. This may be because certain categories have few specific terms that the algorithm can utilise to confidently discriminate. Conversely, certain categories which have very specific words, such as in pressure ulcers (category 0.6), where words such as “pressure” and “grade” are fairly unique in medicine to this topic lead to highly accurate classifications. This has been highlighted in previous work [32]. Similarly, since healthcare professionals write reports in very high level, technical language it regularly contains abbreviations and acronyms which pose a further problem for the classifier [35,36]. More problematic for certain categories, such as “7 *investigations”*, where healthcare professionals are more likely to call a “Full Blood Count” an “FBC”, or a “Positron Emission Tomography scan” a “PET scan” than in other domains like *communication*. However, this can also be seen as a positive since terms that are specific to certain domains are ideal for a classifier. Nuances and ambiguity of language can lead to confusion for the classifier and this has been highlighted as a problem in other NLP/ML applications too [11]. The addition of spelling mistakes causes further issues for the classifier since it treats different spellings as different words and thus classifies them differently. This is regularly a problem in other NLP/ML studies [35].

However, although the number of words may not influence accuracy, when combined with our hardest task (computationally) – the expanded incident categories with 37 possible categories - it may explain why the J48 classifier failed. Decision trees are computationally expensive, needing large amounts of resource (processing and memory) and do not scale to large numbers of classes [37]. In this study, that led to the J48 classifier running out of memory before completing.

One key category which posed problems for the classifier is contained in the 0.2 category, *0.2.1 – no harm from primary care*. This category is used where there is a patient safety incident but it was not caused by an act or omission by primary care. It is likely that the classifier correctly identifies these incidents as, for example, medication incidents but because it was caused by secondary care it is classified as “no harm from primary care” by the PISA study. It is therefore seen as a misclassification, despite the classifier being technically correct.

Strengths and Limitations

This study had several strengths. Firstly, it was the first study of its kind to use UK primary care ~~data~~ incident reports and moreover, was the largest ML/NLP study of patient safety incident reports conducted that we are aware of. Secondly, it used more incident categories than any other study we are aware of, and it was the first of its kind to use not only the information from the reporter, but in addition, the expert applied PISA classification system [7].

There are several broad limiting factors for the overall performance of the study however, often these were out of our (and any studies) control namely the original content of the incident reports, the PISA coding of the incident reports (and their sampling) and inherent limitations of the classifiers themselves.

As seen in other studies on incident analysis, clear definitions can be more important than the size of the training set from which the classifier has to learn [34]. Table 4 shows this clearly and here this study's methodology may have limited the outcome of its classifier. The PISA classification was iteratively developed and contains over 350 different incident categories. It was decided at the outset that there were insufficient data to train a classifier on all 350 categories, due to its hierarchical structure, and therefore to focus on the highest level categories (0 to 10). Whilst this seems at the outset to be simpler for the classifier, it may conversely lead to more confusion since large quantities of incident reports are now grouped by broad vague concepts such as "Medication incidents", "Incorrect use of system" and "Other". The "Incorrect use of system" category is the broadest, ranging from pressure ulcers, through to defensive reporting. To assess if this had caused further confusion the broadest categories - "Incorrect use of system" and "Medication incidents" – were broken down to their next level in the hierarchy which increased the AUROC despite increasing the number of categories from 11 to 31 and at the same time reducing the number of categories available in each category from which to train.

The classifiers used in this study were trained only on the final incident that has directly led to patient harm. However, a single report may contain several interconnected incidents that led to the final outcome. The classifier may correctly identify any number of incidents contained within the report, but if it does not choose the *final*/*primary* incident it will technically get the category wrong. This will require further research. The ultimate category applied to an incident is often subject to much debate and scrutiny, often requiring a third party to cast the final vote [7]. This is seen in numerous studies which used expert-categorised data to train their classifiers, where disagreement between experts was seen in up to 20% of cases [38]. Therefore, we should not expect every incident to have been categorised in exactly the same way due to there being several (albeit highly trained) coders in the original study [16].

The "bag of words" strategy, is a simple and effective approach; however, structure from the text is lost and thus the semantic meaning [12,18]. Negation is lost (e.g. "no allergies") which poses a major problem since it treats the word 'allergies' the same irrespective of the preceding terms and this has been shown to be a problem in other studies [39]. To compensate for this, bigrams and trigrams were utilised in this study which would have attempted to identify the above example. Another solution is to use a semantic processor which can analyse sentences in their entirety [13]. However, even with this approach, sometimes the sentences either side can affect the meaning of the sentence in question, so called 'cross-sentence correlation', which can have a similar effect as negation [40]. Recent works with paragraph vectors have shown improvements on the bag of words model by up to 30% [41].

Comparison with previous work

There has been little research conducted on the use of ML and NLP in automating incident report analysis in healthcare [18]. There has been considerably more research and success with it in incident reports in aviation [34], and notable successes reported for text classification from verbal autopsies[42] which have several similarities with incident reports. Of those studies of safety reports in~~in~~ healthcare, Wong et al. (2013) undertook a study of 227 Canadian medication incident reports, and used a custom classifier based on logistic regression to achieve good accuracy in autonomously categorising incident type [32]. Ong et al. (2010) performed a larger study of 972 incident reports in Australia by focusing on two types of patient safety incident: “inadequate clinical handover” and “incorrect patient identification”. They used Naïve Bayes (NB) and Support Vector Machine (SVM) classifier with excellent results (accuracy up to 97.98% with SVM on patient identification incidents) but noted that the topics chosen had very specific words that the classifier could easily detect which probably lead to their good results [18]. Gupta and Patrick (2013) undertook a larger study of 5448 Australian incident reports, including 13 categories of incident type and utilised NB, SVM as well as the J48 decision tree classifier. They have reported achieving good results in an online presentation, however their detailed methodology has not been published making further comparison difficult. [19]. The largest work in the field (up until now) appears to have been undertaken in Japan, where 15,000 patient safety incident reports were

Commented [MOU6]: Danso et al., 2014.

clustered using cluster analysis to ascertain their incident type but they did not provide statistical or numerical results [23,43]. A recent paper by Wang and colleagues looked at using ML and NLP to categorise Australian incident reports [44]. Their study used fewer incident categories and used a significantly smaller dataset than ours and they too struggled to classify severity level. Wang et al also demonstrated the difference that using balanced datasets makes to the accuracy of the task, although since real world incident report data are inherently imbalanced we did not choose to balance our dataset.

Recommendations for future work

This project is the largest attempt at classifying patient safety incident reports in primary care to date, but further research will be required to achieve the same results on secondary care data. Within the scope of the current dataset future research could focus on examining incident reports in their entirety utilising semantic classifiers[12], and whether sequences of incidents can be extracted, something that has been researched in airline incident report analysis[29]. Although the categorical data routinely collected with each report is often non-specific, as it improved our study's performance it would be prudent to further research how these data can be used to enhance incident report categorisation. Further work around J48, either using reduced categories or superior infrastructure is required, since its "human readable" output allows checking for plausibility by patient safety experts. Improving definitions and increased training examples of select categories will likely further improve the performance.

~~Further research should focus on looking at the incident report as a whole and utilise semantic classifiers[12]. In addition, further research should explore whether it is possible to classify the incident sequence within each report, which has also been identified in airline incident report analysis [29]. Finally, and~~

~~fundamentally, improving the definitions of certain incident categories, coupled with additional examples of incident reports in some categories, is probably the next key step towards fully autonomous classification of incident reports in primary care in NHS England and Wales.~~

~~This project is the largest attempt at classifying patient safety incident reports in primary care to date. However, its focus on primary care may limit its generalisability on other patient safety datasets such as secondary care and further research should focus on ML/NLP in secondary care in the UK.~~

Conclusion

Converting unstructured data to structured data using NLP/ML is challenging across all subject domains [13,40,45]. However, the highly nuanced and technical nature of medical text adds a further dimension of complexity [46]. Whilst this study shows that NLP/ML is not perfect and cannot yet replace manual review entirely [47], it suggests that it can act as a safety net, identifying cases that lead to severe harm and death, that have been incorrectly classified. The ability to determine certain categories accurately can also assist reviewers in those areas to focus on cases that need manual review – saving money and time [48]. It also opens up the possibility of clustering reports that are “near misses” or “no harm”, which are currently too time consuming to work on in healthcare; which is a key strategy used by the airline industry in their successful safety model [49].

References

1. Makary MA, Daniel M. Medical error—the third leading cause of death in the US. *BMJ*; i2139–5.

2. Vincent C, Neale G, Woloshynowych M. Adverse events in British hospitals: preliminary retrospective record review. *BMJ* 2001.

3. Rafter N, Hickey A, Condell S, et al. Adverse events in healthcare: learning from mistakes. *QJM*; 108: 273–277.

4. Panesar SS, deSilva D, Carson-Stevens A, et al. How safe is primary care? A systematic review. *BMJ Quality & Safety*; 25: 544–553.

5. World Health Organization. Conceptual Framework for the International Classification for Patient Safety. 1–153.

6. Stavropoulou C, Doherty C, Tosey P. How Effective Are Incident-Reporting Systems for Improving Patient Safety? A Systematic Literature Review. *Milbank Q*; 93: 826–866.

7. Carson-Stevens A, Hibbert P, Avery A, et al. A cross-sectional mixed methods study protocol to generate learning from patient safety incidents reported from general practice. *BMJ Open*; 5: e009079–8.

8. National Patient Safety Agency (NPSA). Seven Steps to patient safety: full reference guide. *nrls.npsa.nhs.uk/resources*http://www.nrls.npsa.nhs.uk/resources/?entryid45=59787&q=0%c2%acseven+steps+to+patient+safety%c2%ac (2004, accessed 15 June 2016).

9. National Patient Safety Agency (NPSA). Being open: communicating patient safety incidents with patients, their families and care. *nrls.npsa.nhs.ukresourcescollectionsbeing-open*http://www.nrls.npsa.nhs.uk/resources/collections/being-open/?entryid45=83726 (2009, accessed 1 August 2016).

10. Hibbert PD, Healey F, Lamont T, et al. Patient safety's missing link: using clinical expertise to recognize, respond to and reduce risks at a population level. *Int J Qual Health Care*; 28: 114–121.

11. Erhardt RA-A, Schneider R, Blaschke C. Status of text-mining techniques applied to biomedical text. *Drug Discovery Today*; 11: 315–325.

12. Hirschberg J, Manning CD. Advances in natural language processing. *Science*; 349: 261–266.

13. Kimia AA, Savova G, Landschaft A, et al. An Introduction to Natural Language Processing How You Can Get More From Those Electronic Notes You Are Generating. *Pediatr Emerg Care*; 31: 536–541.

14. Melton GB, Hripcsak G. Automated Detection of Adverse Events Using Natural Language Processing of Discharge Summaries. *J Am Med Inform Assoc*; 12: 448–457.

15. Witten IH, Frank E, Hall MA. *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd ed. Morgan Kaufmann Publishers Inc., 2011.

16. Carson-Stevens A, Hibbert P, Williams H, et al. Characterising the nature of primary care patient safety incident reports in the England and Wales National Reporting and Learning System: a mixed-methods agenda-setting study for general practice. *Health Services and Delivery Research*; 4: 1–76.

17. Savova GK, Ogren PV, Duffy PH, et al. Mayo Clinic NLP System for Patient Smoking Status Identification. *J Am Med Inform Assoc*; 15: 25–28.

18. Ong M-S, Magrabi F, Coiera E. Automated categorisation of clinical incident reports using statistical text classification. *Quality and Safety in Health Care*; 19: e55–e55.

19. Gupta J, Patrick J. Automated validation of patient safety clinical incident classification: macro analysis. *Stud Health Technol Inform*; 188: 52–57.

20. Ng AY, Jordan MI. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*; 841–848.

21. Yadav K, Sarioglu E, Smith M, et al. Automated Outcome Classification of Emergency Department Computed Tomography Imaging Reports. *Acad Emerg Med*; 20: 848–854.

22. Japkowicz N, Stephen S. The class imbalance problem: A systematic study. *Intelligent data analysis*; 429–449.

23. Fujita K, Akiyama M, Toyama N, et al. Detecting effective classes of medical incident reports based on linguistic analysis for common reporting system in Japan. *Stud Health Technol Inform*; 192: 137–141.

24. Pollettini JT, Panico SRG, Daneluzzi JC, et al. Using Machine Learning Classifiers to Assist Healthcare-Related Decisions: Classification of Electronic Patient Records. *J Med Syst*; 36: 3861–3874.

25. Frank E, Hall MA, Witten IH. Waikato Environment for Knowledge Analysis (Weka). 2016.

26. Alicante A, Amato F, Cozzolino G. A Study on Textual Features for Medical Records Classification. *Studies in Health Technology and Informatics* 2015. Epub ahead of print 2015. DOI: 10.3233/978-1-61499-474-9-370.

27. Kachites M. Bow: a Toolkit for Statistical Language Modeling, Text Retrieval, Classification and Clustering.

http://www.cs.cmu.edu/~mccallum/bow (1996).

28. Alparslan E, Karahoca A, Bahşi H. Classification of confidential documents by using adaptive neurofuzzy inference systems. *Procedia - Procedia Computer Science*; 3: 1412–1417.

29. Ittoo A, Le Minh Nguyen, van den Bosch A. Text analytics in industry: Challenges, desiderata and trends. *Computers in Industry*; 78: 96–107.

30. Manning CD, Raghavan P, Schutze H. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2009. Epub ahead of print 2009. DOI: 10.1017/CBO9780511809071.

31. Batista GE, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*; 6: 20.

32. Wong Z, Akiyama M. Statistical text classifier to detect specific type of medical incidents. *MedInfo* 2013. Epub ahead of print 2013. DOI: 10.3233/978-1-61499-289-9-1053.

33. Tenório JM, Hummel AD, Cohrs FM, et al. Artificial intelligence techniques applied to the development of a decision-support system for diagnosing celiac disease. *Int J Med Inform*; 80: 793–802.

34. Tanguy L, Tulechki N, Urieli A, et al. Natural language processing for aviation safety reports: From classification to interactive analysis. *Computers in Industry*; 78: 80–95.

35. Penz JFE, Wilcox AB, Hurdle JF. Automated identification of adverse events related to central venous catheters. *Journal of Biomedical Informatics*; 40: 174–182.

36. Savova GK, Fan J, Ye Z, et al. Discovering peripheral arterial disease cases from radiology notes using natural language processing. *AMIA Annu Symp Proc*; 2010: 722–726.

37. Moise I, Pournaras E, Helbing D. Classification and Decision Treeshttps://www.ethz.ch/content/dam/ethz/special-interest/gess/computational-social-science-dam/documents/education/Spring2016/Datascience/Classification%20and%20Decision%20Trees.pdf (2016, accessed 7 December 2016).

38. Hripcsak G, Austin JHM, Alderson PO, et al. Use of Natural Language Processing to Translate Clinical Information from a Database of 889,921 Chest Radiographic Reports1. *Radiology*; 224: 157–163.

39. Hou JK, Chang M, Nguyen T, et al. Automated Identification of Surveillance Colonoscopy in Inflammatory Bowel Disease Using Natural Language Processing. *Dig Dis Sci*; 58: 936–941.

40. Sevenster M, van Ommering R, Qian Y. Automatically Correlating Clinical Findings and Body Locations in Radiology Reports Using MedLEE. *J Digit Imaging*; 25: 240–249.

41. Le QV, Mikolov T. Distributed Representations of Sentences and Documents. *Proceedings of the 31st International Conference on Machine Learning, Beijing, China,* 2014; 32.

42. Danso S, Atwell E, Johnson O. A Comparative Study of Machine Learning Methods for Verbal Autopsy Text Classification. *arXiv.org*; 1402: arXiv:1402.4380.

43. Fujita K, Akiyama M, Park K, et al. Linguistic analysis of large-scale medical incident reports for patient safety. *Stud Health Technol Inform*; 180: 250–254.

44. Wang Y, Coiera E, Runciman W, et al. Using multiclass classification to automate the identification of patient safety incident reports by type and severity. *BMC Med Inform Decis Mak*; 17: 84.

45. Alghoson AM. Medical Document Classification Based on MeSH. IEEE, 2013, pp. 2571–2575.

46. Stanfill MH, Williams M, Fenton SH, et al. A systematic literature review of automated clinical coding and classification systems. *J Am Med Inform Assoc*; 17: 646–651.

47. Warrer P, Hansen EH, Juhl-Jensen L, et al. Using text-mining techniques in electronic patient records to identify ADRs from medicine use. *British Journal of Clinical Pharmacology*; 73: 674–684.

48. Melton GB, Hripcsak G. Automated Detection of Adverse Events Using Natural Language Processing of Discharge Summaries. *J Am Med Inform Assoc*; 12: 448–457.

49. Oster CV Jr, Strong JS, Zorn CK. Analyzing aviation safety: Problems, challenges, opportunities. *Research in Transportation Economics*; 43: 148–164.

Appendix 1 – High level incident categories (from PISA Study)
For further details see Carson-Stevens et al. 2016 [16]

0.~~0~~ Incorrect use of system
0.1 Defensive reporting
0.2 Irrelevant
0.3 Insufficient detail
0.4 Reporting deaths not related to healthcare
0.5 Reporting an incident or patient injury not related to healthcare
0.6 Pressure ulcer
0.7 Healthcare Associated Infection
0.8 Complaint or Coroner investigation
0.9 Appropriate breach of confidentiality
1. Administration
2. Documentation
3.Referrals
4. Diagnosis and Assessment
5. Treatment and Procedures (excluding drugs and vaccines)
6. Medications and Vaccines
6.1 Clinical Treatment Decision - Errors in the treatment decision-making process
6.2 Wrong medication Wrong medication prescribed
6.3 Dispensing medication orders Error in the process of delivering a medication order or inappropriate medication order by a provider working under physician supervision
6.4 Administering medications Error in the process of administering medication to a patient
6.5 Monitoring Medication - Error in the process of monitoring dose-dependent medications, or those with side effects
6.6 Adverse Event - Patient suffered a complication as a result of medication
6.7 Drug Omission - Medication erroneously not given to or not taken by patient
6.8 Patient self-administered overdose - Unintentional drug overdose by patient (self-administered)
6.9 Incorrect storage Medication incorrectly stored
6.10 Medication Timeliness Medication not commenced in a timely fashion
6.11 Vaccines
6.12 Medication unavailable
6.13 Prescription handling. Errors arising from e.g. lost or accidentally shredded prescriptions.
6.14 lost medication
6.15 Inappropriate medication supply - e.g illegal supply of medication
6.16 Unsuitable medication taken by patient Medication erroneously taken by patient e.g. medication taken when stopped by GP or when not recommended
6.17 OTC supply
7. Investigations
8. Communication
9. Equipment
10. Other

Appendix 2 – list of columns of data used

1. Care setting of occurrence (categorical)
2. Location – level 1 (categorical)
3. Location – level 2 (categorical)
4. Location – level 3 (categorical)
5. Incident category – level 1 (categorical)
6. Incident category – level 2 (categorical)
7. Incident category – free text (converted to categorical – see method)
8. Description of what happened (free text)
9. Actions preventing recurrence (free text)
10. Apparent causes (free text)
11. Specialty – level 1 (categorical)
12. Specialty – level 2 (categorical)
13. Degree of harm/severity (categorical)
14. Medical process (categorical)
15. Medical error category (categorical)
16. Approved name of drug 1 (categorical)
17. Proprietary name of drug 1 (categorical)
18. Approved name of drug 2 (categorical)
19. Proprietary name of drug 2 (categorical)
20. Patient age at time of incident (categorical)