Data mining to gain intelligence from routine incident reporting.

Structured Abstract:

Purpose: Incident reporting systems are commonly deployed in healthcare but resulting datasets are largely warehoused. This study explores if intelligence from such datasets could be used to improve quality, efficiency and safety.

Design/Methodology/Approach: Incident reporting data recorded in one NHS acute Trust was mined for insight (n=133,893 April 2005-July 2016 across 201 fields, 26,912,493 items). An a priori dataset was overlaid consisting of staffing, vital signs and national safety indicators such as falls. Analysis was primarily nonlinear statistical approaches using Mathematica V11.

Findings: The organisation developed a deeper understanding of the use of incident reporting systems both in terms of usability and possible reflection of culture. Signals emerged which focused areas of improvement or risk. An example of this is a deeper understanding of the timing and staffing levels associated with falls. Insight into the nature and grading of reporting was also gained.

Practical applications: Healthcare incident reporting data is underused and with a small amount of analysis can provide real insight and application to patient safety.

Value: This study shows insight can be gained by mining incident reporting datasets, particularly when integrated with other routinely collected data.

Keywords: Health Informatics, Patient Safety, Risk Management, Workforce, Staffing, Incident Reporting, Databases, Data Mining, Information and Knowledge Management

Article Classification: Research

Received: 23rd August 2018
Revised:
Accepted:

Introduction

Patient safety is both a national and global priority. In the English National Health Service (NHS), it is estimated that one in ten patients comes to serious harm as a result of their healthcare; half of these incidents of harm are considered preventable (Hogan *et al.*, 2013). This finding is common. Across the world the World health Organization estimates as many as one in ten patients come to harm in high income countries during inpatient care, with almost 50% of these incidents of harm being avoidable (WHO 2019)

The current approach to managing safety in healthcare is to record and measure harm (NHS Quality Observatory, 2013) rather than fully understand the characteristics of its absence (Reason, 2000). Central to the collection of data on harm in the acute health sector is the utilization of electronic databases (Lugg-Widger *et al* 2018).

Electronic incident reporting was introduced into many acute organisations to replace paper-based systems during the late 1990's (Hazan, 2016) and currently such applications are common both across the English healthcare system and internationally. Similar systems have been used in the aviation industry for a number of years to manage risk and have been shown to be beneficial (Hudson, 2003). Despite this, large data sets linking safety to other factors are rarely examined in healthcare beyond survival and morbidity (Howell *et al.*, 2015).

Knowledge discovery through data mining (KDD) is an interdisciplinary area focusing upon methodologies for extracting useful knowledge from large volume data. The widespread use of safety reporting databases has created a need for KDD approaches (IBM Research Series, 2012, Bates *et al* 2018) to investigate the full potential of the data collected.

Most incident reporting data within the acute NHS appears to be warehoused and not mined for insight (Leary & Dix 2018). However, it is likely to contain 'unknown unknowns' - areas which might not already be associated with conventional key performance indicators (KPI) or causes of harm. A recent study found that of 589 quality and safety charts in reports to English NHS Trust boards only 17% (100/589) utilized incident reports (Schmidke *et al.*, 2017).

The challenge of extracting knowledge from data draws upon research in statistics, pattern recognition, machine learning, data visualisation, optimisation, and computing to deliver advanced intelligence (IBM Research Series, 2012, Sivarajah *et al.*, 2017 Fayyad *et al.*, 1996, Brennan and Bracken, 2015). Whilst common in industry it is still fairly uncommon in the NHS.

With a large number of reports, the detection of rare events requires the use of data-mining software which is still immature in healthcare (Rabel *et al.*, 2017). Therefore, an opportunity exists to obtain deeper insight and intelligence from these data which could be used in different ways such as informing quality improvement. Although KDD is uncommon in health there have been attempts to examine the overlap between, for example, incident reporting and complaints (Goldsmith *et al.*, 2015).

Data mining has been used on existing data, such as the electronic health record data in the US, which is gathered from multiple organisations (Almasalha, 2013). The standardisation of data is important as the quality of the data determines the value of the data mining and analysis (Sacristan and Dilla, 2015). Such methods are becoming more common in studies which examine safety (Staggs and Dunton, 2014, Leary *et al.*, 2016, Cook *et al* 2019).

Datix is an incident reporting and risk management platform deployed in approximately 80% of NHS providers across England (Datix, 2019), to which members of staff of any professional group can report both instances of harm and near harm as a patient safety incident. A "patient safety incident" is defined in the NHS as "any unintended or unexpected incident which could have, or did, lead to harm for one or more patients receiving healthcare" (NHS Improvement 2017). The Datix reporting system involves inputting information such as date and time stamps, the intensity of the event, classifications such as ward identifiers and a free text field where more detailed information regarding the event can be noted. Once all the information has been submitted the Datix system produces a report of the incident which others, for example managers, can access.

At University Hospitals Coventry and Warwickshire NHS Trust (UHCW), approximately 130,000 patient safety incidents have been reported into the system since its introduction in 2005. The Trust has adopted two overarching approaches to utilising the information contained in each report. Firstly, individual investigations are conducted to explore the cause of each incident and secondly system-level trend analyses are carried out and are reported at various senior management committees which feed into strategic decision making. Currently, no nonlinear inferential statistics, data mining or pattern recognition techniques are applied. It is hypothesized that by exploring these data using KDD insight will be gained. It is further suggested that analysis of the free text data may lead to insight as to why events occur which may be more useful than simply noting how often they occur. These data utilized are currently in the form of an aggregate output at strategic level. This is common across organisations; a recent study indicated that only six percent of charts reported at board level provide statistics that depict the role of chance in the outcomes (Schmidkte *et al.*, 2017).

The Datix database represents a large, untapped source of knowledge on patient safety. It consists of categorical data derived from a common classification system and free text which in users can describe incidents in their own words. Categorical data such as time, location, type and severity make up the bulk of formal reporting whereas free text is hardly ever utilized apart from individual analysis. Analysis of the entire dataset could provide insight and contribute to alleviating the 'blame culture' (Department of Health, 2015) which is still experienced across the NHS and derive organisation-level relationships that are currently undetectable. This might make visible safety issues that are not currently recognised and may even promote a just culture through transparency (NHS Improvement 2018).

The aim of developing this technique in the most commonly used incident reporting system is to gain what is common in other safety critical industries, using data for insight that allows organisations to improve safety. If successful this technique could be used across other organisations that use similar data capture systems.

Research question and method

Does mining routinely collected incident data give useful intelligence for patient safety? The study was carried out in a large acute NHS Trust in England with a capacity of 1,189 beds. The Datix data set consisted of data from April 2005 to July 2016 and included 133,893 incidents across 201 fields (approaching 27 million items of data). The Datix database was accessed via a locally hosted MySQL database and analysis routines comprised standard SQL query calls to the database for selection and aggregation followed by analysis using the Wolfram Mathematica 11.1 software. Python Scikit-learn (Pedregosa, 2011) was used for free text machine learning. Scikit-learn was used for the text analysis as the algorithms offered proved easier to use when extracting the key features from the Singular Value Decomposition (SVD) process.

Determination of available Datix field use in incident reports

Initial passes through the total data set were aimed at determining which of the available 201 fields were used routinely and robustly within the trust and which tier of the three tier Datix dataset would hold most validity. This analysis revealed that 17 of the available fields were always unused (defining unused as containing either Null, "" or " ") with a further 97 fields unused on 95% of occasions leaving 87 fields which were used on at least five percent of occasions.

To determine the utilization of the Datix system the absolute number of incidents and the average delay between an event occurring and being reported was examined over time.
To determine when events occurred the number of events occurring by hour of the day was examined.

The staffing and routinely collected safety KPI data (for example safety thermometer data spanning January 2013 to June 2014 from UHCW was overlaid with a sub-set of the Datix data from the same period. This time period was chosen as it had the best overlap with the KPIs supplied by UHCW. The analysis aggregates the data on the monthly and ward level (equivalent of 'inc locactual' a first-tier field) with the key limiting factor being the smallest time window available in the KPI data which was episodic. From previous investigation and on consultation with the steering group, the key parameters of interest were the absolute staffing levels and the ratio of health care support workers (HCSW): registered nurses (RN).

Text analysis

A selection of machine learning methods were applied to the analysis of the free form text fields with the aim of investigating if these could be used to determine the grade of event for each incident report. The model compares the predictability of 'high' vs 'very low' event grade to highlight the use at the extremities of the output space.

Machine learning (ML) refers to the field of computational tools and algorithms designed to allow computers to learn without being explicitly programmed. In general, machine learning algorithms are not considered in terms of T-test and F-tests but in how well the model can predict scenarios that were held back during the fit. This is an effect of the analytical approach; it is not a question of which signals are of significant effect but how robust and reliable are the predictions made.

The form of ML applied here is focused on classification, predicting the relative likelihood of competing assignments, in this case via a logistic regression model. The analysis pipeline will comprise the cleaning of the input text, projection into numeric features and then the optimization of the predictive model.

Projection of the text was done via the Latent Semantic Analysis (LSA) approach. This process is described in more detail elsewhere (Landauer *et al.*, 1998) but LSA) is a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text. LSA is an information retrieval technique which analyzes and identifies the pattern in unstructured collection of text and the relationship between them (Li 2018).. In this study we use the implementation offered by the Python Scikit-learn framework.

Ethical and governance considerations

This work was reviewed by the university ethics committee. The data used in this study was anonymized at source by the Trust which removed patient and staff identifiable fields. An algorithm was then constructed to recognize names and remove them from the free text comments. No patient or staff identifiable data left the Trust. This process was approved by the Calidicott guardian of the Trust and the Trust data governance group.

Results

Datix usage

The first aspect of the Datix data considered is how the event reporting system was first implemented, looking at uptake of the tool and impact on safety. This early time effect will be referred to as the uptake or 'burn in' period. Figure 1a) shows the variation in the number of reported incidents within the trust following the installation of Datix. Figure 1b) shows the average delay between an incident occurring and being reported.

In the period 2007 to 2016 there is a clear increase in events reporting, possibly arising from a combination of potential factors such as increasing occupancy, an uptake in reporting culture, improved access to the reporting system, staff training or more incidents occurring. In addition to the number of events, Datix can also be used to track the delay between incidents occurring and being reported – as shown in Figure 1b). This shows a clear decrease in delay.

Figure 1 here

Severity and likelihood

Figure 2 shows the variation over time of events divided by ordinal classification. The data has been aggregated by ordinal variable with those terms not classified by this system combined as ''Null' (e.g. "", " ", "Null", etc.). The first clear feature of reporting is that at the initial introduction of Datix the majority of events were graded as having 'Null' harm, and 'Null' likelihood. It is not until the 2006-2007 time period that there appear to be a concerted effort to use both the severity and likelihood fields.

Over the course of the data, the majority of events appear to be consistently reported at the lowest levels of severity and grade (negligible/minor and very low/ low respectively). Looking across the full data set, excluding 'Null' values, these reporting levels account for approximately 90% of all incidents. On inspection, these low harm fields appear to have increased over time.

Figure 2 here

Relationship between the number of incidents reported by year

The correlation of number of incidents reported as a function of time for the full data set are summarized in Table I. In the majority of cases there appears to be a significant positive correlation – implying more events are being reported over time. The key exceptions to this are the decrease in events reported as 'Null' and 'Rare' likelihood which have significantly decreased in absolute incident count, and the moderate and high grades which remain constant.

Table i here

Breakdown of events over time

In addition to the date stamp of when events occurred, Datix includes a field for the incident time. Figure 3 shows the hour-by-hour breakdown for all events reported with a time stamp, rounded to the nearest hour. The data was fit with a variety of models comprising a step function, offset, and linear background with and without an offset in time. The optimal model, judged using the F-test, is over-laid on Figure 3 and takes the form:

$$Events(t) = \begin{cases} 2200 + \dfrac{5500}{1 + e^{-2(t-8)}}, & t < 8 \\ 2200 + \dfrac{5500}{1 + e^{-2(t-8)}} - 300(t-8), & t > 8 \end{cases}$$

Where *t* is the time post-midnight. The model can be interpreted as comprising a base line of approximately 2000 events, a step function centered at 8am of approximately 6000 events, followed by a gradual decrease until midnight.

Figure 3 here

The hourly breakdown of events can also be applied to the individual types of adverse event occurring on the wards. Figure 4 shows the breakdown by hour for the three most commonly reported adverse events in UHCW over the dataset. These were 'Falls', 'Ulcers' and the 'TAdmin' first tier field (TAdmin is a field associated with treatment errors). The data has been presented in two forms – both as absolute values and as a proportion of the events reported within that time window.

Figure 4 here

From the data displayed in Figure 4 a number of features of interest can be observed. Overall falls are by far the most prominent adverse effect reported and their occurrence remains constant throughout the 24-hour period, however they make up the majority of incidents in the early hours of the morning, around five am. In the period midnight to seven am pressure ulcers and TAdmin errors are unlikely to be reported. There is a steep increase in the number of pressure ulcers and TAdmin events being reported between seven am and 10am. In the final period from 10am to midnight the number of pressure ulcers being reported declines slowly whereas the number of TAdmin errors declines steeply.

Analysis of the integration of incident report data and staffing/ key performance indicator data sets

The relationship between the absolute number of events opened is compared to the achieved staffing level on ward is shown in Figure 5. The data appears to show two distinct regions divided in the region 50 - 60 whole time equivalents (WTE). Below 50 WTE there is a strong positive correlation, with an increase in reporting with staffing and above 60 WTE there is a sudden drop to a lower staffing independent rate. To investigate this behavior the data was fit with a regional dependent linear model of the general form:

$$Events(x) = \begin{cases} m_0\, x + c_0\,, & x < x_B \\ m_1\, x + c_1\,, & x \geq x_B \end{cases}$$

where x_B is the boundary term between the two regions. The optimal model was found to Model 2 which is a first order polynomial below 53.1 WTE and a flat background above. This behavior is included on Figure 5 with the coefficients summarized in Table II.

Table II here

The relationship between the absolute number of events opened is compared to the achieved staffing level on ward is shown in Figure 5. The data shows an increase in reporting with staffing, with a significant positive correlation, most notably in the region 10 – 60 WTE posts. Interestingly, it appears that above 60 WTE the number of events undergoes a rapid decline, though the sparsity of data in this region limits the reliability of any analysis on this region. However, this is worthy of further scrutiny.

Figure 5 here

Prediction of grading from text

The hyper-parameter search resulted in an optimized model with average testing accuracy of 80.3%. A sub set of the models are summarized in Table III, giving the best and worst performing parameters judging by the average of the test accuracy over all cross-validation sets. The key features to the optimal performance appear to be the 'number of topics' (with an increase in accuracy when more features are projected out of the free text) and a lower 'Minimum Document Frequency' (Min DF) implying better performance when uncommon words are included.

The key feature to the optimal performance appears to be the number of topics, with the worst performing being those at the lower end of the numeric features projected out of the SVD method and a lower cut-off for the minimal document frequency (DF) for the tf-idf vectoriser.

Table iii here

The optimal model showed an accuracy of 79% when trialed on the full optimization set and 79% for the validation set. The model shows good performance for both 'high' and 'very low' states.

Implications for Research

Mining routinely collected incident data can provide intelligence for safety in healthcare. This data is collected by many types of healthcare organizations but is rarely used beyond the investigation of serious incidents. It is not utilized fully in healthcare but appears to be a valuable source of insight into both safety and workforce issues. Healthcare organizations appear to lack the resource in terms of time and expertise to extract and analyze these data and therefore automating these processes might make this intelligence more accessible to healthcare decision makers. Fuller use of incident reporting systems, could be employed as a strategy to enhance a culture of learning and improvement within healthcare. Such intelligence can also be used to inform workforce planning, for example in this study higher registered nurse staffing was associated with increased reporting.

Conclusions and recommendations

There was a clear pattern to the uptake period post deployment of Datix in the trust. The majority of fields (around 52%) available within the Datix database remained unused on 95% of entries in this study. When coupled with the time pressures on ward staff this raises a question regarding the optimization of data collection. It may be possible that the current form of input can be simplified by removing some of these apparently vestigial fields. A consensus-building study recommended that it was important to standardize and link data sets (84.6% and 73.1% agreed respectively) and educating staff on the quality of reports was most useful (77% agreed) (Howell *et al.*, 2017).

This study was able to reveal patterns that could be exploited in order to improve care. A number of relationships between staffing, rates of reporting and nature of incidents was revealed and could be tested in a larger dataset and diverse organisations for generalizability. The temporal patterns which show when incidents occur could be used to focus staffing for example. The proportion of falls compared to TADMIN errors and pressure ulcers for example is higher between midnight and seven am, though the absolute number of falls is relatively constant over the 24-hour period. Falls at night have been reported to result in more severe injury (Lopez-Soto *et al.*, 2016) thus using this data to improve care and review staffing levels at times of higher risk is likely to have direct benefit in terms of improving safety and outcomes.

KDD techniques applied to incident reporting systems data provide opportunities for learning at an organizational level. In order to fully understand the implications of traditionally visualized data, statistics should be applied to enable decision makers to more readily distinguish genuine signals from the noise both at the board level and on the front line (Goodwin *et al.*, 2003, Hazan 2016).

An increase in low grade events, while high harm events remain constant, would be consistent with an improvement in the general reporting culture assuming that the lower the harm of an event the less likely it is to be reported (Shaw *et al.*, 2005). In addition, if we consider that the highest severity events have increased, at the lowest rate, yet the grades of the highest are constant this suggests that the safety measures in place are well designed and keep high harm events occurring with low likelihood (as grade is a combination of likelihood and severity). These two in combination suggest that the trust is developing a culture of safety - regularly reporting events across all levels of harm and reducing the highest grades as a proportion of events since 2006.

When patients are becoming mobile in the early morning and overnight they are vulnerable to falling – with other adverse events appearing to either decrease in likelihood or go undetected. Pressure ulcers, despite being likely to have developed over several hours, are rarely detected overnight and are most often detected during the 7am to 10am period which may be due to staff become aware of pressure damage, for example, when washing patients.

The period between seven am and 10 am is clearly where the demand on the staff appears to change the most rapidly with the type and rate of adverse events changing more rapidly than at other times during the 24-hour period. These patterns all demonstrate key features to highlight on the ward level.

As regards staffing levels an increase in events as a function of absolute staffing leads to a series of competing hypotheses – either that registered nurses cause more adverse events to occur or that the presence of more nurses leads to an increase in awareness and reporting. Once a tipping point has been reached there is a falloff in incidents which could be due to increased vigilance.

The machine learning applications employed to investigate the free text comments suggest that firstly there appears to be a routine and reliable difference in the type of language used between events at the extremes of the reporting scale that can be detected by machine learning techniques. Furthermore, in addition to being able to advise on the grading of events (theoretically decreasing the time required to report an incident and increasing consistency of event grading between staff) this form of analysis can also be used to report key phrases and inform briefings on what the underlying causes are at different levels of harm/severity.

It is possible, given a larger data set, that the algorithm could be trained by locality/ medical specialty and used to produce key phrases indicative of each level of harm. Not only serving as a tool to assist at a clinical level but advising managers on underlying topics that may otherwise go unnoticed.

A key point is that these approaches are self-optimizing, given access to a database they can be trained and set to report with minimal operator interface with the key limiting factor being the regularity with which a reasonable quantity of data is input to affect the output state.

Due to the relatively small scale of events reported at the harmful end of the scale the available training set is too small to form a coherent, reliable model. The current limiting factors to increased performance appear to be threefold: firstly, the limited size of data, with only approximately 1,100 events graded as “high”. Secondly the subject specificity of the text encoding could be improved and finally there is the issue of inconsistency in the way in which the text fields are used, with variation in input length, and terminology between staff. Even with these caveats there is a great deal of potential to develop this in future as this was one relatively small dataset.

Fuller use of incident reporting systems, such as Datix, could be employed as a strategy to enhance a culture of learning and improvement within ultra-safe industries.

There remains an issue with data quality and specificity in routinely collected NHS Data (Leary *et al.*, 2017). The Datix common classification system for example is based on medical work and harms which means the nuance of the wider workforce is not collected.

These datasets have huge potential to improve the safety and quality of care for example linking to other patient level data (Leary *et al.*, 2016, de Vos *et al.*, 2018), but largely remain underutilized.

This short study has shown that there is insight to be gained by mining the Datix dataset in the acute setting and that while the incident reporting dataset is often simply warehoused the dataset examined in this study shows that such datasets in general have the potential to inform decision making and reduce harm. In summary, there are three potential benefits within the Datix set.

The first is a deeper understanding of the use of incidence reporting systems both in terms of usability and a possible reflection of culture. The second demonstrates the benefits of integration-by overlaying with other datasets such as vital signs and staffing, signals start to emerge which help focus areas of improvement or management of risk. In addition, this study illustrates that the possibility of using deep machine learning to move to predictive modelling appears realistic with the subsequent development of a dictionary/ontology.

The Datix database represents a large, untapped source of knowledge on patient safety, which could, through the aggregated analysis of the entire database, contribute to alleviating blame culture still experienced across the NHS (Hazan, 2016) and derive organisation-level relationships that currently appear undetectable.

Strengths and limitations of this study

One of the strengths of the study is that it has demonstrated that using data mining methodology to examine patient safety is feasible if suitable data sets are available. It has further shown that overlaying of other data sets to the data collected from an incident reporting system can help focus on areas which are at risk or may require improvement. The study is strengthened by the employment of patient level data and has revealed patterns in the data obtained.
Limitations of the study include the possibility of incomplete or incorrect completion of input into the incident reporting system by users, that the data is not publicly available and that the study is based on a limited data set from a single NHS Trust in England rather than a number of Trusts across the UK.

Ethical approval

This work was reviewed by the university ethics committee. The data used in this study was anonymized at source by the Trust which removed patient and staff identifiable fields. An algorithm was then constructed to recognize names and remove them from the free text comments. No patient or staff identifiable data left the Trust. This process was approved by the Caldicott guardian of the Trust and the Trust data governance group.

Availability of data

The data that support the findings of this study are available from UHCW NHS Trust but restrictions apply to the availability of these data as they are not publicly available. Data are however available from the authors upon reasonable request and with permission of the originating NHS Trust.

References

Almasalha, F., Xu, D., Keenan, G. M., Khokhar, A., Yao, Y., Chen, Y.C., Johnson, A., Ansari, R. and Wilkie, D.J. (2013), 'Data mining nursing care plans of end-of-life patients: a study to improve healthcare decision making'. *International Journal of Nursing Knowledge,* Feb;24(1):15-24.

Bates, DW, Heitmueller, A, Kakad, M, et al. Why policymakers should care about 'big data' in healthcare. Health Policy Technol 2018; 7(2): 211–216

Breault, J., Goodall, C. and Fos, P. (2002), 'Data mining a diabetic data warehouse'. *Artificial Intelligence in Medicine*, Sep-Oct;26(1-2):37-54.

Brennan, P. F. and Bakken, S. (2015), 'Nursing Needs Big Data and Big Data Needs Nursing', *Journal of Nursing Scholarship'.* Volume47, Issue5; 477-484

Cook, R. M., Jones, S., Williams, G. C., Worsley, D., Walker, R., Radford, M., & Leary, A. (2019). An observational study on the rate of reporting of adverse event on healthcare staff in a mental health setting: An application of Poisson expectation maximisation analysis on nurse staffing data. Health Informatics Journal. https://doi.org/10.1177/1460458219874637

Datix. Personal communication regarding number of sites deployed, December 2019

Department of Health. (2015), 'Culture Change in the NHS: Applying the Lessons of the Francis Enquiry', https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/403010/culture-change-nhs.pdf, accessed June 2017.

de Vos, M.S., Hamming, J.F., Chua-Hendriks, J.J.C. and Marang-van de Mheen, P.J. (2018), 'Connecting perspectives on quality and safety: patient-level linkage of incident, adverse event and complaint data'. *BMJ Quality & Safety*, Jul 21. pii: bmjqs-2017-007457. doi: 10.1136/bmjqs-2017-007457. [Epub ahead of print].

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996), 'From Data Mining to Knowledge Discovery in Databases'. *AI Magazine,* Volume 17, No 3 pp37-54.

Goldsmith, P., Moon, J., Anderson, P., Kirkup, S., Williams, S. and Gray, M. (2015), 'Do clinical incidents, complaints and medico legal claims overlap?'. *International Journal of Healthcare Quality Assurance,* Vol 28 (8) 864-871.

Goodwin, L., Vandyne, M., Lin, S., and Talbert, S. (2003), 'Data mining issues and opportunities for buildings nursing knowledge'. *Journal of Biomedical Informatics*, Aug-Oct;36(4-5):379-88.

Hazan, J. (2016), Incident reporting and a culture of safety'. *Journal of Patient Safety and Risk Management,* Volume: 22 issue: 5-6, page(s): 83-87.

Hogan, H., Healey, F., Neale, G., Thomson, R., Vincent, C. and Black, N. (2013), 'Preventable deaths due to problems in care in English acute hospitals: a retrospective case record review study', *BMJ Quality & Safety*, Feb; 22(2):182.

Howell, A.M., Burns, E.M., Bouras, G., Donaldson, L.J., Athanasiou, T. and Darzi, A. (2015), 'Can Patient Safety Incident Reports Be Used to Compare Hospital Safety? Results from a Quantitative Analysis of the English National Reporting and Learning System Data'. *PLoS One*. Dec 9;10(12).

Howell, A.M., Burns, E.M., Hull, L., Mayer, E., Sevdalis, N., and Darzi, A. (2017), 'International recommendations for national patient safety incident reporting systems: an expert Delphi consensus-building process'. *BMJ Quality & Safety*, Feb;26(2):150-163.

Hudson, P. (2003), 'Applying the lessons of high risk industries to health care'. *Quality Safety Health Care*;12(Suppl 1):i7–i12.

IBM Research Series. (2012), 'Knowledge Discovery Through Data Mining', http://researcher.watson.ibm.com/researcher/view_group.php?id=144, accessed December 2014

Landauer, T.K., Foltz, P.W. and Laham, D. (1998), 'An introduction to latent semantic analysis'. *Discourse Processes*, Volume 25, Issue 2-3: pp 259-284

Leary, A., Cook, R., Jones, S., Smith, J., Gough, M., Maxwell, E., Punshon, G. and Radford, M. (2016), 'Mining routinely collected acute data to reveal non-linear relationships between nurse staffing levels and outcomes'. *BMJ Open*; Dec 16;6(12)

Leary, A., Tomai, B., Swift, A., Woodward, A. and Hurst, K. (2017), 'Nurse staffing levels and outcomes – mining the UK national data sets for insight'. *International Journal of Health Care Quality Assurance*, Vol. 30 Issue: 3, pp.235-247.

Leary A, Dix A (2018) Using data to show the impact of nursing work on patient outcomes. Nursing Times [online]; 114: 10, 23-35

Li S (2018) Latent Semantic Analysis & Sentiment Classification with Python Towards DataScience https://towardsdatascience.com/latent-semantic-analysis-sentiment-classification-with-python-5f657346f6a3 accessed December 2019

López-Soto, P.J., Smolensky, M.H., Sackett-Lundeen, L.L., De Giorgi, A., Rodríguez-Borrego, M.A., Manfredini, R., Pelati, C. and Fabbian, F. (2016), 'Temporal Patterns of In-Hospital Falls of Elderly Patients'. *Nursing Research*, Nov/Dec;65(6):435-445.

Lugg-Widger, FV, Angel, L, Cannings-John, R, et al. Challenges in accessing routinely collected data from multiple providers in the UK for primary studies: managing the morass. Int J Popul Data Sci 2018; 3(3): 2.

NHS Quality Observatory. (2013), 'Safety Thermometer', https://www.safetythermometer.nhs.uk/, accessed Feb 2017.

NHS Improvement (2017) Patient Safety Reporting https://improvement.nhs.uk/resources/report-patient-safety-incident/ accessed December 2019

NHS Improvement (2018) A Just Culture Guide https://improvement.nhs.uk/resources/just-culture-guide/ accessed December 2019

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011), 'Scikit-learn: Machine Learning in Python'. *Journal of Machine Learning Research*, 12(Oct):2825−2830, 2011

Rabel, L.I., Gaardboe, O. and Hellebek, A. (2017), 'Incident reporting must result in local action'. *BMJ Quality & Safety*;**26**: 515-516.

Reason, J. (2000), 'Safety Paradoxes and Safety Culture'. *Injury Control and Safety Promotion*, 7(1) pp3-14

Sacristan, D. A. and Dilla, T (2015), 'No big data without small data: learning health care systems begin and end with the individual patient'. *J Eval Clin Pract.* Dec; 21(6): 1014–1017.

Schmidtke, K.A., Poots, A.J., Carpio, J., Vlaev, I., Kandala, N.B., and Lilford, R.J. (2017), 'Considering chance in quality and safety performance measures: an analysis of performance reports by boards in English NHS trusts'. *BMJ Quality & Safety*; Jan;26(1):61-69.

Shaw, R., Drever, F., Hughes, H., Osborn, S. and Williams, S. (2005), 'Adverse events and near miss reporting in the NHS'. *Quality and Safety in Health Care*, Aug;14(4):279-83.

Sivarajah, U., Kamal, M. M., Irani, Z. and Weerakkody, V. (2017), 'Critical analysis of Big Data challenges and analytical methods'. *Journal of Business Research*
Volume 70, January: 263-286

Staggs, V. and Dunton, N. (2014), 'Associations between rates of unassisted inpatient falls and levels of registered and non-registered nurse staffing'. *International Journal for Quality in Health Care*, Feb; 26(1): 87–92.

WHO (2019) Patient Safety Fact File, World Health Organization, Geneva, Switzerland

Table I: Correlation coefficients for the number of incidents reported per day vs time with data aggregated by event rating. A positive value of r indicated an increase in reported events between 2005 and 2016 while a negative value of r indicated a decrease in reported events over the same period. Signals have been interpreted by considering if they would be detected on a smaller data set ('Strong' at n = 100, 'Moderate' at n = 365, 'Weak' for otherwise significant signals) at the α = 0.005 significance level.

Intensity	Rating	Pearson r	P value	Correlation interpretation
Severity	Null	-0.65	<0.001	Strong
	Negligible	0.52	<0.001	Strong
	Minor	0.66	<0.001	Strong

	Moderate	0.40	<0.001	Strong
	Major	0.13	<0.001	Weak
	Catastrophic	0.13	<0.001	Weak
Likelihood	Null	-0.29	<0.001	Moderate
	Almost certain	0.07	<0.001	Weak
	Likely	0.25	<0.001	Moderate
	Possible	0.54	<0.001	Strong
	Unlikely	0.43	<0.001	Strong
	Rare	-0.10	<0.001	Weak
Grade	Null	-0.29	<0.001	Moderate
	Very low	0.46	<0.001	Strong
	Low	0.40	<0.001	Strong
	Moderate	-0.01	0.26	None
	High	0.01	0.34	None

Table II: Comparison of linear model performance within a regional optimized fit. P values are taken from the F-stat, comparing variance with the next most complex model. The three models follow the form of the equation above excluding the marked terms and fitting in order of increasing complexity.

Model Term	Model 1	Model 2	Model 3
c_0	9.7(±0.4)	-3.8(±0.9)	-3.9(±0.9)
c_1	20.0(±0.4)	16.7(±0.7)	19(±2)
m_0	(excluded)	0.59(±0.03)	0.59(±0.03)
m_1	(excluded)	(excluded)	-0.03(±0.02)
x_B	34.2	53.1	53.11
Variance	64.79	51.83	51.75
F-stat	-	174.2	1.07
°Freedom ratio	-	1 / 697	1 / 696
P value	-	< 0.0001	0.30

Table III: Summary of extremities of the model performance for the machine learning pipeline over hyper-parameter space.

	Number of topics	Max DF	Min DF	Mean Test Accuracy
Worst 5:	11	0.94	0.023	0.730
	10	0.99	0.023	0.730
	10	0.72	0.020	0.731
	10	0.85	0.019	0.733
	11	0.87	0.019	0.734
Best 5:	35	0.98	0.0047	0.798
	38	0.99	0.0034	0.799
	38	1.00	0.0036	0.799
	39	0.88	0.0028	0.801
	28	0.87	0.0060	0.803

Figure 1: Break down of incidents as reported into Datix by date of event a) absolute number of events and b) average delay between an event occurring and being reported. Data has been smoothed to the monthly level.

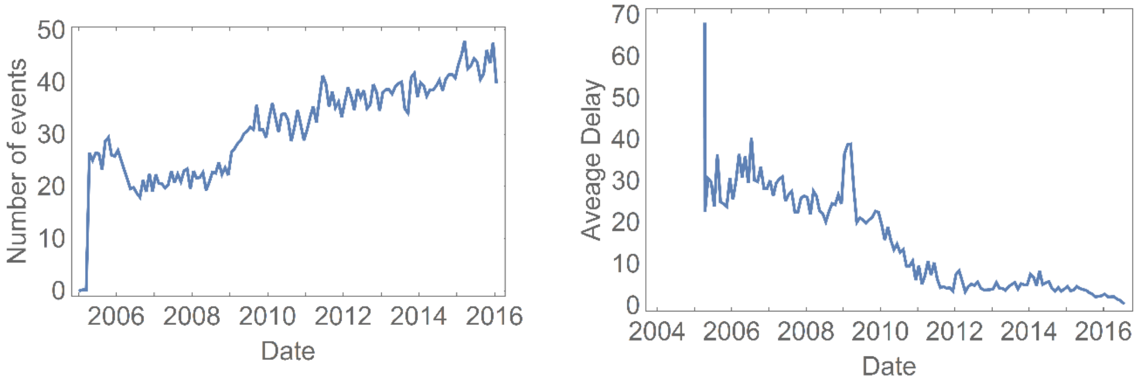


Figure 2: Breakdown of count of intensity of incident for severity (a) and b)), Likelihood (c) and d)) and grade e) and f)). The first figure for each category compares the number of null and not null events while the second compares the number of not null events by category.

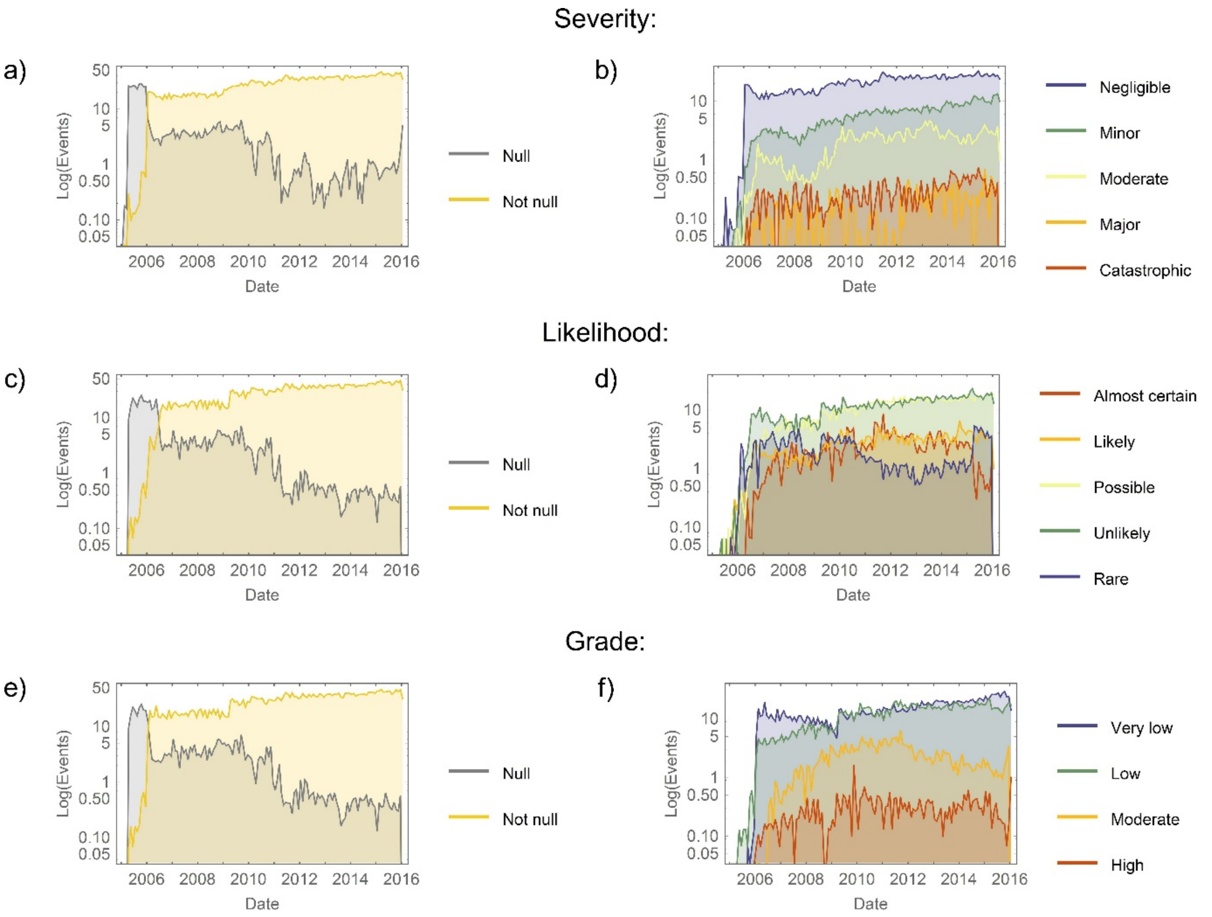


Figure 3: Hour-by-hour breakdown of all events reported with a time stamp. Error bars assume the data follows a Poisson distribution. The highlighted region reflects the period 7 –

10 am when the fastest increase in events occurs. The line represents the best model found from a combination of linear and logistic functions judging quality by the F-test.

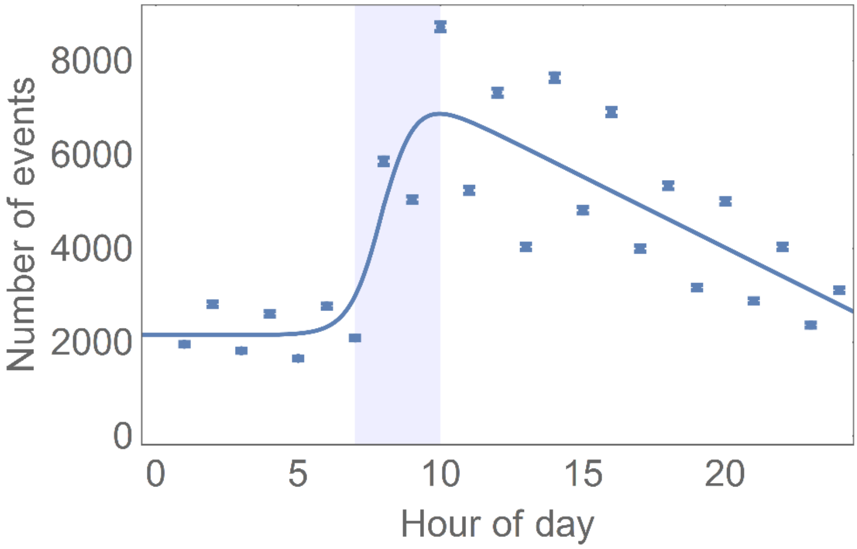


Figure 4: Hour-by-hour event count breakdown for the three most common clinical detail recorded in Datix as a) and b) falls, c) and d) pressure ulcers and e) and f) TADMIN. The first depiction in each category represents the total number of events while the second represents the proportion of events at each time period. Each line represents the best model found from a combination of linear and logistic functions judging quality by the F-test.

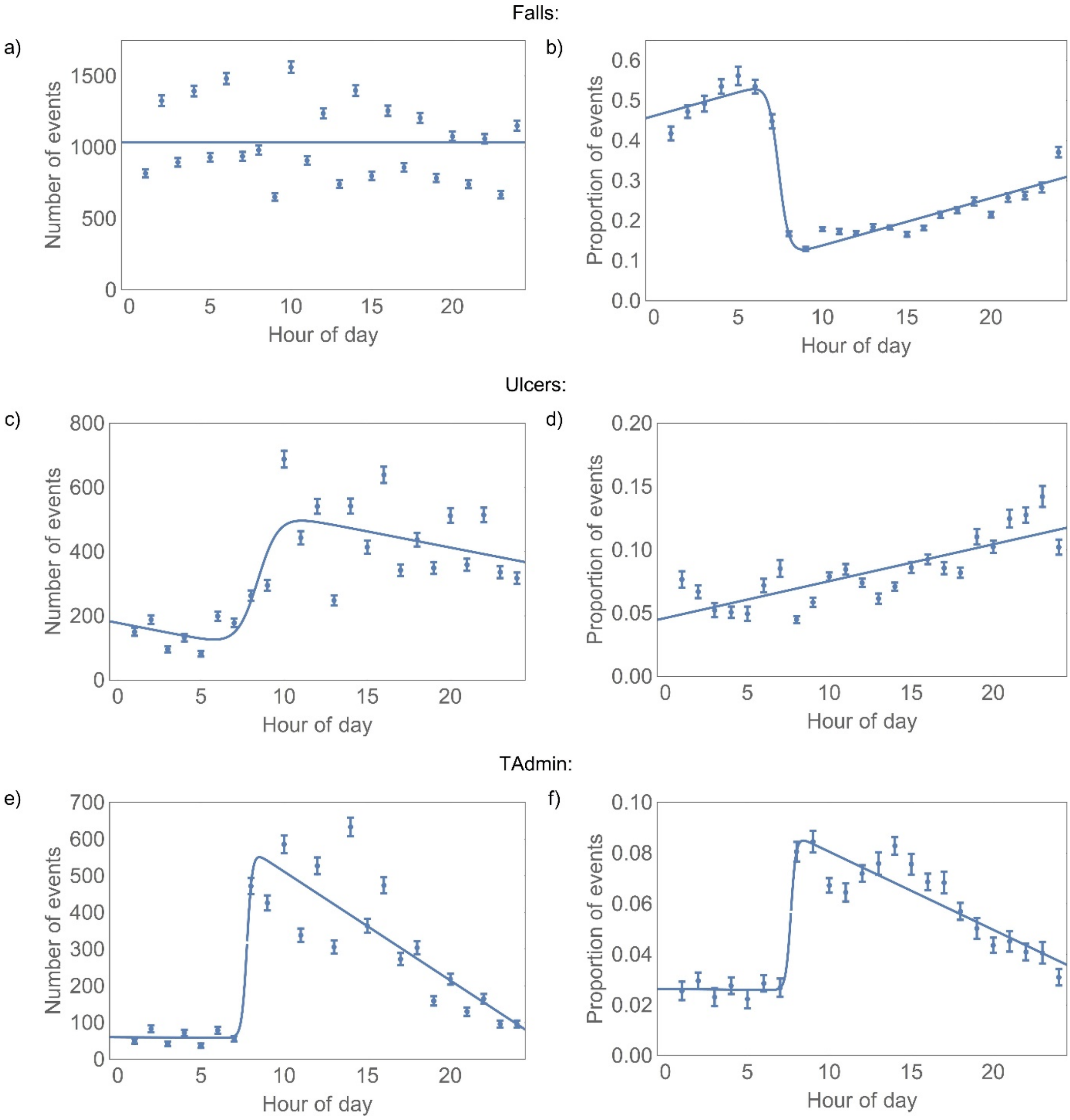
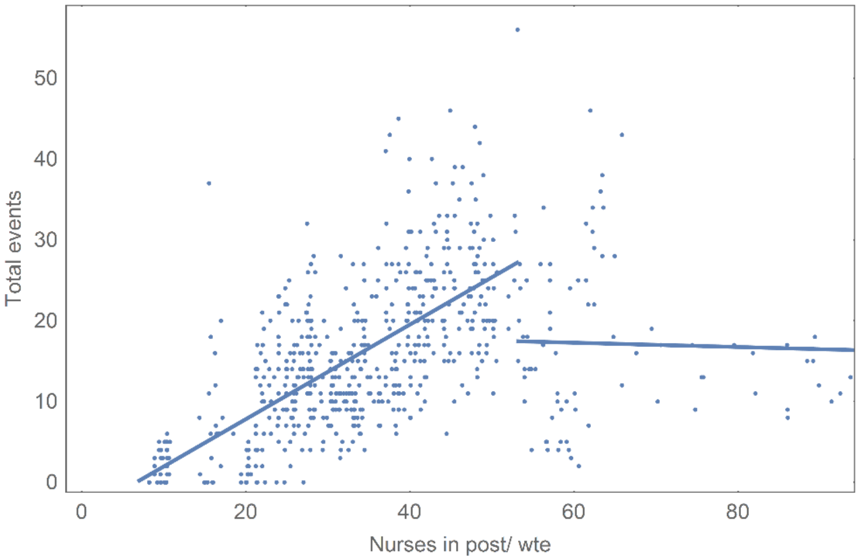


Figure 5: The variation in monthly reporting levels (total events) with nurses in post. The two least square regression lines show the general trend of the observations in regions 10-60 wte and above 60 wte.


Total events
0
10
20
30
40
50
0
20
40
60
80
Nurses in post/ wte