Medication-rights detection using incident reports: A natural language processing and deep neural network approach

Zoie Shui-Yee Wong
St. Luke’s International University, Japan

HY So
Belinda SC Kwok
Mavis WS Lai
David TF Sun
New Territories East Cluster (NTEC), Hong Kong

Abstract
Medication errors often occurred due to the breach of medication rights that are the right patient, the right drug, the right time, the right dose and the right route. The aim of this study was to develop a medication-rights detection system using natural language processing and deep neural networks to automate medication-incident identification using free-text incident reports. We assessed the performance of deep neural network models in classifying the Advanced Incident Reporting System reports and compared the models’ performance with that of other common classification methods (including logistic regression, support vector machines and the decision-tree method). We also evaluated the effects on prediction outcomes of several deep neural network model settings, including number of layers, number of neurons and activation regularisation functions. The accuracy of the models was measured at 0.9 or above across model settings and algorithms. The average values obtained for accuracy and area under the curve were 0.940 (standard deviation: 0.011) and 0.911 (standard deviation: 0.019), respectively. It is shown that deep neural network models were more accurate than the other classifiers across all of the tested class labels (including wrong patient, wrong drug, wrong time, wrong dose and wrong route). The deep neural network method outperformed other binary classifiers and our default base case model, and parameter arguments setting generally performed well for the five medication-rights datasets. The medication-rights detection system developed in this study successfully uses a natural language processing and deep-learning approach.
to classify patient-safety incidents using the Advanced Incident Reporting System reports, which may be transferable to other mandatory and voluntary incident reporting systems worldwide.

**Keywords**
artificial intelligence, hospital incident reporting, medication errors, patient safety, text mining

**Introduction**

The World Health Organization has regarded patient safety as a major public health issue that requires global attention.\(^1^,\(^2\) In March 2017, medication safety was launched as the third global patient-safety challenge that aims to promote medication without harm to the world. Errors in drug management pose a great risk to patients receiving medical care\(^3\) and the basic five ‘rights’ of medication administration remains a widely acceptable standard for safe medication practice. These include the right patient, the right drug, the right time, the right dose and the right route (additional rights include right duration, right documentation, right reason and right response). The overall goal is to assure that the right patient receives the right dose of the right drug via the right route at the right time.

Indeed, medication errors can be potentially prevented in many instances.\(^4\) It is important to learn from the causes, contributory factors, consequences and mitigating actions via incident reporting.\(^5^,\(^6\) Many developed countries/territories (such as the United Kingdom, Australia, Japan and Hong Kong) have set up national/regional medical incident reporting systems.\(^7^–\(^10\) These systems are designed to determine the background and causes of near-misses and actual incidents in a bid to prevent future occurrences. Typical incident report systems rely heavily on the free-text reporting to provide a description of incident background, outcomes and action. Presently many incident reporting systems in the world capture incidents of medication errors using free text without indicating the medication-rights errors structurally. The associated medication-rights information remains in the narrative text part of the incident reports. Analysis of reported incidents enables hospitals to implement measures to minimise risk and develop strategies to prevent recurrence. However, due to the large volume of textual data collected, identification of relevant reports that share the same markers can be very time-consuming and labour intensive. Text classification has proven to be a useful means of interpreting patient-safety reports.\(^7^,\(^8^,\(^10^–\(^13\) A previous probing study demonstrated that certain important event types, such as incidents caused by ‘look-alike, sound-alike’ medication, can be retrieved from a group of medication-incident-related documents using text classifier.\(^13\) The aim of this study was to develop a medication-rights detection system to classify medication incidents using the real-world incident reports collected by the Hong Kong Hospital Authority (HA). It is envisioned that the approach is transferable towards other mandatory and voluntary incident reporting systems worldwide.

The Advanced Incident Reporting System (AIRS),\(^9\) operated by the HA, is a centralised incident reporting system used to collect information on clinical and nonclinical incidents and near-misses in public hospitals in Hong Kong. The system has operated since 2000. Various standard reporting templates are available for the different types of incident that may occur in hospitals, such as medication-related incidents, falls, biomedical events, drug reactions, missing patients and suicide. In addition to the standard templates developed by the HA, text-input boxes are available for reporters to provide supplementary information in free-text form on a wide range of incidents. The AIRS database provides a very valuable source of information on how, why and how often patients are harmed. Compared with other types of incidents/near-misses, medication-related incidents have potentially more severe health consequences, and their analysis should thus be prioritised to improve
the quality of treatment and patient safety in hospitals. Using selected rights markers as class indicators for the free-text incident reports, we attempted to develop a text-based medication-rights detection system using advanced deep-learning methods to aid patient-safety information management. Our study also demonstrates deep-learning methods, can be practical candidate for incident classification and information retrieval, contributing to advance incident report learning.

**Background and significance**

Text classification\(^8,11,13\) remains the state-of-the-art quantitative method to analyse patient-safety reports.\(^5\) These classifiers would require a well-designed feature engineering to achieve reasonable classification performance. However, developing good engineering features is inherently time-consuming and highly domain/subject-specific. Modern deep-learning approaches are capable of bypassing tedious feature engineering steps and learning effective features from random assignments or word embedding methods.\(^14\) Deep Learning, typically achieved by artificial neural network (ANN) methods, makes sense of data by learning multiple levels of representation and abstraction.\(^15\) A ‘deep’ architecture is regarded as those neuron networks which have more than 1 hidden layer.\(^16\) Thanks to the advancement of computation power, deep learning has been receiving high attention in the recent decades. Typical deep-learning methods include deep neural networks (DNNs), recurrent neuron networks (RNNs)\(^17\) and bidirectional encoder representations from transformers (BERTs).\(^18\) Recently, the use of deep-learning algorithms has led to major advances in the extraction of intricate structures from large datasets, particularly those containing high-dimensional data,\(^19,20\) such as textual,\(^21\) image recognition and speech recognition\(^22\) applications. Deep learning, which has been successful in various domains, has also shown to be promising in health informatics applications.\(^17\)

DNN is a kind of the deep-learning method. Fundamentally, DNN is an ANN that attempts to make sense of data in multiple levels corresponding to different levels of abstraction and it comprises multiple hidden layers of neurons between the input and output layers. Starting from an input layer, DNNs consist of layers of interconnected neurons to match the feature space with a classification layer via a nonlinear activation function. DNN, a multilayer perceptron method, uses feedforward networks of connected layers with no loops.\(^23\) These act as classifiers and make different contributions to prediction outcomes. DNN can be used to solve binary and multi-class classification problems. Figure 1 shows the simplified structure of a DNN with a single output node for our study task. Several studies have suggested that DNNs outperform standard classifiers because they have more layers, enabling nonlinearity functions to be nested and the network to represent more complex functions. For instance, Iyyer et al.\(^24\) demonstrated promising results from applying a simple DNN to a typical text-classification problem and identified nonlinear input transformation as the factor responsible for the observed high performance.

Typical statistical classification techniques involve the training of high-dimensional data with very sparse features.\(^12,25\) However, due to the intensive computation required, training a DNN with large number of hidden layers is often time-consuming and it is difficult to reach convergence with some existing packages (such as neuralnet in R). It was only recently that the algorithmic advances of the Java-based H2O platform,\(^23\) which is an open-source machine-learning and deep-learning platform, have allowed supervised training of DNN models more efficiently. The platform runs scalable machine-learning and DNN models and allows installation via R and Python platforms. The H2O R package\(^26\) can be downloaded from CRAN to control the H2O engine and run high-performance deep-learning models. The use of deep learning as well as DNN has not yet received attention as a candidate method of medical incident text classification for improving patient-safety information retrieval. The DNN model we developed in this study is a feedforward ANN model
The key innovation of this study is the exploration of efficient medication-incident classification with DNN method using hospital incidents on five medication rights. The aim of this study was to develop a medication-rights detection system via DNN that is able to automate medication-rights incident identification using free-text incident reports. Classification of incident occurrence is an important means of quantitatively analysing a large amount of mined textual data on reported patient incidents. The state-of-the-art studies used machine-learning models, such as logistic-regression models, support vector machines (SVMs) and decision trees, to model high-level abstractions of textual incident data. The use of DNN, which has had promising outcomes in other complicated machine-learning fields, has to date received little attention as a method of patient-safety information retrieval. We developed the DNN medication-rights classification as a means of analysing AIRS medication-incident reports and to compare the performance of DNN models with that of other standard classifiers. We also aimed to provide practical guidelines for training DNN models to classify patient-safety incidents, which may be transferable to other mandatory and voluntary reporting systems worldwide for effectively understanding medical incidents.

**Materials and methods**

**Datasets**

Our medication-incident textual data were obtained from the New Territories East Cluster (NTEC) through the HA database. The NTEC consists of seven hospitals providing a wide range of medical services. The incident reports were extracted from AIRS version 2 (AIRS v.2) and AIRS version 3 (AIRS v.3) in November 2014. AIRS v.2 was being remodelled to the current AIRS v.3 edition in 2013, with a similar basic framework and common textual input items. However, AIRS v.3 (the latest version) captured less structured items than AIRS v.2 (the former version) to simplify the reporting workload of clinical staff. In the medication-incident template, both versions collected structured data attributes and unstructured free text to record patient demographic, medication name, dosage used, a brief description of the incident and the immediate action taken. We only selected mutual structured
data attributes and unstructured textbox items across the two systems for our analysis (specifically, the medication-rights markers and the free-text summary of the medication incident). The AIRS v.2 reports covered all medication incidents between January 2011 and December 2013, and the AIRS v.3 reports covered medication incidents between January 2014 and September 2014.

We analysed only reports in English; all reports containing Chinese characters were excluded. Only incidents associated with medication errors and near-misses were selected. A clinical near-miss was defined as an incident corrected before it affected the patient, and a clinical incident was defined as an event that affected a patient and did or did not cause harm to the patient. All of the medical incident data were anonymised by truncating the associated patient names and removing other identifying information before analysis. A member of staff at the HA double-checked that the dataset contained no individually identifiable information. Each of the reports addressed one single medication incident that may have involved one or more than one medication-rights markers (coded as wrong patient (WP), wrong drug (WD), wrong time (WT), wrong dose (WE) and wrong route (WR)) at the same time. The medication rights were directly denoted by the same attribute names in the ‘type of event’ reporting section, as shown in the AIRS user interface in Figure 2. The research was approved by the Joint Chinese University of Hong Kong–NTEC Clinical Research Ethics Committee (CREC). The clinical research approval number awarded by the CREC was No: 2014.470.

**Experimental workflow**

The experimental workflow comprised four steps: structuring unstructured textual data by natural language processing (NLP) method, feature selection, DNN model development and performance evaluation. Both AIRS v.2 and v.3 text datasets contained clear indications of medication-rights
markers when the incident reports were captured at the hospital level. The dependent variable (output node) of each medication-rights marker model was a binary response variable (i.e. positive (cases) or negative (no cases)) and depended on the rights marker undertaken. We used AIRS v.2 incident reports to predict the marker class of each AIRS v.3 medication-incident report. Subsequently, we used the trained model to predict the medication-rights marker using the training set text reports (AIRS v.3) and compared the prediction outcomes against the true (originally indicated) medication-rights markers. The workflow is illustrated in Figure 3. The analysis was performed using an R v.3.4.1 64-bit platform with the packages H2O, tm, snowball and Rstem on a desktop personal computer (processor: Intel® Core™ i3-2330M; CPU: 2.2 GHz).

Structuring incident text. We used NLP algorithms written in R to structure the free-text descriptions provided in the AIRS reports. We followed the text-mining procedure given in Wong13 to produce structured summaries of the AIRS free-text data. The original incident reports were processed by removing English stop words, converting words to their stems, converting words to lower case, trimming white space and removing numbers and punctuation. The resulting structured data summarising the term-document matrices of AIRS v.2 and AIRS v.3 were prepared as the training and testing datasets, respectively.

Model input/feature set. The inputs of the models came from the analysed ‘bags of words’ (bow)8 of the training set text reports (AIRS v.2). As a result, different medication-rights markers have different feature sets (e.g. the feature set of WP is different from WE because the free text used in WP is different from that used in WE). These reports were subjected to the text conversion and structuring procedure described above. The structured text outcomes (i.e. bow) were the source of the classifier’s input variables. We developed models for all five markers independently using the

---

**Figure 3.** Experimental workflow.
same experimental workflow. Finally, five bow were prepared for the entire analysis: one bag (bow) each for WP (bow_WP), WD (bow_WD), WT (bow_WT), WE (bow_WE) and WR (bow_WR). Word weighting was determined by calculating term frequency–inverse document frequency (TFIDF).\textsuperscript{27} TF refers to the number of times that each stemmed term occurs in the corpus and the IDF was calculated as $\log \frac{N}{N_t}$, where $N$ denotes the total number of documents in the corpus and $N_t$ the number of documents in which the term appears.

**DNN basic model.** The deep-learning computations were performed in the H2O cluster and initiated by the H2O R package.\textsuperscript{23,26} Beginning with an input layer that matches the feature space, DNN applications require weight matrix and column vector of biases to be computed in each layer and the representation in each layer to be differentiated from the representation in the previous layer. The weighted combination $\alpha = \sum_{i=1}^{j} (w_i x_i + b)$, where bias $b$ accounts for the neurons’ activation threshold, can be defined. Function $f$ represents the nonlinear activation function and it produces signals, $f(\alpha)$, to match the output space for classification.\textsuperscript{23} For each observation, $n$, the objective is to minimise the loss function, that is, $L(W,B \mid n)$. $W$ is the collection of $W_i$ that denotes the weight matrix between $i$ and $i+1$ layers for an $M$ layers network, whereas $B$ is the collection of $b_i$ that refers to the column vector of biases for layer $i+1$. To ensure the stability of the experimental results, we made the DNN model function reproducible by setting a fixed seed and running the model on one thread pool during the training phase.

We used a purely supervised training protocol provided by H2O that allows every node to train on the entire dataset. The default suggested distribution functions that primarily associate with each loss function were employed. We launched the training dataset and testing datasets to the H2O platform and preprocessed and standardised our input layer nodes into a standard normal distribution of $N(0, 1)$ to ensure its compatibility with the activation functions (because activation functions cannot map into the full spectrum of real numbers). The second phase of standardisation process was then carried out after the network propagation phase. The model performance metrics of the trained model were captured using the validation data as the programme output.

**Parameters.** When training a DNN model, a number of possible parameter arguments must be specified. We aimed to explore a choice of algorithms, parameters and settings (including activation functions, hidden-layer size, dropout control and regularisation) for DNN medication-incident classification. We varied the following key parameter arguments one at a time:

- **Number of layers and hidden-layer size:** We varied the level of complexity of the DNN model by the number of hidden layers (1–5) and number of neurons in each hidden layer (50, 100, 150, 200, 250 and 300). We only tested the same number of neurons in each layer if there was more than 1 hidden layer. The base case number of hidden layers was 2 and the number of neurons was 200 in each layer.
- **Activation and loss functions:** We examined the impact of various nonlinear activation functions, namely Tanh, Rectifier and Maxout. The mathematical formulae of the activation

| Function  | Formula |
|-----------|---------|
| Rectifier | $f(-) = \max(0,-)$ |
| Maxout    | $f(.) = \max(w_i x_i + b)$, rescale if $\max f(.) \geq 1$ |
| Tanh      | $f(\alpha) = \frac{(e^\alpha - e^{-\alpha})}{(e^\alpha + e^{-\alpha})}$ |

Table 1. Mathematical formulae for activation functions.
functions are shown in Table 1. The selected activation function was applied to process the nested network values in the output classification layer. We only selected one activation function once at a time for the experiment. The base case activation function is Tanh.

- **Dropout regularisation control and ratio:** In this study, regularisation by dropout\textsuperscript{28} was considered. Dropout control generally allows averaging a large number of models as an ensemble during forward propagation. It helps to improve model overfitting and increase DNN model generalisability. The dropout function is controlled by the activation function argument as ‘TanhWithDropout’, ‘MaxoutWithDropout’ or ‘RectifierWithDropout’. We specified the amount of hidden dropout per hidden layer to be 0.5. The base case incorporates dropouts using Tanh function.

**Other classification methods.** We also compared the performance of the DNN classifier with that of several machine-learning methods (logistic regression, SVMs and decision-tree method), without heavy problem-specific tuning. We followed the same experimental workflow procedure given by Figure 3 for free-text extraction, feature selection and performance evaluation schemes. We developed these classifiers for each medication marker based on the R code used in the previous probing study (the detailed procedures are given in Wong\textsuperscript{13}). Due to a lack of existing results for the non-DNN classifiers, we used stratified (10-fold) cross-validation\textsuperscript{13} for training the AIRS v.2 data, and the resultant models were tested against the AIRS v.3 data. We compared the performance of the classifiers using ROC curves, which provide a graphical display of sensitivity on the \(y\)-axis against \((1 – \text{specificity})\) on the \(x\)-axis by setting the testing probability to 0.5.

**Performance evaluation.** We explored the effects of DNN model settings, including number of layers, number of neurons in each layer and activation algorithms, on the performance of the trained DNN model. We examined classifier performance using the metrics of specificity, sensitivity, F1 score, accuracy and area under the receiver operating characteristic (ROC) curve (AUC). These metrics were reported relative to the threshold of the highest F1 score. Specificity (or true-negative (TN) rate) measures the proportion of negatives that are correctly identified, whereas sensitivity (also known as recall or true-positive (TP) rate) measures the proportion of actual positives that are correctly identified, that is, the fraction of actual positives that are correctly identified. The geometric mean of specificity and sensitivity is also reported. Precision is also considered to be the positive predictive value that represents the fraction of retrieved instances that are relevant. The F1 score is the harmonic mean of precision and recall. Accuracy is a statistical measure of how well a classifier correctly identifies a condition. Since the dataset of this binary classification task is unbalanced, AUC, which indicates the overall ability of a classifier to discriminate between positive and negative cases, would be a fair measure to determine the DNN model performance. A perfect classifier, yielding zero false positives (FPs) and zero false negatives (FNs), would achieve an AUC value of 1.

**Patient and public involvement.** This is a retrospective study based on incidents reports collected from incident reporting system. Patients were not involved in the development of the research question and outcome measures.

**Results**

**Descriptive statistics**

We drew 348 reports from AIRS v.2 and 226 from AIRS v.3 as shown in Table 2. We summarised the information provided on patient gender, patient age, departments involved, event type, severity and the five medication-rights markers in the selected reports. A two-sided Fisher’s test was used
Table 2. Descriptive summary of incidents.

| Structured items                              | AIRS v.2 (n = 348) | %      | AIRS v.3 (n = 226) | %      | p-value |
|-----------------------------------------------|---------------------|--------|---------------------|--------|---------|
| Gender: male                                  | 153                 | 44     | 92                  | 41     | 0.490   |
| Patient age distribution (years)              |                     |        |                     |        |         |
| 0                                             | 10                  | 2.87   | 6                   | 2.65   | 1.000   |
| 1–10                                          | 17                  | 4.89   | 8                   | 3.54   | 0.533   |
| 11–20                                         | 2                   | 0.57   | 3                   | 1.33   | 0.387   |
| 21–30                                         | 7                   | 2.01   | 9                   | 3.98   | 0.196   |
| 31–40                                         | 22                  | 6.32   | 12                  | 5.31   | 0.719   |
| 41–50                                         | 26                  | 7.47   | 14                  | 6.19   | 0.614   |
| 51–60                                         | 38                  | 10.92  | 27                  | 11.95  | 0.788   |
| 61–70                                         | 37                  | 10.63  | 27                  | 11.95  | 0.684   |
| >70                                           | 169                 | 48.56  | 64                  | 28.32  | <0.001  |
| Unspecified                                   | 20                  | 5.75   | 56                  | 24.78  | -       |
| Department involved                           |                     |        |                     |        |         |
| Medicine and therapeutics                     | 80                  | 22.99  | 41                  | 18.14  | 0.175   |
| Surgery                                       | 5                   | 1.44   | 16                  | 7.08   | 0.001   |
| Medicine                                      | 16                  | 4.60   | 14                  | 6.19   | 0.445   |
| Accident and emergency                        | 35                  | 10.06  | 13                  | 5.75   | 0.089   |
| Paediatrics                                   | 17                  | 4.89   | 10                  | 4.42   | 0.843   |
| Psychiatry                                    | 8                   | 2.30   | 9                   | 3.98   | 0.314   |
| Medicine and therapeutics; pharmacy           | 7                   | 2.01   | 9                   | 3.98   | 0.196   |
| Obstetrics and gynaecology                    | 16                  | 4.60   | 9                   | 3.98   | 0.836   |
| Pharmacy                                      | 13                  | 3.74   | 6                   | 2.65   | 0.634   |
| Other                                         | 151                 | 43.39  | 99                  | 43.81  | 0.932   |
| Event type                                    |                     |        |                     |        |         |
| Administration                                | 169                 | 48.56  | 76                  | 33.63  | <0.001  |
| Prescription                                  | 147                 | 42.24  | 66                  | 29.20  | 0.002   |
| Dispensing                                    | 76                  | 21.84  | 35                  | 15.49  | 0.066   |
| Severity index                                |                     |        |                     |        |         |
| 0: Incident occurred but stopped before reaching patients. No consequence | 80 | 22.99 | 76 | 33.63 | <0.001 |
| 1: Incident occurred and reached patient but no injury sustained | 174 | 50.00 | 100 | 44.25 | 0.200 |
| 2: Minor injury                               | 60                  | 17.24  | 32                  | 14.16  | 0.353   |
| 3: Temporary morbidity                        | 16                  | 4.60   | 9                   | 3.98   | 0.836   |
| 4: Significant morbidity                      | 8                   | 2.30   | 3                   | 1.33   | 0.540   |
| 5: Major permanent loss of function/disability | 0 | 0.00 | 0 | 0.00 | 1.000 |
| 6: Death                                      | 1                   | 0.29   | 0                   | 0.00   | 1.000   |
| N: Not applicable                             | 4                   | 1.15   | 6                   | 2.65   | 0.203   |
| X: Not known                                  | 5                   | 1.44   | 0                   | 0.00   | -       |
| Medication-rights markers                     |                     |        |                     |        |         |
| WP: wrong patient                             | 38                  | 10.92  | 18                  | 7.96   | 0.254   |
| WD: wrong drug                                | 70                  | 20.11  | 36                  | 15.93  | 0.227   |
| WT: wrong time                                | 15                  | 4.31   | 5                   | 2.21   | 0.245   |
| WE: wrong dose                                | 44                  | 12.64  | 25                  | 11.06  | 0.602   |
| WR: wrong route                               | 8                   | 2.30   | 5                   | 2.21   | 1.000   |

AIRS: Advanced Incident Reporting System.
Each medication incident may involve one or more than one medication-rights markers. p-values are obtained using a two-sided Fisher’s test.
Table 3. Examples of AIRS WP incident reports (narrative text).

| Example    | Description                                                                                                                                 |
|------------|--------------------------------------------------------------------------------------------------------------------------------------------|
| Example 1  | This patient was given a prescription with the wrong patient name by Dr. K.                                                                |
| Example 2  | Patient A’s demographic data were accidentally used to prescribe medication for patient B; patient B received the wrong medication sheet and obtained the wrong drugs from the pharmacy. |
| Example 3  | A prescription bearing the wrong patient name was sent to the pharmacy. The correct patient name is S.                                      |
| Example 4  | Patient C was admitted to emergency medicine ward (EMW) at 11:00 p.m. One day after Patient C’s admission (around 4 p.m.), a nurse found that another patient’s electronic patient record (ePR) was attached to Patient C’s medical file and discovered that the prescribed drugs shown on patient Ms. C’s medication administration record (MAR) did not belong to the patient’s usual medication list. However, the medication had already been administered as scheduled, according to the MAR. |

AIRS: Advanced Incident Reporting System; WP: wrong patient.

Figure 4. Word cloud for AIRS medication-incident reports (created using the R wordcloud package; size of words stratified by word importance).

Table 3 shows examples of AIRS reports with the WP marker, and Figure 4 visualises ‘clouds of words’ of the AIRS medication incidents. The size and colour of the words are stratified by word importance based on TFIDF, using the R package ‘wordcloud’.

to determine whether the use of AIRS v.2 or AIRS v.3 was significant for any of these categories. The differences found in the proportions of AIRS v.2 and AIRS v.3 reports were nonsignificant for most of the categories (patient gender, severity and the five rights markers) but significant for the >70 patient age group and incidents classified as administration-related (we did not observe any apparent contributing factor for such differences).

Free-text summary

Table 3 shows examples of AIRS reports with the WP marker, and Figure 4 visualises ‘clouds of words’ of the AIRS medication incidents. The size and colour of the words are stratified by word importance based on TFIDF, using the R package ‘wordcloud’.
The 10 terms with the highest TFIDF used to describe medication incidents in the AIRS v.2 and AIRS v.3 reports are listed in Table 4 and include ‘drug’, ‘medi’, ‘nurs’ and ‘dose’. The 10 terms with the highest TFIDF used to describe the group of incidents with the WP marker included ‘drug’, ‘medic’, ‘bed’, ‘check’ and ‘found’. The terms with the highest TFIDF were fairly similar between AIRS v.2 and AIRS v.3, indicating that a limited number of terms were used repeatedly for incident reporting. The important keywords found in AIRS v.2 were also weighted heavily in AIRS v.3. This suggested that text classification using the terms of the training data set (AIRS v.2) as predictor variables for predicting the medication markers of AIRS v.3 was justifiable. We also reported the top 10 TFIDF terms in each of the other classes (WD, WT, WE and WR) in Table S1 of Supplementary Document 1.

DNN model performance with different settings

Table 5 shows the experimental results for WP class label. We used a DNN model with 2 hidden layers and 200 neurons in each layer to test the effects of different activation functions and dropout control. The best AUC value of 0.960 was achieved using a Rectifier algorithm with a dropout ratio of 0.5. Furthermore, we examine the number of hidden layers and neurons under the base case setting. In the WP class label experiments, the highest value for AUC (0.925) was in the model setting with 5 hidden layers and 150 neurons per layer and the highest value for accuracy (0.956) was found using the DNN model with 5 hidden layers and 250 neurons per layer. The average AUC using setting of Tanh with dropout ratio was 0.911 (standard deviation 0.019) and the average accuracy value was 0.940 (standard deviation 0.011). This indicated that the DNN model in general performed well in classifying AIRS v.3 reports based on AIRS v.2 reports. However, it is observed that the performance of different medication-rights detection would perform differently under the same modelling procedure. The optimal accuracy of all the medication rights ranged from 0.820 to 1, as shown in Table 6. These best trained DNN algorithms for each medication right are made publicly available from GitHub.

Comparison with other classification methods

As shown in Figure 5, the ROC curve obtained for the DNN model exceeded the space occupied by the curves for the other classifiers, which implied that the DNN model outperformed the standard classifiers across the entire range of class labels. AUC indicates the efficacy of a trained
Table 5. Experimental results for WP class label.

| DNN | Specificity | Sensitivity | GM      | F1 score | Accuracy | AUC   |
|-----|-------------|-------------|---------|----------|----------|-------|
| No. of hidden layers | Hidden-layer size (neurons) | Specificity | Sensitivity | GM      | F1 score | Accuracy | AUC   |
| 1   | 50          | 0.952       | 0.556   | 0.727    | 0.526    | 0.920  | 0.845 |
|     | 100         | 0.971       | 0.556   | 0.735    | 0.588    | 0.938  | 0.862 |
|     | 150         | 0.971       | 0.556   | 0.735    | 0.588    | 0.938  | 0.875 |
|     | 200         | 0.990       | 0.500   | 0.704    | 0.621    | 0.951  | 0.903 |
|     | 250         | 0.962       | 0.500   | 0.693    | 0.514    | 0.925  | 0.897 |
| Mean |             | 0.971       | 0.528   | 0.716    | 0.570    | 0.936  | 0.881 |
| SD  |             | 0.012       | 0.028   | 0.017    | 0.037    | 0.010  | 0.023 |
| 2   | 50-50       | 0.971       | 0.556   | 0.735    | 0.588    | 0.938  | 0.916 |
|     | 100-100     | 0.990       | 0.500   | 0.704    | 0.621    | 0.951  | 0.908 |
|     | 150-150     | 0.933       | 0.778   | 0.852    | 0.609    | 0.920  | 0.923 |
|     | 200-200     | 0.942       | 0.667   | 0.793    | 0.571    | 0.920  | 0.920 |
|     | 250-250     | 0.990       | 0.500   | 0.704    | 0.621    | 0.951  | 0.894 |
| Mean |             | 0.966       | 0.602   | 0.759    | 0.606    | 0.937  | 0.915 |
| SD  |             | 0.022       | 0.098   | 0.053    | 0.020    | 0.013  | 0.011 |
| 3   | 50-50-50    | 0.952       | 0.667   | 0.797    | 0.600    | 0.929  | 0.918 |
|     | 100-100-100 | 0.981       | 0.556   | 0.738    | 0.625    | 0.947  | 0.920 |
|     | 150-150-150 | 0.990       | 0.500   | 0.704    | 0.621    | 0.951  | 0.906 |
|     | 200-200-200 | 0.990       | 0.500   | 0.704    | 0.621    | 0.951  | 0.915 |
|     | 250-250-250 | 0.971       | 0.611   | 0.770    | 0.629    | 0.942  | 0.923 |
| Mean |             | 0.974       | 0.574   | 0.747    | 0.615    | 0.942  | 0.917 |
| SD  |             | 0.014       | 0.061   | 0.035    | 0.013    | 0.008  | 0.005 |
| 4   | 50-50-50-50 | 0.942       | 0.722   | 0.825    | 0.605    | 0.925  | 0.923 |
|     | 100-100-100-100 | 0.957 | 0.667 | 0.799 | 0.615 | 0.934 | 0.919 |
|     | 150-150-150-150 | 0.981 | 0.556 | 0.738 | 0.625 | 0.947 | 0.924 |
|     | 200-200-200-200 | 0.986 | 0.500 | 0.702 | 0.600 | 0.947 | 0.920 |
|     | 250-250-250-250 | 0.986 | 0.500 | 0.702 | 0.600 | 0.947 | 0.921 |
| Mean |             | 0.973       | 0.574   | 0.745    | 0.608    | 0.941  | 0.921 |
| SD  |             | 0.017       | 0.089   | 0.050    | 0.010    | 0.009  | 0.003 |
| 5   | 50-50-50-50-50 | 0.942 | 0.667 | 0.793 | 0.571 | 0.920 | 0.917 |
|     | 100-100-100-100-100 | 0.957 | 0.667 | 0.799 | 0.615 | 0.934 | 0.919 |
|     | 150-150-150-150-150 | 0.990 | 0.500 | 0.704 | 0.621 | 0.951 | 0.925 |
| Mean |             | 0.976       | 0.556   | 0.734    | 0.608    | 0.942  | 0.921 |
| SD  |             | 0.019       | 0.085   | 0.049    | 0.033    | 0.012  | 0.004 |
| Overall model settings | Mean | 0.972 | 0.567 | 0.740 | 0.601 | 0.940 | 0.911 |
| SD  |             | 0.018       | 0.080   | 0.045    | 0.030    | 0.011  | 0.019 |

(Continued)
classifier in discriminating between positive and negative cases using a training dataset. The AUC of the DNN model, 0.925, was also greater than that of the standard classifiers, again indicating that the latter were outperformed by the DNN model. We compared the accuracy of the models and found that the DNN model was consistently more accurate than the other classifiers across all of the tested class labels (i.e. WP, WD, WT, WE and WR).

Discussion

DNN models have been used in textual NLP and have achieved superior classification accuracy in other application areas. This study is by far the largest scale DNN modelling study examined on practical patient-safety incident reports. In this study, we developed a medication-rights detection system to automate medication-incident identification using free-text incident reports. The findings of this study indicate that a DNN model can be effectively and efficiently used to perform classification tasks on five rights medication incidents. It was shown that DNN method consistently outperformed other text classifiers (including logistic regression, SVMs and the decision-tree method). We suggest that DNN method has potential for the efficient exploration of textual reports of medication incidents. The default base case model and parameter arguments setting generally performed well for our dataset.

Table 5. (Continued)

| Activation function | Base DNN model – No. of layers: 2; Hidden-layer size: 200-200; Tanh with hidden dropout ratio of 0.5 |
|---------------------|-----------------------------------------------------------------------------------------------------|
| Tanh                | Dropout 0.942 0.667 0.793 0.571 0.920 0.920                                                                 |
|                     | No_DropOut 0.976 0.500 0.699 0.563 0.938 0.853                                                        |
| Rectifier           | DropOut 1.000 0.500 0.707 0.667 0.960 0.867                                                         |
|                     | No_DropOut 0.971 0.556 0.735 0.588 0.938 0.823                                                        |
| Maxout              | DropOut 0.957 0.722 0.831 0.650 0.938 0.913                                                         |
|                     | No_DropOut 0.981 0.389 0.618 0.483 0.934 0.807                                                        |
| Overall model settings | Mean 0.971 0.556 0.730 0.587 0.938 0.864                                                          |
|                     | SD 0.018 0.111 0.069 0.061 0.012 0.042                                                            |

DNN: deep neural networks; AUC: area under the curve; SD: standard deviation.
GM refers to geometric mean of sensitivity and specificity.

Table 6. Summary of the optimal model setting and accuracy for each medication right.

| Medication rights | Activation function | Dropout | No. of hidden layers | Hidden-layer size (neurons) | Optimal accuracy |
|-------------------|---------------------|---------|----------------------|-----------------------------|------------------|
| WP                | Rectifier           | Yes     | 2                    | 200                         | 0.960            |
| WD                | Rectifier           | Yes     | 5                    | 150                         | 0.820            |
| WT                | Tanh                | Yes     | 3                    | 50                          | 0.994            |
| WE                | Maxout              | No      | 4                    | 300                         | 0.974            |
| WR                | Tanh                | No      | 4                    | 300                         | 1                |

WP: wrong patient; WD: wrong drug; WT: wrong time; WE: wrong dose; WR: wrong route.
The optimal models (in binary form) are available in the following GitHub site: https://github.com/zoiewong/Optimal-5-medication-rights-models-for-incident-reports.
Both DNN methods and traditional classification methods have advantages and disadvantages. For instance, the coefficient of a logistic-regression model indicates the odds and probability of an event’s occurrence. However, in textual data applications, due to differences in syntactic composition and individual variation in language usage, the use of different wording is not necessarily relevant. Running regression models on such high-dimensional data often violates many underlying model assumptions and is subject to overfitting problems. However, black-box models are much more difficult to interpret than a simpler statistical model. Deep-learning methods usually provide good prediction outcomes, but do not fully address subjective knowledge or thoroughly explain the causal relationships between variables. In free-text example, DNN bypasses the assumption of causative relationships between the terms used and provides outcome predictions based on combined variation in the usage of terms.

We also provided reference guidelines for training DNN models to classify patient-safety incidents and explored several potential model settings, namely number of layers, number of neurons, activation functions and regularisation. This research result does not show any obvious advantages on using various activation function, dropout control, number of hidden layers and neurons across different datasets. Future research should focus on investigating DNN development workflow framework that allows searching for model with optimal performance – it would be beneficial to achieve efficient classification task consistently across sets of incident reports. We found that processing time increased with the number of layers and the number of neurons in the DNN model within our experimental range. Nevertheless, the use of deep-learning training algorithms and deep-learning platform ensured that the most complex model took less than 2 min to run in a typical desktop computer environment.

As recommended by the Institute of Medicine, mandatory and voluntary reporting systems have been implemented in many developed countries to capture clinical near-misses and incidents. An informed culture, which can be fostered by incident reporting, plays an important role in ensuring a safe environment and situational awareness of patient-safety issues. All AIRS reports are currently reviewed and screened manually by designated staff in various subject offices of the HA.
Different readers may interpret textual descriptions differently. Manual review is often a subjective, expensive and time-consuming process that drains human resources. Similar to other reporting systems worldwide, the AIRS system accumulates a large number of free-text reports every year (more than 10,000 reports). Free-text reports carry rich information on incidents and near-misses (such as information on a patient’s diagnosis, reason for admission and the nature of the incident/near-miss) that awaits exploration. Due to the heavy workload of reviewers and the unstructured format of free text, the process of analysing records of similar incidents is slow. Therefore, an efficient and reliable method of managing and categorising the information in incident records is undoubtedly valuable.

WHO Minimal Information Model for Patient Safety (MIM PS) provides a list of the minimal meaningful data features of incident reporting. The suggested incident report structure heavily relies on narrative text to capture the background, progression of event, causes and contributing factors of the incident. Unlike AIRS with the indication of medication rights, many incident reporting systems from various countries do not register medication rights in a structured manner. The indication of the mismanagement related to the medication rights retains in the narrative text part that requires manual efforts to uncover. This increases the difficulties to retrieve similar cases for further analysis. One of our key contribution is that the proposed medication-rights detection system and the training procedure are transferable to understanding the medication-rights errors of other mandatory and voluntary incident reporting systems worldwide. Artificial intelligence-based health informatics applications, which enable efficient information retrieval, show promising potential for incident reporting understanding. Apart from high efficiency and accuracy classification of reports, some possible research applications, such as named entity recognition, are potentially useful to automatically tag entities in report. Our study demonstrates that advanced deep-learning methods for information retrieval and case classification may be of practical assistance.

Although the five ‘rights’ described in this study are recognised target for the safe administration of medication and are also referenced in guidelines for prescribing and dispensing drugs, the rights system has been criticised for offering very little practical guidance on ensuring drug safety. It is essential to reform the ways in which we understand and analyse near-misses and errors, using proper health information technology and a well-designed reporting format. Currently, some patient-safety systems offer tailor-made templates for specific incident categories and we observe a tendency to simplify incident report structures in some countries. The general reporting structure covers both structured items eliciting background and other relevant information on the incidents (such as patient details, hospital and ward information and indices of the severity of harm to the patient) and unstructured space in which users can describe the nature of the event and the immediate response. The learning objectives of these mandatory and voluntary reports have rarely been clearly articulated, despite widespread criticism of incident reporting as an additional burden for healthcare professionals. Synergising with continuous learning from incident reports, it will thus be advantageous to establish clear guidelines for reporting to clarify report structure, ensure the quality of reports, minimise the effort involved in reporting and aid in the assessment of reports. This research offers a scientific solution for analysis of the rich free-text data contained in incident reports, which may facilitate the development of an efficient approach to the exchange of patient-safety information that increases understanding of medical errors.

Deep-learning methods previously were not being used in incident report learning, and one of the contributions of this study is to examine the performance of deep-learning approaches in incident report learning to justify future use of advance related methods in understanding detailed categories and causes of incidents. Currently we employed NLP methods as well as bow for limited feature selection. Future researchers working on patient-safety informatics are advised to further explore...
the information retrieval and feature extraction power of deep learning directly from free-text data and the optimal way to configure the set of parameter arguments. It appears that different medication topics would require different model settings to achieve better prediction outcomes. It would be worthwhile to carry out extensive experiments to examine what kind of model setting scenarios would be more beneficial across different incident reports datasets. Experiments using different network algorithms and different datasets with grid searches may provide insights into the optimised modelling approach. Studies exploring the use of other advanced deep-learning methods, topic-based semantic analysis and incorporating clustering and association rules are potentially useful to improve model performance. Networks with convolutional and pooling layers may be useful for classification tasks in which strong local clues to class membership are needed. Traditional NLP methods may offer a useful means of identifying such clues in different parts of the input. For instance, the Ngram algorithm, which quantifies certain sequences of words and a single key phrase, can help to determine the topic of a document, and association rules may be useful for uncovering relationships between statements or wordings.

Using deep-learning methods for incident report learning would also contribute to classifying the type of medication errors. We envision forthcoming research to automate the identification of medication errors using advanced deep-learning approaches in the future. To achieve this, it would be necessary to overcome some significant hurdles in annotating training incident reports dataset (in a way for revealing the type of incidents and the associated important details of the incidents) based on appropriated named entities related to the medication errors. This probing study successfully showcased the significant potential of introducing deep-learning methods to incident report learning. We envision a promising future of widely using advanced artificial intelligence methods in patient-safety incident report learning.

Apart from the H2O platform, we acknowledge that there are other open-source platforms and libraries potentially suited to deep-learning model development, such as TensorFlow open-source software library. This study only demonstrated the possibility of using deep learning using H2O platform. Due to the limited number of data collected from the real-world AIRS system, we were unable to apply more advanced forms of deep-learning methods to this problem, such as BERT and RNN. It is noted that the DNN algorithms have been trained and tested on data arising from the AIRS system designed to capture the medication rights we set out to analyse. The scalability and generalisability of the built optimal models to other free-text incident reports have yet to be explored. Future research should focus on externally validating the trained models and the experimental workflow using global medication errors’ incident reports collected by other systems in the world.

**Conclusion**

This study developed a medication-rights detection system via DNN to automate medication-incident identification using free-text incident reports and provide reference guidelines for training DNN models to classify patient-safety incidents. The deep-learning method shows promise for the efficient exploration of textual reports of medication incidents. Our findings provide insights into training DNN models to classify patient-safety incidents, which may be transferable to other mandatory and voluntary reporting systems around the world.

**Author contributions**

The study was conceived by Z.S.-Y.W., H.Y.S., B.S.C.K., M.W.S.L. and D.T.F.S.; Z.S.-Y.W., H.Y.S. and D.T.F.S. contributed to the study design; research data or information was retrieved and interpreted by Z.S.-Y.W., B.S.C.K. and M.W.S.L.; and Z.S.-Y.W. led the writing of the article. All authors revised and refined the arguments, and they approved the article.
Declaration of conflicting interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Ethical approval
All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. The research was approved by the Joint Chinese University of Hong Kong–New Territories East Cluster (NTEC) Clinical Research Ethics Committee (CREC). The clinical research approval number awarded by the CREC was ‘2014.470’.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship and/or publication of this article: This research was supported by the Japan Society for the Promotion of Science KAKENHI (Grant No. 18H03336) and the Research Grants Council Theme-Based Research Scheme (Ref. T32-102/14N).

Informed consent
This is a retrospective incident reports analysis study. Incident reports database is contributed by clinical staff. Informed consent was obtained from all individual contributors included in the study (HIJ-19-0169.R1).

ORCID iD
Zoie Shui-Yee Wong https://orcid.org/0000-0003-4499-9779

Supplemental material
Supplemental material for this article is available online.

References
1. WHO. 10 facts on patient safety 2014, http://www.who.int/features/factfiles/patient_safety/en/
2. WHO. Patient safety: making health care safer, 2017, https://apps.who.int/iris/handle/10665/255507
3. Grissinger M. The five rights: a destination without a map. Pharm Therap 2010; 35(10): 542.
4. WHO. Medication without harm, 2017, https://apps.who.int/iris/handle/10665/255263
5. WHO. Minimal information model user guide. WHO, http://www.who.int/patientsafety/topics/reporting-learning/mim/en/ (2016, accessed 14 February 2018).
6. Kohn LT, Corrigan J and Donaldson MS. To err is human building a safer health system. Washington, DC: National Academy Press, 2000.
7. Franklin BD, Panesar SS, Vincent C, et al. Identifying systems failures in the pathway to a catastrophic event: an analysis of national incident report data relating to vinca alkaloids. BMJ Qual Saf 2014; 23: 765–772.
8. Ong MS, Magrabi F and Coiera E. Automated categorisation of clinical incident reports using statistical text classification. Qual Saf Health Care 2010; 19(6): e55.
9. Hospital Authority. Quality and risk management annual report 2012-2013. Report, Hospital Authority, Hong Kong, 2013.
10. Fujita K, Akiyama M, Park K, et al. Detecting effective categories of medical incident reports for patient safety management. In: 2012 Proceedings of PICMET ’12: technology management for emerging technologies, Vancouver, BC, Canada, 29 July–2 August 2012. New York: IEEE.
11. Magrabi F, Ong MS, Runciman W, et al. Using FDA reports to inform a classification for health information technology safety problems. J Am Med Inform Assoc 2012; 19(1): 45–53.
12. Chai KE, Anthony S, Coiera E, et al. Using statistical text classification to identify health information technology incidents. *J Am Med Inform Assoc* 2013; 20(5): 980–985.

13. Wong ZS. Statistical classification of drug incidents due to look-alike sound-alike mix-ups. *Health Informatics J* 2016; 22: 276–292.

14. Jauregi Unanue I, Zare Borzeshi E and Piccardi M. Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition. *J Biomed Inform* 2017; 76: 102–109.

15. Deng L and Yu D. Deep learning: methods and applications. *Found Trends Signal Process* 2013; 7(3–4): 197–387.

16. Delalleau O and Bengio Y. Shallow vs. deep sum-product networks. In: *Proceedings of the advances in neural information processing systems*, Granada, 12–15 December 2011.

17. Ravi D, Wong C, Deligianni F, et al. Deep learning for health informatics. *IEEE J Biomed Health Inform* 2017; 21(1): 4–21.

18. Devlin J, Chang MW, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding, 2018, https://arxiv.org/pdf/1810.04805.pdf

19. Hinton GE, Osindero S and Teh YW. A fast learning algorithm for deep belief nets. *Neural Comput* 2006; 18(7): 1527–1554.

20. LeCun Y, Bengio Y and Hinton G. Deep learning. *Nature* 2015; 521(7553): 436–444.

21. Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch. *J Mach Learn Res* 2011; 12: 2493–2537.

22. Bengio Y. Learning deep architectures for AI. *Found Trends Mach Learn* 2009; 2(1): 1–127.

23. Candel R, Parmar V, LeDell E, et al. Deep learning with H2O, 2017, http://h2o.ai/resources/

24. Iyyer M, Manjunatha V, Boyd-Graber J, et al. Deep unordered composition rivals syntactic methods for text classification. In: *Proceedings of the 53rd annual meeting of the association for computational linguistics*, Beijing, China, July 2015.

25. Goldberg Y. A primer on neural network models for natural language processing, 2015, http://arxiv.org/abs/1510.00726

26. Package ‘h2o’, 2015, https://cran.r-project.org/web/packages/h2o/h2o.pdf

27. Salton G. Developments in automatic text retrieval. *Science* 1991; 253(5023): 974–980.

28. Hinton GE, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors, 2012, https://arxiv.org/pdf/1207.0580.pdf

29. Mytton OT, Greaves FE, Stanton EA, et al. What should WHO be doing about patient safety and technology? *Qual Saf Health Care* 2010; 19(Suppl. 2): i44–i47.

30. WHO. Minimal information model for patient safety incident reporting and learning systems: user guide, 2016, http://www.who.int/iris/handle/10665/255642

31. Wong SY. *Text mining of medication incidents using topic models*. Washington, DC: AMIA, 2013.

32. Zhang S and Elhadad N. Unsupervised biomedical named entity recognition: experiments with clinical and biological texts. *J Biomed Inform* 2013; 46(6): 1088–1098.

33. Leistikow I, Mulder S, Vesseur J, et al. Learning from incidents in healthcare: the journey, not the arrival, matters. *BMJ Qual Saf* 2017; 26: 252–256.

34. Hinton GE and Salakhutdinov RR. Reducing the dimensionality of data with neural networks. *Science* 2006; 313(5786): 504–507.

35. Choi E, Schuetz A, Stewart WF, et al. Using recurrent neural network models for early detection of heart failure onset. *J Am Med Inform Assoc* 2017; 24(2): 361–370.

36. Shima Y and Wong ZSY. Classification scheme for incident reports of medication errors. *Stud Health Technol Inform* 2019; 265: 113–118.