Accepted Manuscript online: 2022-05-29 Article published online: 2022-07-13

Application of a Machine Learning–Based Decision Support Tool to Improve an Injury Surveillance System Workflow

Jesani Catchpoole[1,2,3] Gaurav Nanda[4] Kirsten Vallmuur[2,3] Goshad Nand[1] Mark Lehto[5]

[1] Queensland Injury Surveillance Unit, Royal Brisbane and Women's Hospital, Metro North Hospital and Health Service, Queensland, Australia
[2] Jamieson Trauma Institute, Royal Brisbane and Women's Hospital, Metro North Hospital and Health Service, Queensland, Australia
[3] Australian Centre for Health Services Innovation (AusHSI), School of Public Health and Social Work, Queensland University of Technology, Brisbane, Australia
[4] Purdue University, School of Engineering Technology, West Lafayette, Indiana, United States
[5] Purdue University, School of Industrial Engineering, West Lafayette, Indiana, United States

Address for correspondence Jesani Catchpoole, PhD, L13, Block 7, Queensland Injury Surveillance Unit, Royal Brisbane and Women's Hospital, HERSTON QLD 4029, Australia (e-mail: Jesani.catchpoole@health.qld.gov.au).

Appl Clin Inform 2022;13:700–710.

Abstract

Background Emergency department (ED)-based injury surveillance systems across many countries face resourcing challenges related to manual validation and coding of data.

Objective This study describes the evaluation of a machine learning (ML)-based decision support tool (DST) to assist injury surveillance departments in the validation, coding, and use of their data, comparing outcomes in coding time, and accuracy pre- and postimplementations.

Methods Manually coded injury surveillance data have been used to develop, train, and iteratively refine a ML-based classifier to enable semiautomated coding of injury narrative data. This paper describes a trial implementation of the ML-based DST in the Queensland Injury Surveillance Unit (QISU) workflow using a major pediatric hospital's ED data comparing outcomes in coding time and pre- and postimplementation accuracies.

Results The study found a 10% reduction in manual coding time after the DST was introduced. The Kappa statistics analysis in both DST-assisted and -unassisted data shows increase in accuracy across three data fields, that is, injury intent (85.4% unassisted vs. 94.5% assisted), external cause (88.8% unassisted vs. 91.8% assisted), and injury factor (89.3% unassisted vs. 92.9% assisted). The classifier was also used to produce a timely report monitoring injury patterns during the novel coronavirus disease 2019 (COVID-19) pandemic. Hence, it has the potential for near real-time surveillance of emerging hazards to inform public health responses.

Conclusion The integration of the DST into the injury surveillance workflow shows benefits as it facilitates timely reporting and acts as a DST in the manual coding process.

Keywords
- injury surveillance
- machine learning
- decision support

received
January 31, 2022
accepted after revision
May 26, 2022

DOI https://doi.org/10.1055/a-1863-7176.
ISSN 1869-0327.

Background and Significance

Emergency department (ED)-based injury surveillance systems are used across many countries to capture and monitor injury patterns and trends in the community, though most of these systems face resourcing challenges related to the validation, coding, and utilization of injury data. Previous papers have extensively described the ED-based injury data collection processes and challenges, as well as the iterative development of a machine learning (ML)-based injury classifier.[1–6] This study instead focuses on the applications of the ML-based injury classifier as a decision support tool (DST) to assist the injury coding workforce in validating, coding, and using these data in practice.

Context

The Queensland Injury Surveillance Unit (QISU) collects ED-based injury surveillance data from several participating hospitals across Queensland. QISU data are collected by triage nurses during the initial examination using an injury module triggered when the presentation is flagged as an injury via a Yes/No field. However, the injury data extracted by QISU are often incomplete due to the module being noncompulsory, leaving only the presenting problem narrative as a source of injury details. The missing fields are then completed by QISU coders based on the narrative data using an in-house validation and coding system called Injury Coding System (ICS). In addition to completing missing fields, the QISU coders also review and validate the injury codes assigned by triage nurses. Each record is validated and coded following the National Data Standards–Injury Surveillance (NDS-IS)[7] in the ICS.

QISU is moving toward broadening its injury selection criteria to include all ED cases with injury diagnoses (including cases not flagged affirmatively as injuries by triage nurses). While this will improve QISU data representativeness, using broader criteria will also increase the volume of data that needs validating and coding by QISU coders. With the current 1.9 full-time equivalent coding workforce, the turnaround time of coding is expected to be delayed and the provision of timely injury reporting is unfeasible.

Machine Learning–Based Injury Classifier Overview

The ML-Based Injury Classifier (referred to as "the classifier" herein) was developed in Microsoft Access (as it provides a good visual interface to the data along with database functionalities) for coding External Cause of Injury, Major Injury Factor (MIF), Mechanism of Injury, and Intent of injury based on the narrative of the injury and other fields. The development and refinement of the classifier used in this study have been described in previous publications.[2–4]

The classifier predicts these codes based on two ML models, that is (1) Logistic Regression (LR) and (2) Naive Bayes (NB),[8] which were trained on a large amount of manually coded QISU data from past years (2002 onward), as illustrated in ►Fig. 1.

The NB model can be described as: for a given narrative consisting of a vector of j words, $n = \{n_1, n_2, \ldots, n_j\}$, i possible set of codes (e.g., E-code/MIF) can be assigned represented by a second vector $E = \{E_1, E_2, \ldots, E_i\}$. Using conditional independence assumption, the probability of assigning a particular E-code can be calculated as:

$$P(E_i|n) = \prod_j \frac{P(n_j|E_i)P(E_i)}{P(n_j)}$$

where, $P(E_i|n)$ = probability of code category E_i given the set of n words in the narrative.

$P(n_j|E_i)$ = probability of word n_j given category E_i.
$P(E_i)$ = probability of category i.
$P(n_j)$ = probability of word n_j in the entire word list.
$P(n_j|E_i)$, $P(E_i)$, and $P(n_j)$ are estimated based on their frequency in a training set. $P(n_j|E_i)$ is usually smoothed to reduce the effects of noise by adding a small constant α to the number of times a particular word occurred in a category, as shown below:

$$P(n_j|E_i) = \frac{count(n_j|E_i) + \propto}{count(E_i) + \propto * N}$$

where, count $(n_j|E_i)$ = number of times word n_j occurs in category E_i, *count* (n_j) = number of times word n_j occurs,

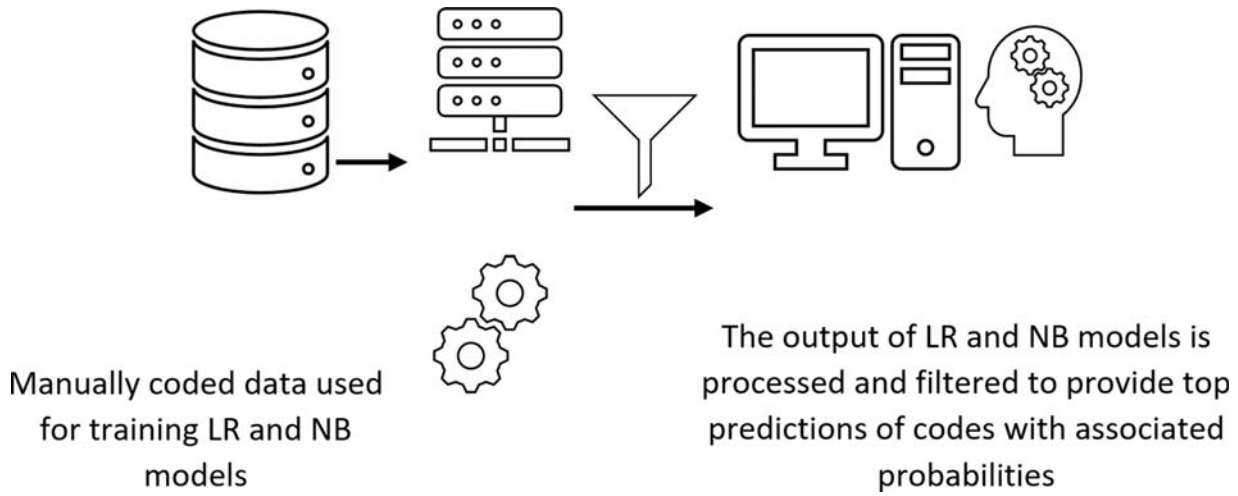

Fig. 1 Schematic diagram of the machine learning-based injury classifier. LR, Logistic Regression; NB, Naive Bayes.

count (E_i) = number of times category E_i occurs, and N = number of training categories.

In the LR model, for a feature vector of words represented by x, prediction represented as y, injury code category represented by k, the total number of categories being K, and θ_k representing the weight vector associated with each category, the probability of a category k being true given x is calculated as:

$$p(y = k|x) = \frac{\exp(\theta_k^T x)}{\sum_{i=1}^{K} \exp(\theta_k^T x)}$$

The optimal weight vectors are calculated iteratively by maximizing the log-likelihood function on the training data:

$$\max_{\theta} \sum_{i=1}^{n} \log p(y_i|x_i, \theta)$$

where n = number of training cases.

These ML models were selected because they output the predicted injury code along with the associated probability scores as shown above, which provides the users with a useful assessment of the model's confidence in a decision support scenario. The classifier offers various options to users, such as providing predictions based on one of the two ML models (LR or NB) or combining the prediction outputs of the two models. In the case any individual model is selected, the prediction output of only that model is considered. If the combined mode of the classifier is selected, then the average of prediction probabilities outputted by the LR and NB models is calculated for each category, and the top-1 or top-3 category with the highest average prediction probability is presented as output to the user.

Related Work

Artificial intelligence (AI)/ML-based health care decision support systems (DSS) have been found helpful for various purposes, such as effectively managing everyday operations in the hospital and providing valuable insights from the electronic health records and literature to health care professionals.[9–15] For example, clinical DSS paired with computerized physician order entry have been found to (1) substantially reduce medication errors,[10] (2) help medical professionals in retrieving highly relevant medical literature to help in formulating diagnoses and applicable treatments,[14] and (3) assist clinicians in classifying anatomical location of catheter location by reading radiology reports.[13] Studies have also been conducted to explore the benefit of ML tools in detecting the onset of sepsis in hospitalized patients[15] and detecting patients with a higher risk of readmissions.[12] Health care DSS can also help hospitals and public health agencies to analyze and monitor trends of health care quality indicators,[9] and predict intensive care unit (ICU) admission and in-hospital death of trauma patients[11] for better prioritization and utilization of resources. Outside the health care system, DSS has also been found useful to automatically classify injury data, particularly occupational injury data.[16–19]

While AI/ML-based DSS offers several benefits, their adoption in health care organizations is sometimes slow due to technology adoption challenges. A previous study recommended that for successful adoption and effectiveness of a DSS in hospitals, it is important to have top management support, active involvement of clinical departments, and robust hospital-wide information infrastructure to collect and process good quality data from multiple sources, among other factors.[9] For successful adoption of the AI/ML-based clinical DSS by staff members, gaining their trust in the system is especially important. Explanation of the recommendations made by clinical DSS with proper reasoning is one of the critical factors for developing trust in the DSS by health care professionals who use the system.[20] Bayesian DSS have been found to be effective in various injury surveillance, health care, and other applications, including (1) learning of motor vehicle accident categories,[21] (2) coding of occupational injury cases,[18] (3) automatic indexing of documents,[22] (4) providing interactive decision support related to print quality to customers,[23] and (5) classification and identification of customer complaints.[24] Similar to the NB model, the LR model also outputs the likelihood of correctness of prediction, and thus, both these models were used to develop the classifier for QISU.

Methods

Machine Learning–Based Decision Support Tool Implementation

The QISU workflow follows a sequence of data processing tasks as described in ►Fig. 2. The injury data from collecting EDs is extracted monthly from a web portal (as text files) and imported into the ICS. The ICS consists of a database that can be interrogated with an interactive form that allows QISU coders to manually validate and code the injury cases one record at a time. After the coding and validation process, the cleaned dataset is exported into the QISU central database. From this database, data can be easily accessed, queried, and analyzed for reporting purposes.

The classifier is being integrated into the QISU's existing workflow in three stages (►Fig. 2). Stage 1 involves setting up a ML-based DST (referred to herein as "the DST") to assist QISU coding workforce in their validation and coding practices. Stage 2 involves the development of a separate data storage for machine-classified data to allow the machine-classified raw injury data to be analyzed and used for more timely injury reporting. Stage 3 includes regular benchmarking of the machine classifier data against the validated data to refine the machine classifier performance iteratively. This paper will focus on evaluating the impact of the ML-based DST trial implementation in stage 1.

The development of the DST in stage 1 involves processing the raw injury data into the classifier prior to the ICS data importation. It is to be noted that no other natural language preprocessing technique, such as stemming, lemmatization, or stop word removal, was applied to maintain the original format and content of the raw data. The classifier automatically reads the delimited text files extracted from the ED information system and assigns predicted codes for four injury variables: external cause, intent, mechanism of injury,

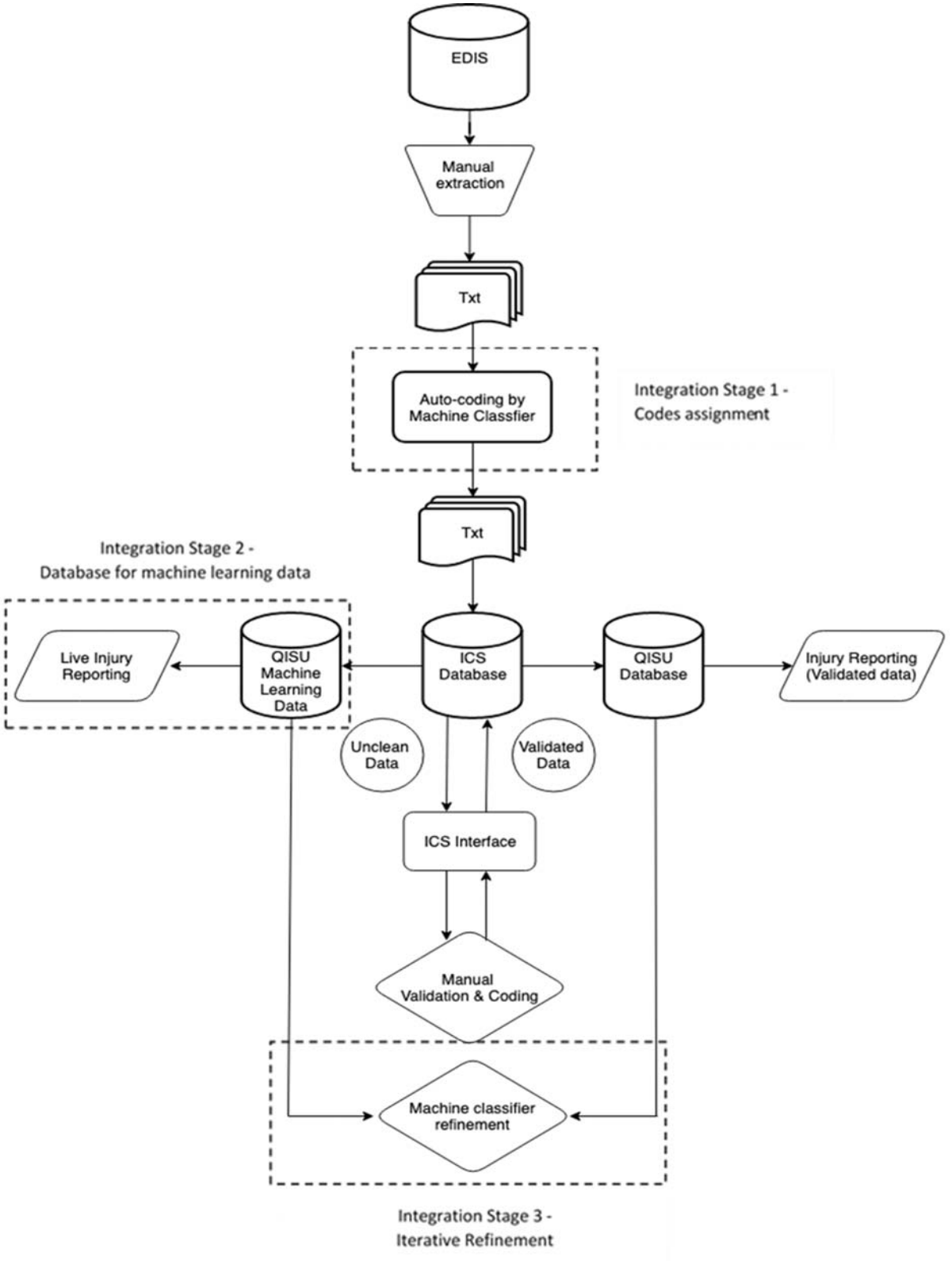

Fig. 2 QISU data workflow. EDIS, Emergency Department Information System; ICS, Injury Coding System; QISU, Queensland Injury Surveillance Unit.

and major injury factor. The classifier was designed only to assign predicted codes where the triage nurses have not completed injury fields. This was done to avoid changes in the existing structure and format of the raw data files. The machine-classified data with predicted codes produced by the classifier is then saved in the same delimited text file format. This will allow the machine classified data to be imported into the QISU cleaning system following the existing workflow for manual validation and coding. The DST consists of predicted codes from the classifier acting as

decision support prompts to QISU coders providing suggested coding for the four injury variables.

Impact Measures

To study the impact of the DST on the coding workforce, a trial implementation was conducted using data from QISU's pediatric hospital collected in March 2020. Data from the month prior (February 2020) was used as a control dataset to compare the outcome pre- and post-DST implementation. The trial dataset was processed through the classifier as described in stage 1 to produce machine-classified data on the four variables as follows: (1) external cause, (2) MIF, (3) mechanism of injury, and (4) intent of injury. The predicted codes were then imported into the ICS as decision support prompts that were then validated by the QISU coders. The validation and coding process remain the same pre- and postimplementation of the DST.

The impact on data accuracy was measured by comparing the number of accurate codes in both February (control) and March (trial) datasets. Two expert coders were assigned to code both datasets independently to generate the gold-standard (GS) codes. The expert coders were blinded to the QISU coders' coding; however, they were not blinded to the machine classifier coding. Disagreements between both expert coders were resolved by discussion to reach an agreement. The codes assigned by QISU coders in both datasets were compared with the agreed GS codes. Discordant cases were regarded as inaccurate coding, and concordant cases were regarded as accurate coding. Concordances between the QISU Coders (QC) and Machine Classifier (MC) were also examined.

The impact on coding efficiency was measured based on the comparison of median coding time in both February (control) and March (trial) datasets. Each time, a record is saved in ICS during the manual coding and validation process, the system automatically assigns a timestamp. These data were used to calculate coding time by sorting the timestamp in chronological order and calculating each record's coding time based on the time difference between records' timestamps. The first records of the day and records after breaks were treated as outliers, and therefore excluded from coding time analysis.

Statistical Analysis

Data management was performed in Microsoft Access, and all statistical analyses were performed using IBM SPSS, version 23. Coding concordance data were analyzed using Chi-square tests. Cohen's Kappa statistics and sensitivity were calculated to test the agreement of the injury coding between the GS and QISU coding. Due to the data's ordinal and positively skewed nature, coding time was analyzed using a nonparametric Mann–Whitney *U*-test and represented using median and interquartile range (IQR).

Results

In total, 3,174 records in February ($n = 1{,}668$) and March ($n = 1{,}506$) were extracted from the ICS. March data were preprocessed through the batch classifier, generating the predicted codes that were used in the DST. February data were included as a control dataset in this study to measure the changes in coding efficiency and data accuracy after implementing the DST in QISU's workflow.

The classifier was designed to read ED data files and only to code records that triage nurses have not coded. In total, the classifier assigned intent codes in 599 injury records, external cause codes in 820 records, and MIF codes in 1,476 records. As specified in the background, the ED data often have empty fields in the injury module as triage nurses are allowed to skip part of or all the injury fields. In the pediatric hospital data used for this study, external cause (46%) and intent of injury (60%) fields appeared to have much higher completion rates compared with the major injury factor field (2%).

Impact on Data Accuracy

The implementation of the DST is expected to improve data accuracy, as it prompts QISU coders to crosscheck code assignments. The findings from the Kappa analysis, comparing the GS and QISU coders coding in both control and trial datasets, show increases in coding accuracy across all three data fields: injury intent (85.4% unassisted vs. 94.5% assisted), external cause (88.8% unassisted vs. 91.8% assisted), and major injury factor (89.3% unassisted vs. 92.9% assisted) as shown in ►Table 1. Although the percentages of accurate coding of injury intent appear to be similar in both control and trial datasets, the kappa value is 5.6% higher in the trial dataset than in the control dataset.

The Intent of Injury Coding Accuracy

Overall, 610 records with machine-classified intent of injury codes were included in the analysis. The intent of injury field includes 11 categories separating unintentional injuries from intentional self-harm and assault injuries. However, the majority of the records are unintentional injuries leaving very small numbers of records in other categories. Therefore, due to the small numbers in the intentional categories in this study, intent categories were grouped into two broad categories, such as (1) unintentional injuries, and (2) intentional injuries.[7]

The majority of the records coded by QISU coders (603 records, 98.8%) were validated by the GS. Out of all these, 590 records were from the concordant group between QISU coders and the classifier, meaning that 13 records were accurately corrected during the manual coding process. The classifier accurately coded one out of seven records that were inaccurately coded by QISU coders. Of all the machine classified intent codes, 591 (96.7%) were in concordance with the GS.

When comparing the 610 machine-classified records with the control dataset, there was almost no difference in the proportion of GS concordance between the two datasets. However, Kappa statistics shows a higher agreement (+5.6%) between the GS and manual coding in the trial dataset (►Table 2). This suggests that the classifier increases the accuracy of the intent-of-injury coding.

Table 1 Coding accuracy unassisted and assisted by the DST

Data fields	Control data (unassisted by the DST) *n* (%)/Sensitivity (95% CI)	Trial data (assisted by the DST)[a] *n* (%)/Sensitivity (95% CI)
Intent of injury (grouped)	***n* = 1,668**	***n* = 610**
Accurate coding	1,653 (99.1)	605 (99.2)
Inaccurate coding	20 (1.2)	7 (0.8)
UAvg sensitivity	0.941 (0.94–0.95)	0.960 (0.95–0.96)
Kappa value	0.889 (0.83–0.94)	0.945 (0.90–0.99)
p-Value	<0.001	<0.001
External cause	***n* = 1,668**	***n* = 820**
Accurate coding	1516 (90.9)	762 (92.9)
Inaccurate coding	152 (9.1)	58 (7.1)
UAvg sensitivity	0.939 (0.898–0.980)	0.942 (0.89–0.99)
Kappa value	0.888 (0.87–0.9)	0.918 (0.90–0.94)
p-Value	<0.001	<0.001
Major injury factor	***n* = 1,668**	***n* = 1,476**
Accurate coding	1,502 (90)	1,378 (93.4)
Inaccurate coding	166 (10)	98 (6.6)
UAvg sensitivity	0.886 (0.850–0.923)	0.939 (0.909–0.971)
Kappa value	0.893 (0.88–0.91)	0.929 (0.91–0.94)
p-Value	<0.001	<0.001

Abbreviations: CI, confidence interval; DST, decision support tool; UAvg, unweighted average
[a]Number of records from the trial dataset included in the analysis varies depending on the number of records classified by the classifier.

Looking at the specific categories in the intent of injury field (ungrouped), in both trial and control datasets, concordance between the GS and QISU coders are almost equal (above 99%) in unintentional injuries coding with the trial dataset showed slightly higher (+0.3%) concordance than the control dataset. In contrast to the unintentional and self-harm injuries, both datasets showed lower GS concordance in assault categories (trial, 77.8% and control, 80%) and other and unspecified intent (trial, 56% and control, 68%) coding. This may be explained by the nature of the data used in the study being from a pediatric hospital. Pediatric injury cases in these two categories are often ambiguous at the point of patient triage, and the documentation regarding intent can be interpreted differently by human coders.[25]

External Cause Coding Accuracy

A total of 820 records with machine-classified external cause codes were included in the analysis. The external cause field includes 30 categories which are grouped into 14 broad categories.[7] Approximately 93% of the external cause coded by QISU coders (762 records) were validated by the GS. Out of all 762 records, 686 records (90%) were from the concordant group between QISU coders and the classifier, and the remaining 76 records were accurately corrected during the manual coding process. The classifier accurately coded 7 out

Table 2 Intent of injury coding concordance

Intent of injury categories (grouped)	Number of concordant cases (% of total)			
	GS and QC (control) *n* = 1,668 *n* (%)	GS and QC (trial) *n* = 610 *n* (%)	GS and MC (trial) *n* = 610 *n* (%)	QC and MC (trial) *n* = 610 *n* (%)
Unintentional	1,590 (99.5)	558 (99.8)	558 (99.8)	557 (99.6)
Intentional self-harm	28 (100)	33 (100)	26 (78.8)	26 (78.8)
Assault	12 (80)	7 (77.8)	6 (66. 7)	6 (66. 7)
Other or unspecified intent	18 (66. 7)	5 (55. 6)	1 (11.1)	4 (44.4)
Overall	1,648 (98.8)	603 (98.8)	591 (96.9)	593 (97.2)

Abbreviations: GS, gold standard coding; MC, machine classifier coding; QC, Queensland Injury Surveillance Unit coders coding.

Table 3 External cause coding concordance

External cause categories (grouped)	Number of concordant cases (% of total)			
	GS and QC (control) $n = 1668$ *n* (%)	GS and QC (trial) $n = 820$ *n* (%)	GS and MC (trial) $n = 820$ *n* (%)	QC and MC (trial) $n = 820$ *n* (%)
Transport	79 (95.2)	37 (100)	33 (89.2)	32 (86.5)
Fall: high or low	625 (93.0)	213 (99.1)	204 (94.8)	203 (94.4)
Threat to breathing	11 (100)	8 (61.5)	7 (53.8)	11 (84.6)
Fire, flames, smoke	1 (100)	2 (100)	(0)	(0)
Exposure to hot object	29 (100)	9 (100)	9 (100)	9 (100)
Poisoning	40 (100)	29 (100)	27 (93.1)	27 (93.1)
Cutting, piercing object	1 (100)	28 (96.5)	24 (82.8)	29 (100)
Animal related	41 (89.1)	23 (95.8)	20 (83.3)	21 (87.5)
Machinery	35 (89.7)	3 (60)	2 (40)	2 (40)
Electricity	3 (50)	2 (100)	1 (50)	1 (50)
Struck by/collision with object	355 (91.0)	211 (91.7)	187 (81.3)	203 (88.3)
Other/unspecified cause	294 (84.5)	197 (87.6)	170 (75.6)	194 (86.2)
Overall	1516 (90.9)	762 (92.9)	684 (83.4)	732 (89.3)

Abbreviations: GS, gold standard coding; MC, machine classifier coding; QC, Queensland Injury Surveillance Unit coders coding.

of 58 records that were inaccurately coded by QISU coders. Of all the machine classified external cause codes, 693 (84.5%) were in concordance with the GS.

When comparing the percentage of GS concordance between the 820 machine-classified records and the control dataset (►Table 3), there was an increase of 2% in external cause coding accuracy from 90.9% (control) to 92.9% (trial). An even higher level of accuracy is shown by Kappa statistics (88.8% unassisted vs. 91.8% assisted).

Within the specific external cause categories, GS concordance is equal at 100% in both trial and control datasets in the fire, flames, smoke, exposure to hot subject, and poisoning categories. The GS concordance is higher in the trial dataset than in the control dataset in all the frequent categories, such as transport (+4.8%), fall (+6.1%), struck by or collision with object (+0.7%), and the other, and unspecified cause (+3.1%). Conversely, the concordance in the trial dataset is lower than the control dataset in the rare categories such as threat to breathing (−38.5%), cutting, piercing object (−3.5%), and machinery (−29.7%). In categories like animal-related and electricity, where the characteristic of the injury is relatively prominent, the trial dataset has higher accuracy than the control dataset.

Major Injury Factor Coding Accuracy

A total of 1,476 records with machine-classified major injury factor codes were included in the analysis (►Table 4). The major injury factor field includes 138 object categories which are grouped into 13 broad categories.[7] Approximately 93% of the MIF coded by QISU coders (1,378 records) were validated by the GS. Out of all 1,378 records, 87.7% (1,209 records) were already coded accurately by the classifier, and 169 records were corrected during the manual coding process. Within the 98 MIF records that were inaccurately coded by QISU coders, 30 records were accurately coded by the classifier. Of all the machine classified injury factor codes, 1,239 (83.9%) were consistent with the GS.

When the GS concordance in both trial and control datasets were compared, there was an increase of 3.4% in MIF coding accuracy from 90% (control) to 93.4% (trial). Similar to injury intent and external cause coding, Kappa statistics shows an even higher level of accuracy in the trial dataset (89.3% unassisted vs. 92.9% assisted)

Within the specific major injury factor categories, the GS concordance range between 71% (miscellaneous) and 99% (food, drink, and personal use item). The GS concordance is higher in the trial dataset than in the control dataset in almost all the broad categories including infant or child's product (+5.6%); furnishing (+0.4%); appliance (+8.1%); utensil; container or rubbish (+11%); sporting equipment (+3.3%); natural object or animal (+3.9%); food, drink, and personal item (+9.5%); chemical substance (+8.3); and other material (+4.7%). On the other hand, the concordance in the trial dataset is lower than the control dataset in several categories such as transport (−2.5%), tool (−7.6%), structure, or fitting (−2.2%), and miscellaneous (−6.8%).

Impact on Coding Efficiency

The impact on coding efficiency was measured based on the time difference between records' timestamps. The median coding time in February and March data was compared with examine the DST impact on coding time. In total, 2,960 records were included in the coding time analysis, and 214 records were excluded as outliers.

Overall, approximately 72% of all the records were coded under 1 minute, 22% within 3 minutes, and only 6% within

Table 4 Major injury factor coding concordance

Major injury factor categories (grouped)	Number of concordant cases (% of total)			
	GS and QC (control) $n = 1{,}668$ *n* (%)	GS and QC (trial) $n = 1{,}476$ *n* (%)	GS and MC (trial) $n = 1{,}476$ *n* (%)	QC and MC (trial) $n = 1{,}476$ *n* (%)
Infants or child's product	108 (91.5)	99 (97.1)	90 (88.2)	92 (90.2)
Furnishing	43 (89.6)	45 (90)	32 (64)	33 (66)
Appliance	176 (89.3)	149 (97.4)	130 (85.0)	131 (85.6)
Utensil, container or rubbish	59 (83.1)	48 (94.1)	35 (68.6)	35 (68.6)
Transport (including mobile machinery)	396 (92.7)	322 (90.2)	325 (91.0)	323 (90.5)
Sporting equipment	165 (94.8)	152 (98.1)	138 (89.0)	139 (89.7)
Tool	10 (90.9)	15 (83.3)	13 (72.2)	14 (77.8)
Natural object or animal	26 (92.9)	30 (96.8)	24 (77.4)	24 (77.4)
Food, drink, personal use item	106 (89.8)	138 (99.3)	134 (96.4)	134 (96.4)
Chemical substance	121 (89.0)	107 (97.3)	95 (86.4)	94 (85.4)
Structure or fitting	8 (88.9)	13 (86.7)	6 (40)	7 (46.7)
Material	190 (90.5)	199 (95.2)	161 (77.0)	165 (78.9)
Miscellaneous	94 (77.7)	61 (70.9)	56 (65.1)	59 (68.6)
Overall	1,502 (90)	1,378 (93.4)	1,239 (83.9)	1,250 (84.7)

Abbreviations: GS, gold standard coding; MC, machine classifier coding; QC, Queensland Injury Surveillance Unit coders coding.

7 minutes. The median coding time for all records is 39 (IQR = 26–65 seconds; ►**Table 5**). The most quickly coded injury records (coded within a few seconds) usually contain the shortest injury narratives with little information about the injury circumstances. Consequently, these records are often assigned to unspecified categories. In contrast, injury records that took the longest to code (a few minutes) were (1) rare injury cases or (2) cases with conflicting information. These injury cases often require further investigation and discussion with other coding staff to make decisions in codes assignment.

Coding times in both trial and control data were analyzed to establish whether the DST improved coding efficiency. The Mann–Whitney *U*-test shows a statistically significant difference in coding time between the two datasets with a 10% reduction in manual coding time after the DST was introduced. The median coding time in March (DST assisted) was 37 seconds (IQR = 25–63) compared with 41 seconds (IQR = 27–68) in February (without DST; $U = 1017058$, $\underline{z} = -3.32$, $p < 0.001$, $r = 0.061$). Although a 4-second reduction in median seems insignificant, cumulatively, it adds up to a considerable increase in the number of records coded. For example, if a coder spends 5 hours coding on a daily basis, completing 400 records unassisted, the coder can complete approximately 80 more records daily with DST. This amounts to an increase of ca. Overall, 1,800 records from one coder on a monthly average.

In addition to the three DST-assisted data fields, other injury fields, such as activity when injured, location of the injury, nature of injury, and body region injured, are also validated and coded by QISU coders during the manual coding process. Although the DST helped speed up the overall coding process, manual coding of these other injury fields is still unassisted.

The length of time spent on coding each injury field may vary depending on the complexity and the number of categories to select from. For example, the injury intent field has 11 categories which is more straightforward and quicker

Table 5 Coding time per record assisted and unassisted by the DST[a]

Coding time per record	Control data (unassisted by the DST) *n* (%)	Trial data (assisted by the DST) *n* (%)	Total *n* (%)
<1 minute	1,071 (70.1)	1,055 (73.6)	2,126 (71.8)
1–3 minutes	359 (23.5)	301 (21)	660 (22.3)
3–7 minutes	97 (6.4)	77 (5.4)	174 (5.9)
Total	1,527 (100)	1,433 (100)	2,960 (100)
Median (IQR)	41 (27–68)	37 (25–63)	39 (26–65)

Abbreviations: DST, decision support tool; IQR, interquartile range.
[a]Mann–Whitney *U*-test = 1,017,058, $z = -3.32$, $p < 0.001$, $r = 0.061$.

to code than the major injury factor field which has 138 categories. Unfortunately, the ICS does not record the length of time spent on coding individual injury fields. Therefore, analysis to compare coding time in each injury field is not feasible. Another factor that may influence coding time is the commonness of the injury case. The QISU coders often memorize frequent codes and are therefore quicker to code than rare codes.

Discussion

Implication for Injury Surveillance

ED-based injury surveillance systems are in use worldwide for monitoring and responding to injury patterns and trends in the community, yet almost universally, systems, such as these are challenged by resourcing constraints and competing demands for health care budgets. Furthermore, producing timely contemporary data to monitor emerging trends is challenged in systems relying on manual coding of such voluminous data, such as the common ED presentation of injury as a cause. This paper described the evaluation of an ML-based classifier and DST which was developed using Queensland's ED-based injury surveillance data and implemented into the workflow of an injury surveillance coding department.

Previous studies suggest that it could be expected that coders would still be adjusting to a new system and lack familiarity with and trust in the new tool.[26–28] Even with this expectation taken into account, in the first month of implementation, the DST was found to increase the efficiency and accuracy of coding in the department. While results are promising so far, further evaluation of the DST's utility for a larger sample of injury surveillance data, including adult injury presentations, is required to quantify the impact in terms of efficiency and accuracy gains.

One of the unexpected benefits of the classifier was that during the implementation period, there was an urgent need for rapid accumulation and analysis of contemporary injury data to support government enquiries regarding the impact of the novel coronavirus disease 2019 (COVID-19) community restrictions on injury frequency and severity. In the routine injury surveillance system, responding to such requests would not be possible due to the coding backlog whereby data are 6 to 12 months' old by the time they have been coded and validated for use. In this context, overall, approximately 150,000 ED records were coded by the classifier in regard to intent and external cause in under an hour (comparing 2019 with 2020 figures), a task that would have taken a coder (averaging one record per minute or 400 records per day) almost a year of solid coding to complete. The raw injury data coded by the classifier could be used to produce a timely report for our government stakeholders. This data proved invaluable to these government departments tasked with responding to safety concerns during this time, providing them with evidence to target their responses to specific areas of concern, including certain transportation devices (motorcycles and bicycles), intentional self-harm, and assault. Hence, continued development and refinement of ML-based classifiers, such as that described in this study, have significant potential for almost real-time biosurveillance of emerging hazards to enhance evidence for public health responses.

Limitations

The study has some limitations. First, the data used for the study was only collected for a short period (2 months). Another evaluation of data accuracy and coding efficiency should be conducted after a longer period of postimplementation. Data from previous years prior to DST implementation should also be used as a control dataset to allow comparison and evaluation of longitudinal effects of the DST. Another limitation of the study is that although the results show some positive impacts of the DST implementation, other secular trends may have also contributed to these outcomes. For example, possible learning by coders over time might influence their ability to code faster and more accurately. A quasiexperimental, such as interrupted time series, should be considered when evaluating a bigger dataset to address the impacts of secular trends. Despite these limitations, in the short term, the integration of the DST into the injury surveillance workflow has shown benefits, as it facilitates timely reporting and acts as a DST in the manual coding process.

Clinical Relevance Statement

Injury surveillance data provide valuable information on injury patterns and trends in the community to injury prevention workers, emergency and trauma clinicians, and government agencies. Integrating the machine learning–based classifier in the injury surveillance system's workflow will increase the completeness of injury data and expedite injury coding for near real-time injury surveillance without putting extra pressure on clinical staff to assign injury coding.

Multiple Choice Questions

1. Which of the following groups was the decision support tool designed for?
   a. Patients
   b. Triage nurses
   c. Injury coders
   d. Doctors

   Correct Answer: The correct answer is option c. The decision support tool was designed to assist injury coders in assigning injury intent, external cause and major injury factor codes.

2. What is the main injury data field used by the machine classifier to predict injury codes?
   a. Injury narrative
   b. External cause
   c. Major injury factor
   d. Age

Correct Answer: The correct answer is option a. The machine classifier was designed to read injury narrative and predict injury intent, external cause, and major injury factor codes.

3. What external cause categories showed higher accuracy when coders were assisted by the machine learning–based decision support tool?
   a. Frequent categories
   b. Rare categories
   c. Machinery categories
   d. Object categories

   Correct Answer: The correct answer is option a. Accuracy of coding was higher in the trial dataset than in the control dataset in all the frequent categories (i.e., transport, fall, struck by or collision with object and the other, and unspecified cause). On the other hand, the coding accuracy in the trial dataset was lower than the control dataset in the rare categories such as threat-to-breathing, cutting, piercing object, and machinery.

4. What injury intent category showed the highest accuracy improvement when coders were assisted by the machine learning–based decision support tool?
   a. Self-harm
   b. Unintentional
   c. Assault
   d. Unspecified

   Correct Answer: The correct answer is option b. The accuracy of unintentional injuries coding was already high in the unassisted environment. However, the coding accuracy increased slightly by 0.3% in the trial dataset.

Protection of Human and Animal Subjects

Human and/or animal subjects were not included in the project. In addition, the study analyzed nonidentifiable data; therefore, consent from patients was not required.

Funding

This work was supported by an Australian Research Council Discovery Grant (grant no.: DP170103136).

Conflict of Interest

J.C. and Go.N. are employees of the Queensland Injury Surveillance Unit. Other authors report no conflict of interest in the research.

References

1 Chen L, Vallmuur K, Nayak R. Injury narrative text classification using factorization model. BMC Med Inform Decis Mak 2015;15 (Suppl 1):S5–S5
2 Nanda G, Vallmuur K, Lehto M. Improving autocoding performance of rare categories in injury classification: is more training data or filtering the solution? Accid Anal Prev 2018;110:115–127
3 Nanda G, Vallmuur K, Lehto M. Semi-automated text mining strategies for identifying rare causes of injuries from emergency room triage data. IISE Trans Healthc Syst Eng 2019;9(02):157–171
4 Nanda G, Vallmuur K, Lehto M. Intelligent human-machine approaches for assigning groups of injury codes to accident narratives. Saf Sci 2020;125:104585
5 Vallmuur K. Machine learning approaches to analysing textual injury surveillance data: A systematic review. Accid Anal Prev 2015;79:41–49
6 Vallmuur K, Marucci-Wellman HR, Taylor JA, Lehto M, Corns HL, Smith GS. Harnessing information from injury narratives in the 'big data' era: understanding and applying machine learning for injury surveillance. Inj Prev 2016;22(Suppl 1):i34–i42
7 National Injury Surveillance Unit. National Data Standards for Injury Surveillance Ver 2.1. Available at: https://nla.gov.au/nla.cat-vn730536
8 Mitchell TM. Machine Learning. 1st ed. Ithaca, NY: McGraw-Hill, Inc.; 1997
9 Chae YM, Kim HS, Tark KC, Park HJ, Ho SH. Analysis of healthcare quality indicator using data mining and decision support system. Expert Syst Appl 2003;24(02):167–172
10 Kaushal R, Shojania KG, Bates DW. Effects of computerized physician order entry and clinical decision support systems on medication safety: A systematic review. Arch Intern Med 2003; 163(12):1409–1416
11 Kong G, Xu D, Yang J, Wang T, Jiang B. Evidential reasoning rule-based decision support system for predicting ICU admission and in-hospital death of trauma. IEEE Trans Syst Man Cybern Syst 2020;51(11):7131–7142
12 Romero-Brufau S, Wyatt KD, Boyum P, Mickelson M, Moore M, Cognetta-Rieke C. Implementation of artificial intelligence-based clinical decision support to reduce hospital readmissions at a regional hospital. Appl Clin Inform 2020;11(04):570–577
13 Shah M, Shu D, Prasath VBS, Ni Y, Schapiro AH, Dufendach KR. Machine learning for detection of correct peripherally inserted central catheter tip position from radiology reports in infants. Appl Clin Inform 2021;12(04):856–863
14 Soldaini L, Cohan A, Yates A, Goharian N, Frieder O. Retrieving Medical Literature for Clinical Decision Support. Paper presented at: European Conference on Information Retrieval, Advances in Information Retrieval; 2015, 2015;538–549
15 Teng AK, Wilcox AB. A review of predictive analytics solutions for sepsis patients. Appl Clin Inform 2020;11(03):387–398
16 Goh YM, Ubeynarayana CU. Construction accident narrative classification: An evaluation of text mining techniques. Accid Anal Prev 2017;108:122–130
17 Goldberg DM. Characterizing accident narratives with word embeddings: Improving accuracy, richness, and generalizability. J Safety Res 2022;80:441–455
18 Nanda G, Grattan KM, Chu MT, Davis LK, Lehto MR. Bayesian decision support for coding occupational injury data. J Safety Res 2016;57:71–82
19 Zhong B, Pan X, Love PED, Ding L, Fang W. Deep learning and network analysis: Classifying and visualizing accident narratives in construction. Autom Construct 2020;113:103089
20 Bussone A, Stumpf S, O'Sullivan D. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. 2015: 160–169. Doi: 10.1109/ICHI.2015.26
21 Lehto MR, Sorock GS. Machine learning of motor vehicle accident categories from narrative data. Methods Inf Med 1996;35(4-5):309–316
22 Zhu W, Lehto MR. Decision support for indexing and retrieval of information in hypertext systems. Int J Hum Comput Interact 1999;11(04):349–371
23 Leman S, Lehto MR. Interactive decision support system to predict print quality. Ergonomics 2003;46(1–3):52–67
24 Choe P, Lehto MR, Shin GC, Choi KY. Semiautomated identification and classification of customer complaints. Hum Factors Ergon Manuf Serv Ind 2013;23(02):149–162

25 McKenzie K, Scott DA, Waller GS, Campbell M. Reliability of routinely collected hospital data for child maltreatment surveillance. BMC Public Health 2011;11(01):8–8
26 Liberati EG, Ruggiero F, Galuppo L, et al. What hinders the uptake of computerized decision support systems in hospitals? A qualitative study and framework for implementation. Implement Sci 2017;12(01):113–113
27 May CR, Mair F, Finch T, et al. Development of a theory of implementation and integration: normalization process theory. Implement Sci 2009;4(01):29–29
28 Mishuris RG, Palmisano J, McCullagh L, et al. Using normalisation process theory to understand workflow implications of decision support implementation across diverse primary care settings. BMJ Health Care Inform 2019;26(01):e100088