Accident Analysis and Prevention 79 (2015) 41–49

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Machine learning approaches to analysing textual injury surveillance data: A systematic review

Kirsten Vallmuur *

Centre for Accident Research and Road Safety – Queensland, School of Psychology and Counselling, Faculty of Health, Queensland University of Technology, Kelvin Grove 4059, Brisbane, Queensland, Australia

ARTICLE INFO

Article history:
Received 14 May 2014
Received in revised form 1 December 2014
Accepted 12 March 2015
Available online 19 March 2015

Keywords:
Text data
Injury surveillance
Injury epidemiology
Text mining
Machine learning

ABSTRACT

Objective: To synthesise recent research on the use of machine learning approaches to mining textual injury surveillance data.
Design: Systematic review.
Data sources: The electronic databases which were searched included PubMed, Cinahl, Medline, Google Scholar, and Proquest. The bibliography of all relevant articles was examined and associated articles were identified using a snowballing technique.
Selection criteria: For inclusion, articles were required to meet the following criteria: (a) used a health-related database, (b) focused on injury-related cases, AND used machine learning approaches to analyse textual data.
Methods: The papers identified through the search were screened resulting in 16 papers selected for review. Articles were reviewed to describe the databases and methodology used, the strength and limitations of different techniques, and quality assurance approaches used. Due to heterogeneity between studies meta-analysis was not performed.
Results: Occupational injuries were the focus of half of the machine learning studies and the most common methods described were Bayesian probability or Bayesian network based methods to either predict injury categories or extract common injury scenarios. Models were evaluated through either comparison with gold standard data or content expert evaluation or statistical measures of quality. Machine learning was found to provide high precision and accuracy when predicting a small number of categories, was valuable for visualisation of injury patterns and prediction of future outcomes. However, difficulties related to generalizability, source data quality, complexity of models and integration of content and technical knowledge were discussed.
Conclusions: The use of narrative text for injury surveillance has grown in popularity, complexity and quality over recent years. With advances in data mining techniques, increased capacity for analysis of large databases, and involvement of computer scientists in the injury prevention field, along with more comprehensive use and description of quality assurance methods in text mining approaches, it is likely that we will see a continued growth and advancement in knowledge of text mining in the injury field.

1. Introduction

Injuries account for 9% of global mortality and it is estimated that for every single death, there are injury hospitalisations numbered in the dozens and emergency presentations in the hundreds (World Health Organization, 2007). To address this burden through prevention and earlier intervention, we need an evidence base from which to identify the risks, causes and circumstances of injury events. Many countries worldwide collect mortality and hospitalisation data in a standardised coded format using international health classifications to enable fatal and serious injury trend reporting (World Health Organization, 2008). However, injury prevention policy and programs need to be informed not just by the most serious cases, but also by the cases that are a frequent burden on the health sector across all severity levels.

Emergency department presentations, occupational health and safety incidents, and incidents requiring emergency responders (police, fire, ambulance) represent potential 'near-miss' cases where more serious injury had the potential to occur and

* Tel.: +61 7 3138 9753; fax: +61 7 3138 5515.
E-mail address: k.vallmuur@qut.edu.au (K. Vallmuur).

http://dx.doi.org/10.1016/j.aap.2015.03.018
0001-4575/© 2015 Elsevier Ltd. All rights reserved.

information from such incidents represent opportunities to focus injury prevention efforts. Information about injuries for a range of severities are available from areas such as emergency departments, workers compensation agencies, occupational health and safety departments, and other emergency responders though the injury circumstances are often collected in less structured formats to mortality and morbidity collections, often including free text items/descriptions/reports as core data fields capturing injury circumstances. While these data are unstandardized, potentially unreliable for trend/frequency estimates, often inconsistently completed and sometimes limited in scope, in some circumstances they may be the only sources of data available for particular cohorts and for a range of severity levels. Furthermore, the information provided in text in such reports has been found to often provide a richness and depth to the understanding of injury causality above and beyond coded data (McKenzie et al., 2010b). Given the enormous resources required to introduce new standardised data collections and the fact that injury surveillance is not the primary purpose of most of these collections, text data is likely to remain one of the only sources of injury information for many years. Furthermore, even in systems dedicated to collecting injury surveillance data such as the National Electronic Injury Surveillance System in the USA, the European Injury Database in Europe, and the Injury Surveillance Systems in Queensland and Victoria in Australia, coded injury data is limited in scope and the text description captured in the database are recognised as a critical element for validation and additional interrogation with many of the published papers described previously and herein drawing on these sources.

As such, research is needed to evaluate the quality of text data and to develop methods for easier (and more consistent) interrogation of these text data. To ensure replicability and comparability of findings, it is important that text search strategies are thoroughly documented. A systematic review of papers using text fields for injury surveillance was published in 2010 by the current author and colleagues and identified 41 papers, 9 of which (7 studies) focused on describing methods for interrogating text data and the bulk of which used the text data for case capture for epidemiological studies (McKenzie et al., 2010a). This paper updates the previous systematic review to synthesise recent research using machine learning approaches to mining text-based injury data in order to demonstrate how the use of text data has changed over the last five years, to describe current practices, and to make recommendations for future research to develop this field.

2. Methodology

2.1. Study question

How are machine learning approaches being used for analysing textual injury surveillance data and what are the strengths, limitations and potentials of these techniques?

2.2. Search strategy

The electronic databases PubMed Cinahl, Medline, Google Scholar, and Proquest were searched for peer reviewed papers using the search phrase: ("text mining" OR "data mining" OR "text analytics" OR "machine learning" OR "semantic analysis") AND ("injury surveillance" OR "injury epidemiology") (anywhere in the full text of the paper) with a restriction of publication dates to 2010–2014. This identified 125 peer-reviewed English language papers to be screened for inclusion/exclusion. Snowballing from bibliographies of relevant papers and citations to included papers was used to identify further papers not identified in the original search.

2.3. Inclusion/exclusion criteria

The following criteria were used to screen papers for inclusion in the systematic review:

1. The paper was published in a peer-reviewed journal.
2. The study used a health-related database which included pre-hospital/ambulatory databases, injury surveillance databases, emergency department (ED) information systems, emergency responder data, hospital information systems, mortality databases or occupational health and safety databases.
3. The main focus of the research was injury, not other acute or chronic diseases.
4. At least one of the study objectives was to use machine learning approaches to analyse textual injury data.

Only peer-reviewed journal articles were included and grey literature was excluded as the aim of the research was to evaluate the extent to which text mining research has developed in the last five years and peer-reviewed journal articles are the best quality sources widely accessible for researchers to build on prior techniques. Abstracts of all papers located using the described search strategy were screened and abstracts which did not meet the inclusion criteria were excluded from further scrutiny.

This yielded 15 potential papers (as of October 2014) for detailed screening (details: 6 papers from ScienceDirect, 1 paper from Pubmed (Medline and CINAHL only identified duplicates of Pubmed paper), and 8 unique papers from Google Scholar. After applying the selection criteria to the 15 full English language papers, all papers fulfilled the inclusion criteria. An additional paper was identified by snowballing from the 15 full papers, with a final selection of 16 papers used in the systematic review.

2.4. Synthesis of study results

Papers were reviewed and summarized in tabular form. Focus, databases, methods, quality assurance techniques, strengths and limitations were identified for each paper. Due to heterogeneity between studies meta-analysis was not performed.

3. Results

There were 16 papers which were identified through the search strategy which used machine learning approaches to analyse textual injury data (See Table 1).

3.1. Focus of papers

The most common focus of papers was on injuries occurring while working with half of the papers reviewed discussing occupational injuries (Cheng et al., 2010; McKenzie et al., 2010a; Marucci-Wellman et al., 2011; Bertke et al., 2012; Nenonen, 2013; Abdat et al., 2014; Verma et al., 2014; Zhou et al., 2014) with a mixture of databases examined for work-related injuries including occupation incident databases (Cheng et al., 2010; Abdat et al., 2014; Verma et al., 2014; Zhou et al., 2014), workers compensation claim databases (Marucci-Wellman et al., 2011; Bertke et al., 2012; Nenonen, 2013), and an emergency department database (McKenzie et al., 2010a). The remainder of the papers focused on a mixture of population groups including consumers, children, and elderly people using data sources such as consumer product safety complaint records (Pan et al., 2012, 2014), veterans health databases (Womack et al., 2010; McCart et al., 2013), specific injury hospital registry data (Berchialla et al., 2010, 2012; Hirata et al., 2013), and firefighter near miss reporting data (Taylor et al., 2014). The injury data elements which were the

Table 1
Summary of papers.

Author, Year	Focus of paper	Target population	Injury data element/s	Database	Machine learning approach	Quality assurance	Strengths of technique	Limitations of technique
(Berchialla et al., 2010, 2012)	Use foreign body registry data to identify risk factors to predict serious outcomes of hospitalisation (yes/no)	Children	Outcome	Foreign body registry data	Fully automated Bayesian Network developed from qualitative and quantitative data with no expert guidance, using only the probability distributions learnt from the data	Statistical assessment of sensitivity and specificity of model (using 10-fold cross validation) to predict hospitalisation (2010 and 2012 article) and complications (2012 article). No independent dataset used for evaluation of model	High sensitivity (93.4%) and specificity (93.8%) of predicting hospitalisation outcome. Ability to use the model to make inferences about likely outcomes for new cases based on most probable outcome (i.e. hospitalisation risk)	Complexity of interpreting conditional probability tables to illustrate outcomes. Inability to validate model on an independent dataset
(Cheng et al., 2010)	Understand occupational construction incident patterns using data mining of safety incident data	Construction industry workers	Contributory risk factors	Occupational incident database	1347 incident information coded into categories and analysed using association rules	Use of cutoff in association rules analysis of 14% rule support and 80% confidence	Patterns identified where inadequate knowledge of safety of both workers and management	Findings are not able to be generalised outside of the target population in the absence of data from other domains
(McKenzie et al., 2010a)	Comparison of three approaches to identifying cases of work-related (vs non-work-related) emergency department presentations	Workers	Activity	Emergency department-based injury surveillance system	Keyword search, index search, and content analytic text mining	Comparison of results of each method with gold standard manual coding assigned at triage	Content analytic method had higher sensitivity (77%) and high specificity (95%), compared to a basic keyword search (58% sensitivity, 99% specificity)	Gold standard' source data also needs validation and limitations of documentation affect ability to create a strong predictive model
(Womack et al., 2010)	Compared two text processing techniques to review radiology reports to identify fractures	Radiology patients	Nature of injury	Veterans health data	SQLServer (structured query language) and NegEx text processing	Gold standard established by dual coding of 400 records by two clinicians. Comparison of text processing results with gold standard	NegEx had no false positives or false negatives (even with a small number of target cases 13/400 true fractures). SQLServer had poor recall (0.15) but good precision (1.00)	SQLServer did not address negation well, using an overly conservative approach to any negation terms in text (even if not in relation to fracture presence) leading to high false negative rate
(Marucci-Wellman et al., 2011)	Coded injury mechanism (approx 40 categories) for wholesale/ retail trade workers compensation claims to assess ability to use algorithm to autocode some cases and flag cases in need of manual review	Wholesale/ retail industry workers	Mechanism	Compensation claim	Fuzzy and Naïve Bayesian models developed on training set of 11,000 cases and tested on test dataset of 3000 cases. Evaluation of two semi-automated machine coding approaches for selecting cases for manual review to enhance accuracy and minimise manual review number	Comparison of results of semi-automated machine coding approaches with gold standard	Fuzzy and naïve models agreed on 64% of cases achieving sensitivity of 0.85, higher sensitivity than using naïve model (0.7). Screening cases where models agreed and naïve prediction >.89 improved sensitivity of model 0.90, but required 50% manual coding	Use of more advanced filters to select cases for manual review could optimise the accuracy/manual review balance
(Bertke et al., 2012)	Coded broad injury mechanism (Falls vs musculoskeletal disorder) for workers compensation claims to assess ability to use algorithm to autocode cases	Workers	Mechanism	Compensation claim	Naïve Bayesian model developed on training set of 2400 cases and tested on test dataset of 7732 cases	Dual coding by trained coders to develop gold standard categories for training set. 10% random sample of test dataset manually coded by trained coder to assess sensitivity, specificity and positive predictive value	Using injury code and narrative marginally improved prediction overall (88.4% to 89.9%) but substantially improved prediction of musculoskeletal disorders (85.4% to 90.3%)	Accuracy only demonstrated for very broad common categories and for one sector and further work needed for more specific category coding, coding of infrequent causes or coding of data from different sectors

Table 1 (*Continued*)

Author, Year	Focus of paper	Target population	Injury data element/s	Database	Machine learning approach	Quality assurance	Strengths of technique	Limitations of technique
(Pan et al., 2012)	Using text classification to enable mining of association rules from consumer product complaint records compiled from USA, Europe, and China to identify	Consumers	Mechanism, object	Consumer product complaint records	Compilation of cases used a web spider to search web pages for consumer product complaint data. Cases preprocessed to extract features and association mining used to classify cases into categories with explicit association rules	Ten experts evaluated the association rules to identify whether technique was accurate in text classification amd association mining	The association mining technique showed greater than 90% precision when evaluated by experts	More detail in the methods and results is needed to allow replicability of technique. Implications and applications of technique need further elaboration
(Hirata et al., 2013)	Prediction of common and potential risks of consumer products using accident data, by designing situation graphs of agents, actions and products to model accident situations	Children	Mechanism, object	Burns registry data	Probabilistic latent semantic indexing using narrative text to create a 'situation graph' to show typical action-agent-product scenarios to enable modelling of potential risk; compared to a feature-based approach using latent features of class predefined by researcher	Not described	Using features of the different situations, predictions about potential risks with products with similar features to those involved in the accident were able to be visually displyed in situation graphs	Considerable work needed to develop the platform to enable analysis requiring large amounts of expert user input before analysis of data is possible.
(McCart et al., 2013)	Identification of falls from text recorded in outpatient records using statistical text mining	Elderly	Mechanism	Veterans health data	70% of cases from largest hospital site used as training dataset, and three models were tested:logistic regression using logit boost, support vector machine, support vector machine-cost (which provides a model with the best balance of sensitivity and specificity). Models then applied to test data from four sites to examine AUC, accuracy, sensitivity, specificity, positive and negative predictive value	Written guidelines to describe a fall were developed to create a reference standard which was tested using three clinicians to annotate a dataset of 150 cases as fall/not-fall, evaluating cohens k for small dataset to ensure concordance. Clinicians then coded 10 out of every 1000 documents classified by machine learning to calculate an overall k	All three models achieved AUC's greater than 0.95 including across different hospital sites to show the generalizability of findings	Context dependent terms (i.e. use of negation, patient history) caused classification errors and recommendation that natural language processing needs to be used in conjunction with text mining models to reduce likelihood of errors
(Nenonen, 2013)	Analysis of factors associated with occupational falls using data mining	Workers	Mechanism	Compensation claim	Almost 49,000 cases with a code indicating a slip/stumble/fall were extracted from database and decision trees and association rules were applied to the data to identify factors influencing the fall event	Use of cutoff in association rules analysis of 20% rule support and 70% confidence in line with other research in this area	Provided a good visualisation of factors impacting on falls and showed applicability for use with a large dataset to represent the relationships within dataset	Requires both data mining methodological ability and domain knowledge, and guidance and choices made by researcher about the data mining process is affected by domain knowledge and may influence results
(Abdat et al., 2014)	Identify recurring occupational accident scenarios through exploration of	Construction/ metallurgical industry workers	Contributory risk factors	Occupational incident database	Three occupational accident experts extracted generic factors from 143 accident reports. Bayesian networks built	Required agreement from all three experts before a factor included in network building phase	Ability to reduce 143 accident reports into 30 generic factors which built 8 common scenarios. Provided	Small number of cases from specific industries were examined due to need for significant input from experts, which may not be

Table 1 (*Continued*)

Author, Year	Focus of paper	Target population	Injury data element/s	Database	Machine learning approach	Quality assurance	Strengths of technique	Limitations of technique
	narrative text describing serious and fatal occupational accidents				using coded accident reports and expert knowledge, and common scenario clusters identified		a useful analysis framework approach and rich summary of recurring scenarios for targeting industry specific prevention activities	generalisable to other industries nor representative of the full range of scenarios in these industries
(Pan et al., 2014)	Identification and coding of safety factors pertaining to electric shock near miss and injury incidents from a range of web-based consumer sources (such as RAPEX, CPSC and product safety databases in China and Japan)	Consumers	Contributory risk factors	Consumer product complaint records	Named entity recognition used to parse unstructured data. Bayesian Network developed from narrative and coded data, safety factors extracted and key factors identified by knowledge reasoning junction tree approach	Comparison of key factors identified by algorithm with rankings from 10 experts of key factors	Effective extraction of the safety factors (into risk and impact factors) from unstructured text from multiple sources, and high correlation between predicted high risk factors and expert rankings of high risk factors	Complexity of replicating method for different products or on different databases, which requires high degree of information science and domain knowledge
(Taylor et al., 2014)	Coded injury outcome and injury mechanism (precipitating and proximal mechanisms assigned from almost 40 categories) for both near miss and injury fire and emergency events collected in a Firefighter Near Miss Reporting System	People involved in incident reported by fire fighters	Mechanism, outcome	Firefighters near miss reporting system	Fuzzy and Naïve Bayesian models developed on training set of 764 cases and tested on test dataset of 1285 cases	Triple coding by researchers to develop gold standard categories for training dataset. Comparison of Bayesian models with gold standard. Random sample of 300 test dataset manually coded by three researchers to assess prediction strength	Fuzzy model outperformed naïve model achieving 74% accuracy compared to 68% accuracy. Fuzzy model more suitable for longer narratives as it chooses words which are most predictive, rather than using all words as is used in naïve model which has been shown to be stronger predictor for short narratives	Training dataset could have been larger to achieve higher accuracy as each increment of 100 cases into trainined dataset improved prediction, with around 1000 cases suggested by authors to be ideal. Modifications of the preprocessing techniques suggested also to improve the prediction
(Verma et al., 2014)	Understand steel plant occupational incident patterns using data mining of safety incident data	Manufacturing industry workers	Contributory risk factors	Occupational incident database	843 incident information coded and verified by a safety expert and analysed using association rules	Use of cutoff in association rules analysis of 1% rule support and 34% confidence. Use of 'multiple loop rules generation' to provide a causal chain of events	Unique patterns identified which provided relevant information for management to enable implementation of safer processes	Rules which are generated are highly specific to the population being investigated and need to be tested on other datasets to claim applicability outside of this study. Findings reliant on quality of documentation of supervisors recording the incident
(Zhou et al., 2014)	Use of network theory to provide a better understanding of the pattern of subway construction accidents	Construction industry workers	Contributory risk factors	Occupational incident database	241 incidents analysed and visualised using spider software to extract types of accidents (nodes) and examine the connection of nodes (edges)	Calculation of path length and diameter of nodes, clustering coefficent and betweenness centrality to deal with complex network	Visualisation of primary and secondary accidents allowed for identification of potential areas of intervention for early warning system	Findings are highly reliant on data and reflective only of patterns within population of interest and need to be continually rerun as new data is available to ensure accurate representation of patterns

target of machine learning prediction were mechanism of injury, contributory risk factors, objects/products involved, outcome, nature of injury and activity.

3.2. Machine learning approaches

A mixture of machine learning approaches was used from the more basic content analytic/text processing type approaches (McKenzie et al., 2010a; Womack et al., 2010) through to complicated network analytic (Berchialla et al., 2010; Berchialla et al., 2012; Abdat et al., 2014; Pan et al., 2014; Zhou et al., 2014) and statistical discriminative modelling approaches (Hirata et al., 2013; McCart et al., 2013). The most common methods used were Bayesian probability or Bayesian network based methods (Berchialla et al., 2010, 2012; Marucci-Wellman et al., 2011; Bertke et al., 2012; Abdat et al., 2014; Pan et al., 2014; Taylor et al., 2014; Zhou et al., 2014) and binary decision methods such as decision trees and association rule mining (Cheng et al., 2010; Pan et al., 2012; Nenonen, 2013; Verma et al., 2014).

Network analytic and association rule methods are commonly used to visualise and deduce common injury scenarios and patterns using largely categorical data extracted from text/coded data fields. For example, Abdat et al. (2014) conducted Bayesian network analysis using coded accident reports and expert knowledge to identify common recurring occupational accident scenarios, and Pan et al. (2014) identified safety factors pertaining to electric shock for near miss and injury incidents from a range of web-based consumer sources using Bayesian networks and a knowledge reasoning junction tree approach.

The Bayesian probability-based methods and simpler content analytic/text processing approaches often focus on predicting certain injury categories/states (such as mechanism, nature of injury, activity), the purpose being to predict membership of categories. Taylor et al. (2014) provides an example of a Bayesian probability-based method, in their research which categorised injury outcome and injury mechanism (precipitating and proximal mechanisms assigned from almost 40 categories) for both near miss and injury fire and emergency events collected in a Firefighter Near Miss Reporting System. Examples of content analytic/text processing approaches are McKenzie et al., 2010a study which compared three approaches (keyword search; index search and content analysis) to identifying cases of work-related (vs non-work-related) emergency department presentations; and Womack's et al. (2010) study which compared the accuracy of text processing techniques (SQL and NegEx) to identify fractures from radiology reports.

3.3. Quality assurance methods

The three main approaches to assess the quality of the techniques were comparison of findings with gold standard manually coded data, evaluation of the findings by content experts and using statistical techniques to set strict criteria in the modelling process to reduce the likelihood of extraneous results.

Sensitivity and specificity, positive and negative predictive values, and area under the curve were commonly reported when comparing to gold standard codes. For some studies, the entire dataset was already manually coded with the purpose of the study being to evaluate the ability to 'autocode' cases from text fields (McKenzie et al., 2010a; Marucci-Wellman et al., 2011), while for other datasets with only text data available, expert coders assigned codes to a sample of cases to enable the computer to learn the patterns and autocode the remainder of the data (Bertke et al., 2012; Taylor et al., 2014).

Expert review of scenario patterns was a common quality assurance method for association rule and network analytic techniques. Expert review ranged from three experts (Abdat et al., 2014) to ten experts (Pan et al., 2012, 2014) examining the patterns identified and providing qualitative responses to indicate their agreement with the commonality of scenario patterns identified.

Statistical criteria for extraction of common patterns is also an approach to assure the quality of the results. Different approaches are used dependent on the machine learning technique used and within the technique the reported criteria appears to vary considerably, influenced by the researchers needs in terms of specificity and sensitivity of results. An example from association mining research showed that the cutoff level used for two aspects (rule support and confidence) ranged from a low of 1% rule support and 34% confidence (Verma et al., 2014), through to a requirement for 20% rule support and 70% confidence (Nenonen, 2013) and 14% rule support and 80% confidence (Cheng et al., 2010).

3.4. Strengths of techniques

Machine learning techniques were found to provide valuable findings across all papers, with the main benefits being (1) high precision and accuracy in assigning broad categories to text data, (2) the ability to identify and visualise discrete injury patterns from a large amount of data with inclusion of both indirect and direct injury mechanisms in scenario extraction, and (3) the ability to make inferences about likely future outcomes based on models developed from existing data. These strengths will be discussed in turn with examples from the literature provided below.

In general, the smaller the number of categories being predicted the higher the precision and accuracy outcomes for those studies which evaluated models against gold standards/expert reviewers. Research using foreign body registry data for children which used Bayesian networks to model and predict binary hospital admission outcomes achieved 93% sensitivity and 94% specificity (Berchialla et al., 2010, 2012). Similarly, McCart et al. (2013) reported accuracy (measured by AUC) in the order of 0.95 in predicting whether an outpatient presentation was fall-related or not from Veterans Health outpatient records and Bertke et al. (2012) reported specificity of 90% and sensitivity of 88% in distinguishing falls from musculoskeletal disorders or other mechanisms. The accuracy of prediction of injury mechanism category was slightly lower (though still notably high) when a larger number of mechanism categories were predicted, with Marucci-Wellman et al. (2011) and Taylor et al. (2014) reporting overall sensitivities of 85% and 74%, respectively for Bayesian probability-based models predicting approximately 40 mechanism categories.

Secondly, the papers which used network analysis and association rule techniques to depict injury scenarios generally reported positive findings in regards to the techniques ability identify and visualise common injury scenarios from large databases. Several papers examined occupational incident reports to identify the common injury scenarios with Cheng et al. (2010) identifying inadequate knowledge of safety practices by workers and management, Nenonen (2013) illustrating the precipitating and contributory factors for occupational falls, Abdat et al. (2014) depicting the scenarios surrounding serious and fatal injuries in the construction sector, and Verma et al. (2014) extracting unique injury patterns from steel plant incident data. All of these authors noted the value of this approach in identifying scenarios which can be used by management to better target safety and prevention initiatives in the workplace. Furthermore, the incorporation of both indirect and direct injury mechanisms in the scenario extraction was seen as a valuable outcome of the technique to enable both primary and secondary prevention strategies to intervene at various stages of the causal chain (Abdat et al., 2014; Verma et al., 2014).

Finally, several papers illustrated the ability to make inferences about likely future outcomes based on machine learning models developed from existing data. Berchialla at al (2010, 2012) described how the developed model from foreign body registry data was able to used when new cases of foreign body ingestions presented by matching up the new case features with the previously modelled features to predict the likelihood of hospital admission. Hirata et al. (2013) used latent semantic indexing and feature-based approaches with burns data from hospital records to depict relationships between agents, actions and products to model accident situations and make predictions about potential risks with consumer products with similar features. Zhou et al. (2014) used a network analytic approach to modelling subway construction accidents to enable the visualisation of primary and secondary accidents which they suggest could be used as an early warning system and to target potential areas of safety intervention.

3.5. Limitations of techniques

There were limitations to the machine learning techniques however, with the limitations able to be categorised into four main areas: (1) problems with generalizability of results, (2) source data issues, (3) complex model application challenges, (4) limitations in the integration of domain and data mining knowledge. Examples of each of these limitations are provided in reference to the relevant papers below.

As machine learning techniques are highly dependent on the underlying data, research which uses highly specific datasets which are confined to small samples of cases and/or data from distinct domains/populations suffers from limitations of generalizability of results. Several authors acknowledge these limitations in their papers, suggesting that further research is needed to test the model on a larger dataset or within different populations. Verma et al. (2014) stated that the rules which are generated are highly specific to the population being investigated and need to be tested on other datasets to claim applicability outside of this study, and Zhou et al. (2014) commented that the findings are highly reliant on data and reflective only of patterns within population of interest and need to be continually rerun as new data is available to ensure accurate representation of patterns. Abdat et al. (2014), using only 143 accident reports from the construction/metallurgical industry acknowledged that the small number of cases from the specific industries which were examined (due to the need for significant input from experts in the quality assurance process), may not be generalisable to other industries nor representative of the full range of scenarios in these industries.

Similar to the previous point, the underlying source data quality and nuances affects the quality of the modelling process. Modelling relies on the quality of documentation recorded in the text as well as the accuracy of the coded data fields from which machine learning is based, and both McKenzie et al. (2010b) and Verma et al. (2014) discussed these issues. Several authors also discussed the inherent difficulties in processing unstructured text data, mentioning natural language processing issues of the use of negation, describing patients history as opposed to current circumstances (Womack et al., 2010; McCart et al., 2013).

Thirdly, the more complex machine learning technique, the greater difficulty in interpreting the output of the model and the ability to apply the findings in practice. Berchialla et al. (2010, 2012),) reported difficulty in interpreting the conditional probability tables generated through the Bayesian network technique, but suggested that a graphical representation of the relationships facilitated the interpretability of results. Three highly technical papers (Pan et al., 2012, 2014; Hirata et al., 2013) which were reviewed described complex extraction, modelling, expert review and knowledge building processes which would be very difficult to replicate in the absence of advanced data mining knowledge and content expertise.

Related to the previous point, successful applications of machine learning to injury data requires an integration of domain knowledge and technical expertise throughout each phase of the machine learning process, with the accuracy, precision and applicability of the results highly dependent on both knowledge sources. The sustainability of specialised domain and data knowledge resources coupled with ongoing time commitment to sufficiently complete machine learning projects is potentially challenging in many domains. As machine learning is an iterative process of pre-processing of data, model development, refinement and retesting under different criteria, many authors acknowledged the limitations of their presented models and the need for further refinement improvement and testing of algorithms, as well as ongoing refinement using new cases to ensure the applicability of models (Marucci-Wellman et al., 2011; Bertke et al., 2012; Nenonen, 2013; Taylor et al., 2014; Zhou et al., 2014).

4. Discussion

This paper synthesised recent research since the previous systematic review to describe machine learning approaches to mining text-based injury. Based on this summary, this final section outlines how the use of text data has changed since the previous systematic review and provides recommendations when preparing machine learning papers and suggests opportunities for future research to develop this field.

The most common focus of research using text mining was still occupational injury, with half of the papers focused on this domain. This has not changed substantially since the last systematic review with 46% of papers using text mining to report on occupational injuries. Not surprisingly, occupational incident databases and workers compensation databases comprised almost half of the data sources used, which was similar to the previous systematic review. It is not clear why occupational injury is consistently the main foci of these studies, however it could be argued that prevention of occupational injuries has a more direct financial imperative to industry and organisations than other fields of injury prevention (e.g. child injury, sports injury, violence prevention etc.), and that machine learning may be more common in other areas of business and management and hence a more accepted technique for data analysis. Largely, papers described similar foci and databases to the previous systematic review for the remainder of papers, apart from the two new inclusions of consumer product safety incident data sources (Pan et al., 2012, 2014) and a firefighter near miss data source (Taylor et al., 2014). There was less reliance on emergency department data in the machine learning space compared to the previous systematic review, with simpler techniques such as keyword searching and coding which were described in the previous systematic review more commonly used with emergency department data. However, emergency departments are increasingly recognised as a useful source of injury surveillance data and work has been done in many jurisdictions over the last ten years to improve the collection of data in emergency departments, which often includes injury (or generic presenting problem) text descriptions (Annest et al., 2008; Chow et al., 2012; Eurosafe, 2014; Gray and Finch, 2014). As collection of emergency department presentation data improves and access to and capacity to store and analyse large datasets becomes more widespread with the move to electronic health records, text mining methodologies which enable efficient and reliable case identification and classification of causes will be an invaluable injury surveillance tool. Further research evaluating the accuracy of machine learning approaches to the brief unstructured text fields recorded in emergency department data is needed.

Only three papers in the previous systematic review described Bayesian/clustering approaches to exploring text data. Hence, it was evident that over the last five years, with 16 papers reviewed, that there has been an increasing complexity in the applications of machine learning to the injury area, with the use of more advanced Bayesian networks, association rules and statistical discriminative modelling approaches to identify patterns in injury scenario text.

The previous systematic review highlighted several best practice approaches to the use of narrative text, and these points will be discussed in reference to the current review. The first recommendation was for all studies regardless of text extraction technique, to clean and parse the data prior to analysis to ensure removal of misspellings and check for consistency of terminology and abbreviations. There was limited discussion of this step in the process in any of the papers reviewed. Only one of the text methodology papers mentioned that attempts were made to clean the misspellings (Taylor et al., 2014), while other papers that mentioned misspellings chose to use the raw data in the modelling process to assess how well the data can be modelled on noisy data (Lehto et al., 2009; Marucci-Wellman et al., 2011; Bertke et al., 2012). However, previous research has found misspellings in text descriptions in health databases of between 11 and 19%, and abbreviation use in 20% of cases which varied by site, and text normalization was found to significantly improve the sensitivity of detection of cases compared to use of raw text (Shapiro, 2004). Further research is needed to evaluate the importance of text normalization in the context of injury surveillance data and compare the strength of classification algorithms under both conditions (raw data compared to normalized data). At the very least, inclusion of a rationale for using raw data or a description of the process and software used to clean and parse the data should be encouraged in the methodology section of any papers describing text mining approaches to injury surveillance.

Training datasets and iterative development and review to refine algorithms, with comparison to gold standards and use of test datasets were recommended in the previous paper for machine learning methodologies. Three of the machine learning methodology papers provided a comprehensive outline of this process and are ideal examples of strong methodolologies for text mining of injury data which future research could develop from (Marucci-Wellman et al., 2011; Bertke et al., 2012; Taylor et al., 2014).

Both the current and previous systematic review papers acknowledged the value of text mining for gathering and visualising patterns in the data to enable the depiction of chain of events in injury incidents, though the current review included the ability to make inferences about future events based on models as another benefit of machine learning. Both the current and previous papers recognised the limitations of poor source data and the lack of detailed descriptions of methodologies affecting the replicability of studies. The current review of machine learning approaches also identified the problems with generalizability of models based on very specific case data, the difficulty interpreting and using complex models in practice, and the requirements for a solid integration of domain and technical knowledge throughout each stage of the machine learning process.

The growth in machine learning approaches in injury surveillance is encouraging and studies show promising results and useful applications. However, for continued growth in this field, there is a need for more complete description of methodologies and for further comparative studies which evaluate the accuracy and applicability of a range of machine learning approaches in different contexts. Furthermore, the continued improvement of the quality, consistency and completeness of source data cannot be overlooked and research which explores methods for better collection of injury text data should also continue to be encouraged.

In conclusion, the use of narrative text for injury surveillance has grown in popularity, complexity and quality over recent years. With advances in data mining techniques, increased capacity for analysis of large databases, and involvement of computer scientists in the injury prevention field, it is likely that we will see a continued growth and advancement in knowledge of text mining in the injury field. To facilitate this development and expansion of text mining methods, attention is needed to the use of best practice approaches when designing and describing these research studies as outlined throughout the two systematic review papers.

Acknowledgement

Associate Professor Kirsten Vallmuur is the recipient of an Australian Research Council Future Fellowship (Project number FT120100202).

References

Abdat, F., Leclercq, S., Cuny, X., Tissot, C., 2014. Extracting recurrent scenarios from narrative texts using a Bayesian network: application to serious occupational accidents with movement disturbance. Accid. Anal. Prev. 70, 155–166.

Annest, J.L., Fingerhut, L.A., Gallagher, S.S., Grossman, D.C., Hedegaard, H., Johnson, R., Kohn, M., Pickett, D., Thomas, K.E., Trent, R.B., 2008. Strategies to improve external cause-of-injury coding in state-based hospital discharge and emergency department data systems: recommendations of the cdc workgroup for improvement of external cause-of-injury coding. MMWR 57, 1–15.

Berchialla, P., Scarinzi, C., Snidero, S., Gregori, D., 2010. Adaptive Bayesian networks for quantitative risk assessment of foreign body injuries in children. J. Risk Res. 13 (3), 367–377.

Berchialla, P., Snidero, S., Stancu, A., Scarinzi, C., Corradetti, R., Gregori, D., 2012. Understanding the epidemiology of foreign body injuries in children using a data-driven Bayesian network. J. Appl. Stat. 39 (4), 867–874.

Bertke, S., Meyers, A., Wurzelbacher, S., Bell, J., Lampl, M., Robins, D., 2012. Development and evaluation of a Naïve Bayesian model for coding causation of workers' compensation claims. J. Saf. Res. 43 (5), 327–332.

Cheng, C.W., Lin, C.C., Leu, S.S., 2010. Use of association rules to explore cause–effect relationships in occupational accidents in the Taiwan construction industry. Saf. Sci. 48, 436–444.

Chow, C.B., Leung, M., Lai, A., Chow, Y.H., Chung, J., Tong, K.M., Lit, A., 2012. Development of an electronic emergency department-based geo-information injury surveillance system in Hong Kong. Injury 43 (6), 739–748.

Eurosafe, 2014. EuroSafe injury data-programme. http://www.eurosafe.eu.com/csi/eurosafe2006.nsf/wwwVwContent/l3aim-aa.htm, (accessed 01.02.14.).

Gray, S., Finch, C., 2014. Victorian emergency department data for injury surveillance: how useful is it? Br. J. Sports Med. 48 (7), 601.

Hirata, A., Kitamura, K., Nishida, Y., Motomura, Y., Mizoguchi, H., 2013. Accident-data-aided design: visualizing typical and potential risks of consumer products by data mining an accident database. Proceedings of the 2013 IEEE/SICE International Symposium on System Integration, Kobe, Japan.

Lehto, M., Marucci-Wellman, H., Corns, H., 2009. Bayesian methods: a useful tool for classifying injury narratives into cause groups. Inj. Prev. 15 (4), 259–265.

Marucci-Wellman, H., Lehto, M., Corns, H., 2011. A combined Fuzzy and Naıve Bayesian strategy can be used to assign event codes to injury narratives. Inj. Prev. 1–8.

McCart, J.A., Berndt, D.J., Jarman, J., Finch, D.K., Luther, S.L., 2013. Finding falls in ambulatory care clinical documents using statistical text mining. J. Am. Med. Inf. Assoc. 20, 906–914.

McKenzie, K., Campbell, M.A., Scott, D.A., Discoll, T.R., Harrison, J.E., McClure, R.J., 2010a. Identifying work related injuries: comparison of methods for interrogating text fields. BMC Med. Inf. Decis. Making 10, 19.

McKenzie, K., Scott, D., Campbell, M., McClure, R., 2010b. The use of narrative text for injury surveillance research: a systematic review. Accid. Anal. Prev. 42 (2), 354–363.

Nenonen, N., 2013. Analysing factors related to slipping, stumbling, and falling accidents at work: application of data mining methods to Finnish occupational accidents and diseases statistics database. Appl. Ergon. 44, 215–224.

Pan, S., Wang, L., Xia, G., 2012. Mining association rules from consumer product safety cases based on text classification. J. Convergence Inf. Technol. 7 (9), 422–430.

Pan, S., Wang, L., Wang, K., Bi, Z., Shan, S., Xu, B., 2014. A knowledge engineering framework for identifying key impact factors from safety-related accident cases. Syst. Res. Behav. Sci..

Shapiro, A., 2004. Taming variability in free text: application to health surveillance. CDC MMWR Suppl. 53 (Suppl), 95–100.

Taylor, J.A., Lacovara, A.V., Smith, G.S., Pandian, R., Lehto, M., 2014. Near-miss narratives from the fire service: a Bayesian analysis. Accid. Anal. Prev. 62, 119–129.

Verma, A., Khan, S.D., Maiti, J., Krishna, O.B., 2014. Identifying patterns of safety related incidents in a steel plant using association rule mining of incident investigation reports. Saf. Sci. 70, 89–98.

Womack, J.A., Scotch, M., Gibert, C., Chapman, W., Yin, M., Justice, A.C., Brandt, C., 2010. A comparison of two approaches to text processing: facilitating chart reviews of radiology reports in electronic medical records. Perspect. Health Inf. Manage. 7, 1–12.

World Health Organization, 2007. Preventing Injuries and Violence: A Guide for Ministries of Health. WHO Press, Geneva.

World Health Organization, 2008. World Report on Child Injury Prevention. WHO Press, Geneva.

Zhou, Z., Irizarry, J., Qiming, L., 2014. Using network theory to explore the complexity of subway construction accident network (SCAN) for promoting safety management. Saf. Sci. 64, 127–136.