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