Accident Analysis and Prevention 42 (2010) 354–363

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Review

The use of narrative text for injury surveillance research: A systematic review

Kirsten McKenzie[a,*], Deborah Anne Scott[a], Margaret Ann Campbell[a], Roderick John McClure[b]

[a] National Centre for Health Information Research and Training, Queensland University of Technology, Brisbane, Queensland, Australia
[b] Monash University Accident Research Centre, Monash University, Melbourne, Victoria, Australia

ARTICLE INFO

Article history:
Received 23 February 2009
Received in revised form
23 September 2009
Accepted 25 September 2009

Keywords:
Narrative text
Injury surveillance
Text mining
Health data

ABSTRACT

Objective: To summarise the extent to which narrative text fields in administrative health data are used to gather information about the event resulting in presentation to a health care provider for treatment of an injury, and to highlight best practise approaches to conducting narrative text interrogation for injury surveillance purposes.
Design: Systematic review.
Data sources: Electronic databases searched included CINAHL, Google Scholar, Medline, Proquest, PubMed and PubMed Central. Snowballing strategies were employed by searching the bibliographies of retrieved references to identify relevant associated articles.
Selection criteria: Papers were selected if the study used a health-related database and if the study objectives were to a) use text field to identify injury cases or use text fields to extract additional information on injury circumstances not available from coded data or b) use text fields to assess accuracy of coded data fields for injury-related cases or c) describe methods/approaches for extracting injury information from text fields.
Methods: The papers identified through the search were independently screened by two authors for inclusion, resulting in 41 papers selected for review. Due to heterogeneity between studies meta-analysis was not performed.
Results: The majority of papers reviewed focused on describing injury epidemiology trends using coded data and text fields to supplement coded data (28 papers), with these studies demonstrating the value of text data for providing more specific information beyond what had been coded to enable case selection or provide circumstantial information. Caveats were expressed in terms of the consistency and completeness of recording of text information resulting in underestimates when using these data. Four coding validation papers were reviewed with these studies showing the utility of text data for validating and checking the accuracy of coded data. Seven studies (9 papers) described methods for interrogating injury text fields for systematic extraction of information, with a combination of manual and semi-automated methods used to refine and develop algorithms for extraction and classification of coded data from text. Quality assurance approaches to assessing the robustness of the methods for extracting text data was only discussed in 8 of the epidemiology papers, and 1 of the coding validation papers. All of the text interrogation methodology papers described systematic approaches to ensuring the quality of the approach.
Conclusions: Manual review and coding approaches, text search methods, and statistical tools have been utilised to extract data from narrative text and translate it into useable, detailed injury event information. These techniques can and have been applied to administrative datasets to identify specific injury types and add value to previously coded injury datasets. Only a few studies thoroughly described the methods which were used for text mining and less than half of the studies which were reviewed used/described quality assurance methods for ensuring the robustness of the approach. New techniques utilising semi-automated computerised approaches and Bayesian/clustering statistical methods offer the potential to further develop and standardise the analysis of narrative text for injury surveillance.

[*] Corresponding author at: National Centre for Classification in Health, School of Public Health, Queensland University of Technology, KELVIN GROVE 4059, Queensland, Australia. Tel.: +61 7 3138 9753; fax: +61 7 3138 5515.
E-mail address: k.mckenzie@qut.edu.au (K. McKenzie).

0001-4575/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2009.09.020

Contents

To understand and describe the circumstances surrounding injury events, core data are needed to describe the conditions preceding the event, the details of the event, and the outcomes of the event. Administrative health databases, such as those collected in primary care practices, emergency departments, hospitals, mortality registration data and workers compensation databases collect various coded data and narrative text fields for the routine monitoring and analysis of injury causation and incidence. However, the value of these data to inform injury research, policy and practice is variable across systems, with the focus of most routinely collected health information systems being the collection of clinical/diagnostic data and/or administrative data for planning/billing purposes.

Many of these systems, particularly those relying on aggregate coded data, lack the detail and granularity needed to understand the complexity of the injury event and design effective injury prevention initiatives. Many morbidity and mortality databases use the International Statistical Classification of Diseases and Related Health Problems (ICD), the international standard health classification for coding diseases and other health issues (World Health Organization (WHO) 1994). Criticisms of ICD coding systems have argued that external cause codes have an inflexible structure providing incomplete coverage and insufficient detail to identify important injury factors (Pointer et al., 2003).

Several authors have promoted and described the value of narrative text for providing extra detail to supplement routine coded data (Langley, 1998; Jones and Lyons, 2003) or for providing the required information to classify an injury post data collection if a dataset has not been coded at the point of care (Wellman et al., 2004). Many administrative health databases contain narrative text fields that can supply the additional information on injury events that a coding system may lack the granularity to provide.

However methods of obtaining the information from the narrative text are inconsistent and vary from study to study and depend on the field of research. Approaches range from basic keyword searches of text strings (Hume et al., 1996; Collins et al., 1999; Fradkin et al., 2005; Farmakakis et al., 2007; Hammig et al., 2007; Sikron et al., 2007) through to complex statistical approaches using Bayesian methods and technological methods using language processing techniques (Wellman et al., 2004; Muscatello et al., 2005; Brooks, 2008). There is currently no recommended approach to inform researcher and practitioners in their use of narrative text, and no focused delineation of the current state of the art to support developments in the field.

A systematic review of the literature was conducted to summarise the extent to which narrative text fields in administrative health data are used to gather information about the events precipitating presentation to health care providers for treatment of injury. The aim was to synthesise the information obtained from this literature review in order to describe current optimal practice and make recommendations for future research to develop the potential of this form of injury surveillance.

1. Methodology

1.1. Study question

Can narrative text fields in administrative health databases be used for extracting information for injury surveillance purposes?

1.2. Search strategy

The electronic databases PubMed, EBSCO Host, Proquest, were searched for peer-reviewed papers using the following search phrase: (“text mining” OR “narrative text” OR “text descrip*” OR “text field”) AND injury. This identified 57 papers to be screened for inclusion/exclusion. Google Scholar was searched using the above search phrase and this search located 2520 papers. A modified search of (“text mining” OR “narrative text” OR “text descrip*” OR “text field”) AND “injury surveillance” located 130 papers to be screened for inclusion/exclusion. Snowballing was used to identify 1 further paper (which was not identified from the original search) from bibliographies of relevant papers.

1.3. Inclusion/exclusion criteria

The criteria for including papers in the systematic review were as follows:

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, hospital information systems, mortality databases or occupational health and safety databases.

Table 1
Summary of injury epidemiology papers using narrative text field.

Author, year	Injury focus	Population/industry	Activity	Database	Extraction method	Quality assurance	Strengths of narrative text	Limitations of narrative text
(Bentley et al., 2002, 2005)	General	Forestry	Work	OH&S	Coding	NA	Detailed circumstance information	Inconsistent recording → errors in interpretation
(Bentley et al., 2007)	Adventure sports	General	Sport	Compensation claim	Coding	NA	Case identification	Missing text/inconsistent recording → underestimate of cases
(Bulzacchelli et al., 2008)	General	Manufacture	Work	OH&S	Coding	10% sample recode by independent coders	Detailed circumstance information	Inconsistent recording → errors in interpretation
(Bunn et al., 2008)	Tractor	Farming	Work	OH&S	Keywords and coding	NA	Detailed circumstance information	Missing text/inconsistent recording → underestimate of cases
(Burt and Overpeck, 2001)	Sports	General	Sport	ED	Coding	Coded by 2 independent coders and compared	Ability to code activity	Lack of guidelines for recording affecting data quality
(Collins et al., 1999)	Forklift	Automotive	Work	OH&S	Keywords	NA	Case identification	NA
(Dement et al., 2003)	Nail-gun	Construction	Work	OH&S	Keywords and coding	NA	Case identification and detailed circumstance information	NA
(D'Souza et al., 2007)	Ladder	General	General	ED	Coding	NA	Detailed circumstance information	Lack of detail in narrative limiting extraction of meaningful information
(Farmakakis et al., 2007)	Ingestion of objects	Children	General	ED	Keywords	NA	Detailed circumstance information	NA
(Finch et al., 1998)	Sports	General	Sport	ED	Coding	NA	Ability to code activity	NA
(Fordyce et al., 2007)	Burns	Electricity workers	Work	OH&S	Coding	NA	Detailed circumstance information	Lack of detail in narrative limiting extraction of meaningful information
(Fradkin et al., 2005)	Golf equipment	Children	Sport	ED	Keywords	Manual review with inconsistent or miscoded cases removed	Case identification	NA
(Hammig et al., 2007)	Basketball	Adults	Sport	ED	Keywords	Original coded data recoded and inconsistent cases removed	Case identification	Missing text/inconsistent recording → underestimate of cases
(Hendricks and Layne, 1999)	Fast food restaurant	Adolescents	Work	ED	Coding	NA	Detailed circumstance information	Lack of guidelines for recording affecting data quality
(Hume et al., 1996)	Trampoline	General	General	ED	Keywords	Manual review and removal of incorrect cases	Case identification	NA
(Husberg et al., 2005)	General	Construction	Work	Trauma reg	Review	NA	Detailed circumstance information	NA
(Kemmlert and Lundholm, 2001)	Slip/trip/fall	General	Work	OH&S	Coding	NA	Detailed circumstance information	Missing text/inconsistent recording → underestimate of cases
(Langley, 1998)	General	General	General	Hospital data	Keywords	NA	Detailed circumstance information	Lack of guidelines for recording affecting data quality
(Lipscomb et al., 2003)	Carpentry	Construction	Work	OH&S	Review	NA	Detailed circumstance information	NA
(Lombardi et al., 2005)	Welding	Construction	Work	OH&S	Keywords	Manual review and removal of incorrect cases	Case identification and detailed circumstance information	Missing text/inconsistent recording → underestimate of cases

Table 1 (*Continued*)

Author, year	Injury focus	Population/industry	Activity	Database	Extraction method	Quality assurance	Strengths of narrative text	Limitations of narrative text
(McGeehan et al., 2006)	Escalator	Children	General	ED	Coding	NA	Detailed circumstance information	Missing text/inconsistent recording → underestimate of cases
(Qazi et al., 2001)	Curling iron	General	General	ED	Coding	NA	Detailed circumstance information	NA
(Saltzman et al., 2005)	Partner violence	General	General	ED	Review	Coded by 2 independent coders and compared	Case identification	Missing text/inconsistent recording → underestimate of cases
(Sikron et al., 2007)	TV tipover	Children	General	Trauma reg	Keywords	NA	Case identification	Inconsistent recording → errors in interpretation
(Simon et al., 2006)	Sports	Children	Sport	ED	Coding	NA	Ability to code activity	NA
(Smith et al., 2006)	Ladder	General	Work	OH&S	Keywords and coding	Manual review and removal of incorrect cases	Case identification and detailed circumstance information	Missing text/inconsistent recording → underestimate of cases
(Warner et al., 1998)	General	Automotive	Work	OH&S	Keywords	NA	Detailed circumstance information	NA

3. The study objectives were to:

- Use text fields to assess accuracy of coded data fields for injury-related cases OR.
- Use text field to identify injury cases OR.
- Use text fields to extract additional information on injury circumstances not available from coded data OR.
- To describe methods/approaches for extracting injury information from text fields.

As the aim of this research was to describe current approaches and optimal practice when utilising narrative text fields, the inclusion criteria strictly focused on peer-reviewed journal articles and grey literature was excluded. The rationale for this was that a) peer-reviewed journal articles represent the higher quality sources of information for synthesis and b) information available in peer-reviewed articles is widely accessible to enable researchers to build on prior techniques and approaches to using narrative text fields.

Abstracts of all papers located using the described search strategy were screened by 2 authors (DS and KM) and papers which did not meet the inclusion criteria were excluded from further scrutiny. This yielded 52 potential papers for detailed screening by these 2 authors (DS and KM) (Details: 10 from Pubmed (out of 13), 12 unique papers from EBSCO Host (out of 32 abstracts; 12 duplicate papers), 1 unique paper from Proquest (out of 12 abstracts; 5 duplicates), 28 unique papers from Google Scholar (out of 130 abstracts; 13 duplicates), 1 paper identified by snowballing). After review and discussion of papers, the final review included 41 papers that met all inclusion criteria.

1.4. Synthesis of study results

Papers were reviewed and summarised in tabular and text form. Due to heterogeneity between studies meta-analysis was not performed.

1.5. Results

The published, peer-reviewed literature contains a number of studies conducted using narrative text in health administrative data. Our study identified 3 main genres of studies (as identified by the studies aims/objectives) conducted around the use of the narrative text:

1. Studies which used the narrative text for injury epidemiological purposes to select cases for analysis and/or to extract additional information to 'add value' to coded data, or 'injury epidemiology papers' (28 papers).
2. Studies which used the narrative text to retrospectively validate coded data, or 'coding validation papers' (4 papers).
3. Studies which specifically describe methods for automating/standardising the classification of narrative text for injury cases into defined injury-specific categories, or 'text extraction methodology papers' (7 studies, 9 papers).

1.6. Injury epidemiology papers

There were 28 papers identified through the search strategy which used narrative text specifically for the purpose of selecting cases and/or extracting additional information to describe the epidemiological trends in the injury topic of interest where there was insufficient coded data to identify these factors (see Table 1). The focus of these papers and sources of data are summarised in Table 1, with the majority of these papers examining occupational injuries (14 out of 28 papers).

1.6.1. Methods for extracting and using text information

The injury epidemiology papers could be categorised into 2 groups in terms of the use of text information, with 5 papers using text information to enable the selection of cases only and 23 papers reporting on the use of text information to extract additional details regarding circumstances of injuries.

The 5 papers using text information to select cases used basic keyword searches of narrative text fields to enable the selection of cases which were unable to be identified via coded data (Hume et al., 1996; Collins et al., 1999; Lombardi et al., 2005; Hammig et al., 2007; Sikron et al., 2007). Once cases were selected, the patterns and trends in injuries could then be examined by variables of interest, though no further coding of narratives was undertaken for these studies.

Of the 23 papers using the text for additional information, 16 papers used the text for firstly selecting cases for review and secondly extracting additional details regarding the injury event, while 7 papers used the text information for extracting additional circumstances details only. Sixteen of the 23 papers undertook manual coding of the narratives to enable the capture of data from the text for use in further quantitative analyses, with the information that was coded ranging from simple classification of type of sport undertaken in 4 papers (Finch et al., 1998; Burt and Overpeck, 2001; Simon et al., 2006; Bentley et al., 2007), or simple classification of mechanism of injury (Qazi et al., 2001; McGeehan et al., 2006; D'Souza et al., 2007), to more complex classification of elements (all occupational injury papers) such as initiating event, equipment, objects, tasks being performed, mechanism of injury, etc. (Hendricks and Layne, 1999; Kemmlert and Lundholm, 2001; Bentley et al., 2002, 2005; Dement et al., 2003; Smith et al., 2006; Fordyce et al., 2007; Bulzacchelli et al., 2008; Bunn et al., 2008). These papers used standardised classification systems for the coding of these more complex elements such as the American National Standards Institute (ANSI) codes, Bureau of Labor Statistics (BLS) codes, Standard Industrial Classification (SIC) of occupation, International Statistical Classification of Diseases and Related Health Problems (ICD) codes, and International Classification of External Causes of Injury (ICECI). The remaining 7 papers provided limited information regarding the process for extracting additional information from the text (Langley, 1998; Warner et al., 1998; Lipscomb et al., 2003; Fradkin et al., 2005; Husberg et al., 2005; Saltzman et al., 2005; Farmakakis et al., 2007).

1.6.2. Quality assurance methods

Only 8 of the 28 papers described quality assurance methods which were undertaken to ensure case selection was appropriate and/or coding was reliable. Five papers reported that a manual review of selected cases was undertaken to ensure cases had been selected appropriately, with miscoded or wrongly selected removed from subsequent analysis (Hume et al., 1996; Fradkin et al., 2005; Lombardi et al., 2005; Smith et al., 2006; Hammig et al., 2007). To assess the reliability of coded data, two studies used two independent coders for the coding of all selected records (Burt and Overpeck, 2001; Saltzman et al., 2005), and one study used an independent coder to assign codes for a sample of 10% of records (Bulzacchelli et al., 2008), with inconsistencies in coding identified and consensus reached.

1.6.3. Strengths and limitations of narrative text

The major strengths of narrative text which were highlighted within the injury epidemiology papers reviewed were that narrative text:

1. allowed the identification of cases which were unable to be identified through a specific code;
2. enabled the coding of more specific elements than coded within the coding system;
3. improved the capture of more specific detail regarding the circumstances leading up to the injury event, objects involved, tasks performed, mechanisms, etc.

The most important limitations according to the authors of the injury epidemiology papers which were reviewed included the lack of guidelines for recording information in the text field and subsequent inconsistency in the documentation of information in the text field which affect the data quality and could lead to errors in interpretation, the lack of detail for some cases affecting the ability to extract meaningful information regarding the case, and the extent of missing information leading to underestimates of the true magnitude of the injury and/or risk factor under investigation.

1.7. Coding validation papers

There were 4 papers which were identified through the search strategy which used narrative text specifically for the purpose of validating coded data (Smith and Langley, 1998; Amoroso et al., 2000; Jones and Lyons, 2003; Gillam et al., 2007; Indig et al., 2008).

1.7.1. Focus of papers and sources of data

The coding validation papers used a range of sources and focused on different injury aspects. Two papers used Australian emergency department text data to validate coded data fields, with Gillam et al. exploring the validity of coded injury surveillance using text fields (Gillam et al., 2007) and Indig et al. focusing on the capture of alcohol involvement in coded data compared to text fields (Indig et al., 2008). Smith et al. examined drowning fatalities to validate coded external cause data compared to text fields in a mortality database in New Zealand (Smith and Langley, 1998), while Amoroso et al. assessed the quality of injury information from coded and text data in military hospital databases (Amoroso et al., 2000).

1.7.2. Methods for extracting and using text information

Coding validation papers used a variety of methods including keyword searches, qualitative reviews and/or manual coding to extract information from text fields. Smith et al. used a keyword search of 'drown' in the mortality text field and compared the cases identified through this method to those identified using ICD-9 codes for drowning, and they found that narrative text identified almost 18% more drowning cases than code alone using this method (Smith and Langley, 1998). Similarly, Indig et al. used a detailed keyword search with a range of alcohol terms/synonyms to identify cases where alcohol involvement was documented in text fields and compared this to cases identified using ICD diagnostic codes for alcohol involvement, and the text from all cases was manually coded to identify the presences of drug use, aggression, police involvement, injuries, and mental health issues (Indig et al., 2008). Diagnostic codes only identified 24% of the alcohol-related cases with the text search algorithm identifying 76% of alcohol-related presentations.

Amoroso performed a qualitative review of injury cases recorded in a military hospital database whereby a single epidemiologist evaluated the accuracy of coded data and comprehensiveness of text data by examining both fields to see whether a) the reviewer agreed with the original codes and b) whether additional information was recorded in the free text for specific items of interest (Amoroso et al., 2000). They found that, while the narrative text provided useful additional information, there was a lack of consistency in the recording of these data thereby limiting the utility of these data for case selection purposes.

Gillam et al. used a systematic manual approach to the validation of coded data by having 4 health professionals, who were

Table 2
Summary of text methodology papers.

Author, year	Aim of study	Database	Available data	Methods	Quality assurance	Results/conclusion regarding narrative text
(Williamson et al., 2001)	To examine value of narrative text for examining patterns of work-related fatalities across 3 countries	National datasets of occupational fatalities, Australia, United States, New Zealand	Type of Occurrence Classification System (TOCS) codes; Brief narrative injury text field	Developed text search then tested text search on 200 records and modified text search to include more terms until no accuracy gains found	Text search algorithm revised through iterative search and manual review and comparison with gold standard manual coding	Narrative coding using text search more sensitive for some causes; Text-based search showed good specificity but underestimated true number of cases; Accuracy in text searches improved using a text search dictionary.
(Jones and Lyons, 2003)	To explore if narrative text from ED is able to be interrogated systematically for injury surveillance purposes	All Wales Injuries Surveillance System, Wales United Kingdom	Coded external cause data; Brief narrative injury text fields	Manual search of records and extraction location, activity and intent into new variables; Index of terms used for automated search of narratives to identify additional cases. Tested sensitivity and specificity of algorithm compared to code	Development of algorithm on one dataset and testing on new dataset	Use of narrative enabled capture of additional information to improve records coded as ‘other/unspecified’; Coded data more sensitive detecting work injuries than narrative; Narrative more sensitive than code for sport-related injuries; Narrative improves surveillance data but not replacement for coding
(Lincoln et al., 2004)	To identify if narrative text from safety reports has enough contributory information for injury prevention initiatives	Army Safety Management Information System database, United States	Narrative injury text coded using BLS Occupational Injury and Illness Classif and ICECI	Reconstruction template developed to extract information from text for injury elements; Manual coding of information; Presented as hazard scenario	Relevant elements extracted by three authors and coded by one person	Narrative text provided additional info to code for task, contributory factor, precipitating mechanism, and primary source; Systematic approach for use, coding, and presentation of narrative provided valuable info to inform prevention initiatives
(Wellman et al., 2004)	To examine accuracy of automated classification method of text for external causes of injury	National Health Interview Survey, United States	ICD-9 codes; Brief narrative injury text field	Text parsed and indexed; Keyword synonyms grouped; Computer trained using fuzzy Bayesian approach; Comparison of categories assigned by algorithm and manual coding	Comparison of computer classification with gold standard codes assigned by coder	Reasonable classification accuracy using automated approach; Multiple-word model higher sensitivity and specificity than single-word model; Automatic assignment to broad categories better accuracy than for specific categories; Probability thresholds can be set high in Bayesian classifier to filter out difficult cases for manual coding
(Bondy et al., 2005; Glazner et al., 2005; Lipscomb et al., 2004)	To classify narrative text using Haddon’s Matrix to identify factors involved in injury event	Workers Comp Claim database for Denver Airport construction, United States	Mechanism of Injury Event (MOIE) adaptation of ANSI MOI; Brief narrative injury text field	Use of structured Haddon’s Matrix abstraction tool to manually code contributory factors into human factors, objects, environment and organization by pre-event/event and post-event factors	Coded by 2 independent coders, interrater reliability test, rules developed and modified to improve consistency; Discrepancies mediated by safety expert	Use of theoretical framework for coding leads to more complete coding and identification of contributory factors; Limited narrative information for coding all aspects; Establishing causal chain of events difficult; Need for clear rules and instructions for coders of text to ensure consistency

Table 2 (*Continued*)

Author, year	Aim of study	Database	Available data	Methods	Quality assurance	Results/conclusion regarding narrative text
(Muscatello et al., 2005)	To describe use of automated syndromic surveillance system for emergency department presentations	Department of Health, ED surveillance database, New South Wales Australia	Emergency department information system codes; Presenting problem and nurse assessment text fields	Naïve Bayesian classification of free text into likely 'syndrome' from presenting problem/nurse assessment fields (based on training dataset using ICD codes and text fields to establish syndrome categories)	Comparison of automatic syndrome classification with diagnosis-based syndrome to assess sensitivity	Broad categorisation of cases from text strongly correlated with categorisation from diagnosis; Triage text provides more real-time information and more likely to be completed than diagnosis code; Text classification algorithms can be used to assign cases to infinite number of categories not confined to codes available
(Brooks, 2008)	To use SAS Text Miner to enable the mining of text fields for more detailed analysis of occupational injuries	Workers Compensation Claim database for Victoria, Australia	Narrative text fields 'Injury and Accident Text' and 'Claims Text' from Victorian WorkCover Authority	Manual review to add codes to text where able to supplement text; Iterative manual review of clusters to build stop lists/synonym lists and iterative data clustering using SAS; manual review to establish clusters	Manual review at each step to ensure validity and recall of data; Assessment of criterion-related validity to compare coded data and data derived from text fields	Text-mining software relies on quality of documented text; text-mining software may have errors in 2–4 times the number of cases than coded data and suggest large datasets needed for text mining to work after removing problem data (~20% in this study)

blinded to the original codes, use the injury description text field to code the injury circumstances using ICECI and compare and discuss codes until consensus was reached between the 4 recoders (Gillam et al., 2007). The codes reached on consensus were then compared to the originally assigned ICECI codes to explore areas of consistency and discrepancy in the coding. There was agreement between the original codes and the recoded data using the text field of approximately 92% for intent and 79% for cause of injury. Errors in coding by the original coders were identified by the narrative review and recoding process, which highlighted discrepancies.

1.7.3. *Quality assurance methods*

Only one of the coding validation papers described methods for ensuring the reliability of the data that was recoded, with Gillam et al. reporting that four independent coders coded the extracted text and final codes were decided through a process of consultation and consensus (Gillam et al., 2007). While the other papers assessed the validity of coded data by text review, they did not describe the quality assurance methods regarding their approach to text extraction or recoding.

1.7.4. *Strengths and limitations of narrative text*

The major strengths of narrative text which were highlighted within the coding validation papers reviewed were that narrative text:

1. allowed the identification of more cases than were identified the use of codes alone;
2. provided additional details regarding the circumstances surrounding the injury event than could be extracted by the code;
3. provided a means of recoding data retrospectively for specific research of interest or using different classification schema;
4. enabled the identification of systematic errors in coding and variations in the terminology used by clinical staff when documenting information.

The most important limitations according to the authors of the coding validation papers which were reviewed included the lack of consistency around the recording of information in the text fields. This lack of consistency stemmed from both a lack of guidance for staff in the completion of these text fields and variations in the terminology used by clinical staff when documenting information.

1.8. *Text extraction methodology papers*

There were 9 papers which were identified through the search strategy which had the objective of providing detailed methodologies for the extraction of text information for use in injury surveillance (see Table 2).

1.8.1. *Methods for extracting and using text information*

The methods for extracting and using text information that are described ranged from manual coding using structured abstraction tools and semi-automated text search algorithms, to more complex statistical/technical approaches using Bayesian/clustering methods.

The two occupational health and safety studies (one of which had elements of the study described in 3 separate papers) (Lincoln et al., 2004; Lipscomb et al., 2004; Bondy et al., 2005; Glazner et al., 2005) used structured abstraction tools to facilitate the complete coding of injury circumstances from text fields. Lincoln et al. developed a 'reconstruction template' from army safety reports and information was manually coded for activity, task, contributing factor, precipitating mechanism, primary/secondary source, exposure/event, nature of injury, and outcome using the Bureau of

Labor Statistics (BLS) Occupational Injury and Illness Classification and the ICECI (Lincoln et al., 2004). Cases were then presented in 'hazard scenario' diagrams (using MS Excel) for the frequency of cases by injury event/exposure, precipitating mechanism, task, and contributory factor. Bondy and colleagues (Lipscomb et al., 2004; Bondy et al., 2005; Glazner et al., 2005) used workers compensation data from the Denver International Airport construction and applied Haddon's matrix to develop a 'structured abstraction tool' to manually code contributory factors into human factors, objects, environment and organization by pre-event/event and post-event factors. These narratives were coded using an adaptation of the American National Standards Institute (ANSI) Mechanism of Injury (MOI) codes, to code both the underlying mechanism and the direct mechanism involved in the injury. Narratives and codes were imported into QSR N5 software for review and quality assurance of coding.

Williamson et al. (Williamson et al., 2001) and Jones and Lyons used a combination of manual and automated methods for interrogating text data (Jones and Lyons, 2003). In both studies a text search algorithm was developed via a process of manual review and testing on subsets of records. Williamson et al. used an abbreviated version of established classification system, the Type of Occurrence Classification System (TOCS) to form the index for the text search algorithm, while Jones and Lyons used a manual review process to extract relevant search terms to develop an index. In both studies, this index was used to develop a text search algorithm for automated searching of text fields which was then tested on a new dataset to enable the comparison of cases identified from coded fields and those identified from text fields. Jones and Lyons found that coded data were more sensitive in detecting work-related injuries than narrative text, as key word searches of the narrative did not identify the cases were the 'place of work' was stated; while in contrast, narrative text searches were more sensitive than coded data for sport-related injuries as both location and cause of the injury code be extracted from the text data to capture the type of sport (Jones and Lyons, 2003). Similarly, Williamson et al. found that narrative text was more sensitive for some causes than other causes depending on the accuracy and specificity of the search algorithm (Williamson et al., 2001).

The three studies using more complex statistical approaches for the classification/identification of information from text field also incorporated manual review and refinement of techniques as a key aspect of the process (Wellman et al., 2004; Muscatello et al., 2005; Brooks, 2008). In all three studies, the first stage of the process involved a data cleaning element which included some/all of the following fixes: misspellings were identified and corrected, abbreviations were expanded, 'stop/drop word' (such as and, or, the, etc.) which added unnecessary noise were removed, negation words were linked to the subsequent word (i.e. 'no seatbelt' became 'no-seatbelt'). Secondly, text was parsed and keywords/phrases were identified and synonyms were grouped with key words. Thirdly, practice datasets which had the target category information coded were used to enable Bayesian classifying software to identify clusters of words with a high probability of association with the target category. Manual review of the clusters was undertaken in an iterative approach to refine the clusters and 'train' the software to improve the accuracy of the categorisation process. Wellman undertook this exercise using both a single-word only list and a multiple-word list, and found that the multiple-word list had a higher sensitivity and specificity for the assignment of cases to categories. The final step was then to run the classifier on a test dataset (not used in the development and training process) and conduct further reviews of the data to assess the accuracy, sensitivity and specificity of computer assigned categories against those assigned by manual coders.

1.8.2. Quality assurance methods

All seven 'text methodology' studies described approaches for ensuring the quality and reliability of text extraction and coding. Each study, regardless of the interrogation method described, reported that an iterative process of extraction, manual review, revision and re-extraction was undertaken. Two studies (four papers) described the use of several independent coders/data extractors to assess the reliability of coded data (Lincoln et al., 2004; Bondy et al., 2005; Lipscomb et al., 2004; Glazner et al., 2005). Three studies described the process of identification/classification of cases using semi-automated methods which were then compared with a gold standard manual coding approach to assess the validity of auto-assigned codes/categories (Williamson et al., 2001; Wellman et al., 2004; Muscatello et al., 2005). Three papers discuss the process of developing search algorithms based on 'training' datasets and then assessing the accuracy of these algorithms on 'test' datasets (Jones and Lyons, 2003; Muscatello et al., 2005; Brooks, 2008).

1.8.3. Strengths and limitations of narrative text

The major strengths of narrative text which were highlighted within the text extraction methodology papers reviewed were that:

1. an important first step in the process of interrogating text data is the cleaning of the data and the development and refinement of keyword lists and indexes through expert manual review;
2. systematic approaches for interrogating text data extracted valuable additional information which was not available in coded form and reduced the amount of manual coding necessary for specific causes of interest;
3. automated classification can remove some of the variability inherent in human coding and is more easily replicated on large datasets than manual coding enables;
4. automated classification methods can be used to assign more simple cases to categories and to filter out the more complex cases for manual review.

The most important limitations according to the authors of the text extraction methodology papers were that:

1. extraction methods had a variable specificity and sensitivity across injury causes;
2. automated extraction approaches rely on standardised consistent data entry to gain the most accuracy in the classification process, which is often not the case with injury text fields;
3. irrelevant words in the dataset creates 'noise' which inhibits the ability of an automated classifier to accurately assign cases to categories.

2. Conclusion

Manual review and coding approaches, text search methods, and statistical tools have been utilised to extract data from narrative text and translate it into useable, detailed injury event information. These techniques can and have been applied to administrative datasets to identify specific injury types and add value to previously coded injury datasets. However, there is considerable variation in the approaches used to interrogate narrative text and limited documentation regarding the methodologies used and quality assurance undertaken to ensure the validity of conclusions.

The major approaches to narrative text mining for injury surveillance which have been used include: a) manual review and recoding methods using relevant standardised classification systems to capture additional information from text fields, b) keyword searches using either individual words or detailed indexes of words to select cases and identify additional information of interest, and c) semi-automated computer-based approaches using Bayesian/clustering principles to categorise cases based on broad injury elements of interest.

The main strengths of narrative text-based approaches to injury surveillance are: narrative text mining enables the identification of cases which are unable to be identified through a specific code, coding of more specific elements than that coded within the coding system is possible and text fields allow for the recoding of data using alternative classification schema retrospectively depending on the focus of the study, narrative text allows for capturing sequential chain-of-event information which is not able to be fully captured in single codes, and narrative text can be used to identify systematic errors in coding and limitations of the classification systems.

The major weaknesses with using narrative text for injury surveillance are: the lack of guidelines for recording information in the text field and subsequent inconsistency in the documentation of information in the text field which affect the data quality and could lead to errors in interpretation, the lack of detail for some cases affecting the ability to extract meaningful information regarding the case, and the extent of missing information leading to underestimates of the true magnitude of the injury and/or risk factor under investigation. Furthermore, only a few studies thoroughly described the methods which were used for text mining and less than half of the studies which were reviewed used/described quality assurance methods for ensuring the robustness of the approach.

This review has highlighted several systematic approaches to the use of narrative text which could strengthen the quality of information produced from narrative text for injury surveillance. Firstly, for all studies regardless of the approach used, a systematic approach towards the cleaning and parsing of data prior to interrogation are important to ensure the consistency of terminology, removal of misspellings, expansions of abbreviations and other inconsistencies in the text. Secondly, for studies utilising keyword search methods or semi-automated computerised approaches to text mining, a vital next step is the creation of detailed indexes developed through manual expert review by several researchers with content knowledge of the area of concern, along with supplementation of index lists using relevant standardised classification systems. This phase involves an iterative development and review process (manual and/or semi-automated) whereby search algorithms based on indexes are assessed in terms of specificity and sensitivity to ensure appropriate levels of accuracy are attained depending on the purpose of the study (sometimes specificity needs have to be balanced against sensitivity needs (Williamson et al., 2001)). For studies involving a manual review and recoding approach, the use of a standardised classification system for the coding of data undertaken by two or more coders, along with strict guidelines for the extraction of data, inclusion and exclusion criteria, and a structured data collection form are recommended. Bayesian clustering approaches have been suggested as possible alternatives to classification of relatively well structured text for simple categories, to reduce the number of cases requiring manual review. For studies using these semi-automated approaches, specific attention is needed in the cleaning, parsing and structure of the text field, as text fields lacking in structure and consistency limit the ability to use semi-automated approaches because the inconsistent words create too much 'noise' in the dataset for the machine learning to work effectively. Furthermore, for semi-automated approaches the use of a training dataset developed through an iterative development and review process to assess and refine the clusters identified, with comparison to gold standard categories (assigned by expert coders), along with a test dataset to assess the validity and generalisability of categories is recommended.

Given the importance of using systematic approaches to text mining for narrative text to be used most effectively for injury surveillance, and the current paucity of methodology papers in this area, more emphasis needs to be placed on using and describing approaches to text mining in methods sections or via designated methodology papers in this field. These methodology papers need to thoroughly document the approach used, data collection tools developed, index lists, quality assurance methods, and measures of specificity and sensitivity (McCullough and Smith, 1998). Documentation of detailed methodologies will enable the development and refinement of text-mining approaches for injury surveillance and prevention research.

While systematic approaches to the interrogation and interpretation of narrative text is an important consideration when using these data for injury surveillance, it is also important to recognise that the text fields can only be useful if the information is provided accurately and thoroughly in the first instance. Future practical applications of research work on narrative texts could be used to inform improvement to data collection methods and identification of structured approaches for recording of text data at the point of presentation. Improving the input of these data will reduce the amount of preparation work needed prior to interrogating narrative text fields and improve the consistency and reliability of the data extracted.

This review aimed to highlight best practise approaches to conducting narrative text interrogation to improve the use of these data for injury surveillance purposes. With the move towards electronic health records, where large amounts of text information will be available more routinely, alternatives sources of information regarding injuries will be available beyond the routinely used coded data. With the development and refinement of systematic approaches for interrogating narrative text fields, new opportunities exist to gather valuable supplementary information to inform injury prevention initiatives.

References

Amoroso, P.J., Smith, G.S., et al., 2000. Qualitative assessment of cause-of-injury coding in U.S. military hospitals: NATO standardization agreement (STANAG) 2050. American Journal of Preventive Medicine 18 (3 Suppl.), 174–187.

Bentley, T.A, Page, S.J., et al., 2007. Adventure tourism and adventure sports injury: the New Zealand experience. Applied Ergonomics 38 (6), 791–796.

Bentley, T.A., Parker, R.J., et al., 2005. Understanding felling safety in the New Zealand forest industry. Applied Ergonomics 36 (2), 165–175.

Bentley, T.A, Parker, R.J., et al., 2002. The role of the New Zealand forest industry injury surveillance system in a strategic ergonomics, safety and health research programme. Applied Ergonomics 33 (5), 395–403.

Bondy, J., Lipscomb, H., et al., 2005. Methods for using narrative text from injury reports to identify factors contributing to construction injury. American Journal of Industrial Medicine 48 (5), 373–380.

Brooks, B., 2008. Shifting the focus of strategic occupational injury prevention: mining free-text, workers compensation claims data. Safety Science 46 (1), 1–21.

Bulzacchelli, M.T., Vernick, J.S., et al., 2008. Circumstances of fatal lockout/tagout-related injuries in manufacturing. American Journal of Industrial Medicine 51 (10), 728–734.

Bunn, T.L, Slavova, S., et al., 2008. Narrative text analysis of Kentucky tractor fatality reports. Accident; Analysis and Prevention 40 (2), 419–425.

Burt, C.W., Overpeck, M.D., 2001. Emergency visits for sports-related injuries. Annals of Emergency Medicine 37 (3), 301–308.

Collins, J.W, Smith, G.S., et al., 1999. Injuries related to forklifts and other powered industrial vehicles in automobile manufacturing. American Journal of Industrial Medicine 36 (5), 513–521.

D'Souza, A.L., Smith, G.A., et al., 2007. Ladder-related injuries treated in emergency departments in the United States, 1990–2005. American Journal of Preventive Medicine 32 (5), 413–418.

Dement, J.M, Lipscomb, H., et al., 2003. Nail gun injuries*** among construction workers. Applied Occupational and Environmental Hygiene 18 (5), 374–383.

Farmakakis, T., Dessypris, N., et al., 2007. Magnitude and object-specific hazards of aspiration and ingestion injuries among children in Greece. International Journal of Pediatric Otorhinolaryngology 71 (2), 317–324.

Finch, C, Valuri, G., et al., 1998. Sport and active recreation injuries in Australia: evidence from emergency department presentations. British Journal of Sports Medicine 32 (3), 220–225.

Fordyce, T.A., Kelsh, M., et al., 2007. Thermal burn and electrical injuries among electric utility workers, 1995–2004. Burns 33 (2), 209–220.

Fradkin, A.J, Cameron, P.A., et al., 2005. Children's misadventures with golfing equipment. International Journal of Injury Control and Safety Promotion 12 (3), 201–203.

Gillam, C., Meuleners, L., et al., 2007. Electronic injury surveillance in Perth emergency departments: validity of the data. Emergency Medicine Australasia: EMA 19 (4), 309–314.

Glazner, J, Bondy, J., et al., 2005. Factors contributing to construction injury at Denver International Airport. American Journal of Industrial Medicine 47 (1), 27–36.

Hammig, B.J., Yang, H., et al., 2007. Epidemiology of basketball injuries among adults presenting to ambulatory care settings in the United States. Clinical Journal of Sport Medicine: Official Journal of the Canadian Academy of Sport Medicine 17 (6), 446–451.

Hendricks, K.J, Layne, L.A., 1999. Adolescent occupational injuries in fast food restaurants: an examination of the problem from a national perspective. Journal of Occupational and Environmental Medicine/American College of Occupational and Environmental Medicine 41 (12), 1146–1153.

Hume, P.A., Chalmers, D.J., et al., 1996. Trampoline injury in New Zealand: emergency care. British Journal of Sports Medicine 30 (4), 327–330.

Husberg, B.J, Fosbroke, D.E., et al., 2005. Hospitalized nonfatal injuries in the Alaskan construction industry. American Journal of Industrial Medicine 47 (5), 428–433.

Indig, D., Copeland, J., et al., 2008. Why are alcohol-related emergency department presentations under-detected? An exploratory study using nursing triage text. Drug and Alcohol Review, 1–7.

Jones, S.J., Lyons, R.A., 2003. Routine narrative analysis as a screening tool to improve data quality. Injury Prevention: Journal of the International Society for Child and Adolescent Injury Prevention 9 (2), 184–186.

Kemmlert, K, Lundholm, L., 2001. Slips, trips and falls in different work groups–with reference to age and from a preventive perspective. Applied Ergonomics 32 (2), 149–153.

Langley, J., 1998. Loss of narrative data in New Zealand Health Statistics public hospital discharge injury files. Australian Epidemiologist 5 (4), 18–20.

Lincoln, A.E, Sorock, G.S., et al., 2004. Using narrative text and coded data to develop hazard scenarios for occupational injury interventions. Injury Prevention: Journal of the International Society for Child and Adolescent Injury Prevention 10 (4), 249–254.

Lipscomb, H.J., Dement, J.M., et al., 2003. Direct costs and patterns of injuries among residential carpenters, 1995–2000. Journal of Occupational and Environmental Medicine/American College of Occupational and Environmental Medicine 45 (8), 875–880.

Lipscomb, H.J, Glazner, J., et al., 2004. Analysis of text from injury reports improves understanding of construction falls. Journal of Occupational and Environmental Medicine/American College of Occupational and Environmental Medicine 46 (11), 1166–1173.

Lombardi, D.A., Pannala, R., et al., 2005. Welding related occupational eye injuries: a narrative analysis. Injury Prevention: Journal of the International Society for Child and Adolescent Injury Prevention 11 (3), 174–179.

McCullough, P.A, Smith, G.S., 1998. Evaluation of narrative text for case finding: the need for accuracy measurement. American Journal of Industrial Medicine 34 (2), 133–136.

McGeehan, J., Shields, B.J., et al., 2006. Escalator-related injuries among children in the United States, 1990–2002. Pediatrics 118 (2), e279–e285.

Muscatello, D.J, Churches, T., et al., 2005. An automated, broad-based, near real-time public health surveillance system using presentations to hospital Emergency Departments in New South Wales, Australia. BMC Public Health 5, 141–1141.

Pointer, S., Harrison, J.E., et al., 2003. National Injury Prevention Plan Priorities for 2004 and Beyond: Discussion Paper. Injury Research and Statistics Series. Adelaide, AIHW.

Qazi, K., Gerson, L.W., et al., 2001. Curling iron-related injuries presenting to U.S. emergency departments. Academic Emergency Medicine: Official Journal of the Society for Academic Emergency Medicine 8 (4), 395–397.

Saltzman, L.E, Mahendra, R.R., et al., 2005. Utility of Hospital Emergency Department Data for Studying Intimate Partner Violence. Journal of Marriage and Family 67, 960–970.

Sikron, F., Glasser, S., et al., 2007. Children injured following TV tipovers in Israel, 1997–2003. Child: Care, Health and Development 33 (1), 45–51.

Simon, T.D., Bublitz, C., et al., 2006. Emergency department visits among pediatric patients for sports-related injury: basic epidemiology and impact of race/ethnicity and insurance status. Pediatric Emergency Care 22 (5), 309–315.

Smith, G.S, Langley, J.D., 1998. Drowning surveillance: how well do E codes identify submersion fatalities. Injury Prevention: Journal of the International Society for Child and Adolescent Injury Prevention 4 (2), 135–139.

Smith, G.S., Timmons, R.A., et al., 2006. Work-related ladder fall fractures: identification and diagnosis validation using narrative text. Accident; Analysis and Prevention 38 (5), 973–980.

Warner, M, Baker, S.P., et al., 1998. Acute traumatic injuries in automotive manufacturing. American Journal of Industrial Medicine 34 (4), 351–358.

Wellman, H.M., Lehto, M.R., et al., 2004. Computerized coding of injury narrative data from the National Health Interview Survey. Accident; Analysis and Prevention 36 (2), 165–171.

Williamson, A, Feyer, A.M., et al., 2001. Use of narrative analysis for comparisons of the causes of fatal accidents in three countries: New Zealand, Australia, and the United States. Injury Prevention: Journal of the International Society for Child and Adolescent Injury Prevention 7 (Suppl. 1), i15–20.

World Health Organization (WHO), 1994. International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10). WHO, Geneva.