Injury, Int. J. Care Injured 46 (2015) 891–897

Contents lists available at ScienceDirect

Injury

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

Making the most of injury surveillance data: Using narrative text to identify exposure information in case-control studies

Janessa M. Graves [a,b,*], Jennifer M. Whitehill [c], Brent E. Hagel [d,e,f], Frederick P. Rivara [b,g,h]

[a] College of Nursing, Washington State University, Spokane, WA, USA
[b] Harborview Injury Prevention and Research Center (HIPRC), University of Washington, Seattle, WA, USA
[c] Department of Health Policy and Management, School of Public Health and Health Sciences, University of Massachusetts Amherst, Amherst, MA, USA
[d] Department of Paediatrics, University of Calgary, Calgary, Alberta, Canada
[e] Department of Community Health Sciences, University of Calgary, Calgary, Alberta, Canada
[f] Alberta Children's Hospital Research Institute for Child and Maternal Health, Calgary, Alberta, Canada
[g] Department of Pediatrics, School of Medicine, University of Washington, Seattle, WA, USA
[h] Department of Epidemiology, School of Public Health, University of Washington, Seattle, WA, USA

ARTICLE INFO

Article history:
Accepted 14 November 2014

Keywords:
Head injuries
Epidemiology
Case-control study
Helmet
Recreation/sports
Narrative text

ABSTRACT

Introduction: Free-text fields in injury surveillance databases can provide detailed information beyond routinely coded data. Additional data, such as exposures and covariates can be identified from narrative text and used to conduct case-control studies.
Methods: To illustrate this, we developed a text-search algorithm to identify helmet status (worn, not worn, use unknown) in the U.S. National Electronic Injury Surveillance System (NEISS) narratives for bicycling and other sports injuries from 2005 to 2011. We calculated adjusted odds ratios (ORs) for head injury associated with helmet use, with non-head injuries representing controls. For bicycling, we validated ORs against published estimates. ORs were calculated for other sports and we examined factors associated with helmet reporting.
Results: Of 105,614 bicycling injury narratives reviewed, 14.1% contained sufficient helmet information for use in the case-control study. The adjusted ORs for head injuries associated with helmet-wearing were smaller than, but directionally consistent, with previously published estimates (e.g., 1999 Cochrane Review). ORs illustrated a protective effect of helmets for other sports as well (less than 1).
Conclusions: This exploratory analysis illustrates the potential utility of relatively simple text-search algorithms to identify additional variables in surveillance data. Limitations of this study include possible selection bias and the inability to identify individuals with multiple injuries. A similar approach can be applied to study other injuries, conditions, risks, or protective factors. This approach may serve as an efficient method to extend the utility of injury surveillance data to conduct epidemiological research.

Introduction

Administrative health and injury surveillance databases often include narrative text fields that provide additional detailed information beyond routinely coded data. Researchers use free text to validate coded data retrospectively or to obtain supplemental information about patients, illnesses, injuries, comorbidities, outcomes, or health services received [1–6]. Secondary use of data from free text illustrates one way to extend the value of electronic health information for application in clinical research, quality improvement, or public health surveillance [7]. To our knowledge, however, researchers have not used narratives or free text to obtain additional information for a case-control study.

The National Electronic Injury Surveillance System (NEISS) is an electronic database of injury information from a national probability sample of U.S. emergency departments (EDs), managed by the Consumer Product Safety Commission (CPSC). NEISS provides coded information on body part injured and injury type and has been widely used in epidemiological studies to analyze injury mechanisms for many sports. NEISS does not systematically capture information on known risk factors (e.g. alcohol use) or protective factors (e.g. helmets or other protective equipment).

* Corresponding author at: Washington State University College of Nursing-Spokane, P.O. Box 1495, Spokane, WA 99210, USA. Tel.: +1 509 324 7257; fax: +1 509 324 7341.
E-mail address: janessa.graves@wsu.edu (J.M. Graves).

http://dx.doi.org/10.1016/j.injury.2014.11.012
0020–1383/

However, since January 1, 2002, NEISS data have included 142-character narratives that provide additional detail on the injury and circumstances around its occurrence. For some injuries such as bicycling, narrative text may provide important etiological information about the injury and whether protective gear was used. Although some studies have used NEISS narrative data to ascertain the activity at the time of injury, they have not been widely used to evaluate the prevalence or effectiveness of protective devices, such as helmets.

Many head injuries can be prevented through improved use of protective gear, especially helmets. Several landmark case-control studies have shown that in bicycling, appropriate helmet use reduces head injuries by up to 85% [8–10]. These studies provide "gold standards" for estimates of the protective effect of helmets.

The primary objective of this study was to validate the use of narrative-derived exposure information by comparing odds ratios (ORs) for the association between bicycle helmet use and ED-reported head injuries obtained from NEISS with ORs previously reported in the literature. As an exploratory analysis, we also estimated ORs for head injuries associated with helmet use for sports-related injuries with NEISS narratives that report helmet use. Finally, we investigated the factors associated with report of helmet information in the narrative text.

Methods

We obtained NEISS data from the Consumer Product Safety Commission for sports-related injuries that presented to NEISS hospital emergency departments from 2005 to 2011. Injury information provided by NEISS includes age, sex, ethnicity, body part, diagnosis, discharge disposition, location of incident, consumer product(s) associated with injury, and 142-character narrative text field. Diagnosis code refers to the "most severe and specific diagnosis" given by the attending physician and is classified into one of thirty categories [11].

Study subjects included patients with any injury reported in the NEISS database between 2005 and 2011 that involved activity, apparel, or equipment associated with bicycling and other sports that typically involve helmets (skateboarding, in-line skating, snow skiing, snowboarding, horseback riding, unpowered scooters, ice hockey, football, lacrosse, mopeds (including motorized mini-bikes, scooters, or skateboards), all-terrain vehicles (ATVs), or two-wheeled, powered, off-road motorized bikes (motocross)). If two sports were listed for an injury, the narrative was reviewed by authors (JMG and JMW) to determine the attributable cause of injury.

Case ascertainment

We defined cases as head injuries and controls as non-head injuries. NEISS provides coded data for body part affected. Head injuries included any injury to the head, face, mouth, eye, or ear. Non-head injuries included injury to any other body part (neck, extremity, trunk, multiple body parts or the entire body, or not recorded).

Exposure ascertainment

For this study, exposure was defined as use of a helmet and was ascertained from NEISS free text. Narratives were reviewed using text-search algorithms in Microsoft Excel (Microsoft Corporation, Redmond, WA) to identify those containing possible iterations of the word *helmet*, including abbreviations, truncations, and common misspellings (Appendices). In narratives identified as containing helmet information, algorithms then differentiated between narratives describing a "helmeted," "unhelmeted," or "helmet use unknown." For example, if the prefix "unhel-" was identified in the narrative, the narrative was classified as unhelmeted. The category "helmet use unknown" described narratives that included mention of helmet but lacked detail as to whether a helmet was worn or not, for example: "helmet NS" (helmet not specified), "helmet?", "unknown if hel-". After categorization of narratives, a 10% random sample of all narratives containing the word *helmet* was reviewed by authors (JMG and JMW) to evaluate and improve the performance of the text-search algorithms. We also reviewed a 1% random sample of narratives that did not contain the word *helmet* to confirm that the algorithms were not systematically missing narratives where helmets were mentioned. Narratives that suggested *definitive* helmet use were those that indicated the individual as either helmeted or unhelmeted.

Statistical analysis

Helmet reporting (any helmet mention vs. no mention) and helmet status (helmeted, unhelmeted, or helmet use unknown) were described across injury types. These data were not weighted by population weights and do not produce national estimates.

For bicycling injuries, unadjusted and adjusted ORs and 95% confidence intervals were calculated using logistic regression for the effect of helmet use compared with non-use. We used multiple logistic regression to adjust for age and sex using STATA/MP 13.0 (College Park, TX). ORs and 95% confidence intervals were compared to previously published OR estimates for bicycling head injuries and helmet use [8–10].

As an exploratory analysis, we reviewed narratives for motorized and non-motorized sports in which participants typically use helmets, as defined above. For sports with at least 10% of narratives indicating definitive helmet use (helmeted or unhelmeted), we calculated unadjusted and adjusted ORs for head injury.

Proportions were used to compare definitive helmet use in NEISS narratives. These descriptors were calculated for bicycling injuries and injuries from other sports for which at least 10% of narratives indicated definitive helmet use. Variables included patient age, sex, body part, diagnosis, and discharge disposition. Ethnicity and location of incident were not included as factors due to high percentages of missing data (e.g., 27% of bicycling injuries were missing race data).

We conducted one sensitivity analysis. Rather than including "body location not recorded" as a separate category, we excluded these observations, reran analyses, and compared results to full models for bicycle injuries.

This study was exempt from human subjects review by our institutional board because it was performed using publicly available, de-identified data.

Results

Bicycling injuries

There were 105,614 bicycling injury narratives reported in NEISS from 2005 to 2011, of which 14,925 (14.1%) referenced helmet use. Of narratives that referenced helmets, 5270 (35.3%) were categorized as helmeted, 7287 (48.8%) as unhelmeted, and 2368 (15.9%) with helmet mentioned but use unknown.

A 10% random sample of bicycling injury narratives containing the word *helmet* resulted in 1493 narratives, all of which were reviewed for validation of the text-search algorithm. Among the reviewed narratives, 1486 (99.5%) mentioned the word *helmet* in the context of riding a bicycle. Four narratives did not contain the word *helmet* (the algorithm identified misspellings of other words), and three mentioned helmet but not in the context of riding a bicycle (e.g., bicycle helmet fell off shelf resulting in injury). Among