Preventive Veterinary Medicine 113 (2014) 417–422

Contents lists available at ScienceDirect

Preventive Veterinary Medicine

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

Mining free-text medical records for companion animal enteric syndrome surveillance

R.M. Anholt[a,*], J. Berezowski[b], I. Jamal[c], C. Ribble[a,d], C. Stephen[a,d]

[a] Faculty of Veterinary Medicine, Department of Ecosystem and Public Health, University of Calgary, Alberta, Canada T2N 4N1
[b] Veterinary Public Health-Institute, Vetsuisse Faculty, University of Bern, Schwarzenburgstrasse 155, CH-3097 Bern, Switzerland
[c] AQL Management Consulting, Edmonton, Alberta, Canada T6J 2R8
[d] Centre for Coastal Health, Nanaimo, British Columbia, Canada V9R 5S5

ARTICLE INFO

Article history:
Received 12 September 2013
Received in revised form 7 January 2014
Accepted 14 January 2014

Keywords:
Electronic medical record
Veterinary
Informatics
Surveillance
Text-mining
Enteric syndrome

ABSTRACT

Large amounts of animal health care data are present in veterinary electronic medical records (EMR) and they present an opportunity for companion animal disease surveillance. Veterinary patient records are largely in free-text without clinical coding or fixed vocabulary. Text-mining, a computer and information technology application, is needed to identify cases of interest and to add structure to the otherwise unstructured data. In this study EMR's were extracted from veterinary management programs of 12 participating veterinary practices and stored in a data warehouse. Using commercially available text-mining software (WordStat™), we developed a categorization dictionary that could be used to automatically classify and extract enteric syndrome cases from the warehoused electronic medical records. The diagnostic accuracy of the text-miner for retrieving cases of enteric syndrome was measured against human reviewers who independently categorized a random sample of 2500 cases as enteric syndrome positive or negative. Compared to the reviewers, the text-miner retrieved cases with enteric signs with a sensitivity of 87.6% (95%CI, 80.4–92.9%) and a specificity of 99.3% (95%CI, 98.9–99.6%). Automatic and accurate detection of enteric syndrome cases provides an opportunity for community surveillance of enteric pathogens in companion animals.

1. Introduction

The use of electronic medical records (EMR) by private practice veterinarians provides an opportunity to carry out pet disease surveillance in clinical settings because it is possible to access large numbers of clinical records without the manual labor required to access paper-based records. Sorting, filtering, and analyzing electronic records for a health event of interest require informatics. Informatics is the application of computer and information technologies to structure and analyze information and to improve problem solving and decision making in health care (Shortliffe and Blois, 2006). Informatics has been used to organize large volumes of biomedical data in human EMR's to search for useful patterns and information (Benin et al., 2005; Bloom et al., 2007; Penz et al., 2007; Chen et al., 2008; Birtwhistle et al., 2009; Singh et al., 2012). It has become a priority of public health agencies around the world to use these tools to develop automated surveillance systems to enhance early detection of public health threats (Friede et al., 1995; Yasnoff et al., 2000; Friedlin et al., 2008). Informatics has also been advocated for surveillance in veterinary medicine (Smith-Akin et al., 2007; Johnson et al., 2011; Santamaria and Zimmerman, 2011).

Electronic medical records are often documented in free text with little structure and without standardized clinical coding or fixed vocabulary. Exploiting the information

* Corresponding author at: Box 30, Site 1, RR1, Airdrie, Alberta, Canada T4B 2A3. Tel.: +1 403 860 0208; fax: +1 403 210 9740.
E-mail address: rmanholt@ucalgary.ca (R.M. Anholt).

0167-5877/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.prevetmed.2014.01.017

in free text medical records for surveillance requires the use of text-mining applications. Text mining is a component of natural language processing (NLP) and its goal is to convert free text into computer-analyzable language. The process extracts predefined types of information from text by applying statistical algorithms (trained or rule based) or by recognizing the target using word and phrase matching and rules (untrained or linguistic approach) (Mooney and Bunescu, 2005; Meystre et al., 2008; Raja et al., 2008). Information extraction is followed by analysis of the structured data that has been retrieved to identify clusters and relationships between concepts. Text mining should be valuable for case detection in surveillance if it can be performed with sufficient accuracy. Previous work has demonstrated that the performance of a text miner can be validated using established methods for measuring the accuracy of a diagnostic test (Chapman et al., 2004, 2005; Penz et al., 2007).

Collection of pre-diagnostic data that has been grouped together by a specific set of clinical features (syndromes) is known as syndromic surveillance. The appeal of syndromic surveillance is its potential to provide an early warning before laboratory-based data is available (Henning, 2004). Empirical treatment of cases without an etiological diagnosis is common in companion animal veterinary medicine (Boothe, 2012). Therefore syndromic surveillance can also provide a more comprehensive collection of data when a laboratory diagnosis is not pursued.

Enteric syndrome was the target syndrome in this study. It is a common problem in companion animals. A study found 13.9% of new insurance claims for dogs in the UK were due to enteric problems (Lidbury et al., 2008). Diarrhea, a common manifestation of enteric syndrome, has a complex pathophysiology with both infectious and non-infectious causes. Animal surveillance has provided the data necessary to better understand the epidemiology of enteric diseases and to monitor the distribution, spread, and changes in endemic diseases (Salmon, 2003). Further, there are many diagnostic tests and treatments available for managing acute diarrhea in pets, however there is little evidence of the diagnostic and therapeutic value of many of these procedures (Weese, 2011). The collection and interpretation of outcome-specific data is essential to develop clear guidelines for cost-effective approaches to the diagnosis and management of infectious diarrhea in pets (Holmes and Ramey, 2007).

Diarrhea from acute enteric infections is also one of the most commonly diagnosed illnesses in people in the developed world (Cheng et al., 2005). Of the 11 Canadian nationally notifiable enteric, food and waterborne pathogens (http://www.phac-aspc.gc.ca/bid-bmi/dsd-dsm/duns-eng.php), 6 (*Cryptosporidium* spp., *Giardia* spp., *Campylobacter* spp., *Clostridium difficile*, *Salmonella* spp., verotoxigenic *Escherichia coli*) are potentially companion animal zoonoses (Weese and Fulford, 2011). *Toxoplasma* spp. and some helminths are not reportable but they are also zoonotic enteric pathogens that can be acquired from companion animals (Weese and Fulford, 2011). Many human enteric disease outbreaks cannot be traced to food or other sources (Jones et al., 2004; Ravel et al., 2009). Pets are acknowledged sources of enteric infection for people, however, the attributable fraction of exposure to pets in human cases of enteric illness in Canada is largely unknown. To better link patterns of animal disease with human illness requires new integrated capacities for gathering and assimilating information from diverse sources and disciplines (Anholt et al., 2012).

The EMR and text-mining applications may be useful for detecting patterns of disease in animals and identifying changes that could signify increased disease risk to animals and people. Text mining has been applied successfully in human surveillance systems (Hutwagner et al., 2003; Chapman et al., 2004, 2005; Bradley et al., 2005; Penz et al., 2007; Friedlin et al., 2008). The use of text mining to retrieve targeted electronic medical records is relatively new in veterinary epidemiology. The overall objective of this project was to determine if text mining of data from veterinary EMR could provide automated, timely, and complete case detection for syndromic surveillance in companion animals as a proof of concept study. Specific objectives were to (1) develop a text-mining tool that can be used to automatically classify and retrieve cases of enteric syndrome from warehoused free text companion animal medical records; and (2) evaluate the effectiveness of the text-mining tool by estimating its diagnostic sensitivity and specificity.

2. Methods

2.1. Data

Twelve private veterinary practices with completely computerized medical records from the communities of Calgary, Cochrane, Airdrie, Chestermere, Strathmore and Okotoks in the province of Alberta, Canada were recruited for this study. Data extraction software was developed to achieve the automatic extraction and loading of the records to a data warehouse at the University of Calgary. A total of 296,086 electronic medical records were extracted from the veterinary management programs at the participating practices from January 1, 2007 to December 31, 2009. All pet species are represented in the data; the species distribution was 71.4% dogs, 27.3% cats and 1.3% for all other pet species (fish, birds, reptiles, amphibians and small mammals). The features of the practices, data extraction methodology, data warehouse and data quality, including the demographic and geographic representativeness of the data, have been described (Anholt et al., 2014).

The data were extracted as .csv files and each record included the date the animal was seen, the practice identification number, the patient identification number, the patient species, breed, sex and date of birth. The first 3 digits of the owner's home postal code provided spatial data. The appointment schedule, medical notes as recorded by the veterinarian or their staff (chief complaint, history, results of the physical exam, recommendation for further diagnostic procedures or referral, laboratory results, clinical assessment, and details of the surgical or medical management of the case) and prescription data for each case were combined into one variable named 'Note', in the data file. All of the records were in free text. There was no