Applied Veterinary Informatics: Development of a Canine Data Repository to Uncover Disease-Breed Associations

Mary Regina Boland\textsuperscript{1,4}, MA, MPhil, PhD, Margret L. Casal\textsuperscript{5}, DVM, MS, PhD, Marc Kraus\textsuperscript{3}, DVM, Anna R Gelzer\textsuperscript{5}, Dr. med.vet., PhD

\textsuperscript{1}Department of Biostatistics, Epidemiology & Informatics; \textsuperscript{2}Institute for Biomedical Informatics; \textsuperscript{3}Center for Excellence in Environmental Toxicology, Perelman School of Medicine, University of Pennsylvania
\textsuperscript{4}Department of Biomedical and Health Informatics, Children’s Hospital of Philadelphia
\textsuperscript{5}Department of Clinical Studies and Advanced Medicine, School of Veterinary Medicine, University of Pennsylvania

Abstract

Companion animals can be used to study the pathogenesis of human disease because, like humans, they develop disease spontaneously. We developed an algorithm to construct a repository from veterinary hospital records. Our framework contains two phases: a semantic and a domain-specific phase. We validated our algorithm against known breed-disease predispositions in our population of 84,405 dogs from 194 distinct dog breeds. We observed the expected breed-disease associations ($P<0.05$). We also report novel disease-breed associations.

1. Introduction

Understanding the origins of disease, including both environmental and genetic etiologies requires the use of good models. Canines are a useful model for studying many diseases, including cardiovascular diseases [1]. Pets are important because they develop disease spontaneously mirroring disease progression in humans [2]. To study disease prevalence and perform comparative analyses between humans and dogs, it is necessary to develop an accurate and validated data repository for clinical data obtained during routine veterinary care at the Matthew J Ryan Veterinary Hospital of the University of Pennsylvania (PennVet). Informatics methods are required to develop and validate data repositories [3, 4]. Research data repositories use data collected during routine clinical care. Unfortunately, data recorded during clinical care are often not collected for research purposes and therefore data entry errors occur frequently, and disease related terms can be used inconsistently [5]. Methods, including outlier detection, are often used in the human medical context to identify data anomalies and other issues with data collection [6]. Many of these techniques need to be tailored to the specific context. However, much prior work has been conducted in the human medical context with one study finding less than 2 articles published per year in the clinical veterinary informatics space [7], therefore much information is still to be learned. Some informatics work has been conducted in the veterinary context [8], but recent work has focused mainly in the Natural Language Processing domain [9, 10] and utilizing the UK/Australia VetCompass database [11].

This study describes the development of a veterinary informatics method that enables the construction of a canine data repository validated using PennVet data, a veterinary hospital within the United States of America. This data repository could be used for disease related studies [1], including identifying novel disease-breed associations.

2. Methods

We obtained data for canines treated at PennVet from their clinical records system. All canines were extracted from the system by extracting all records where the animal field was ‘canine’. The veterinary context differs in several key ways from the typical human medical system. For example, patient names (e.g., Fido) and birth dates of pets are not required to be accurate for billing purposes. Therefore January 1\textsuperscript{st} is often filled in for birth date if the information is absent. Therefore, we removed all canine records where January 1\textsuperscript{st} was listed. We primarily removed these records because we were interested in investigating birth-related disease associations as part of our future work. It is possible that some of these records could be useful for other researchers, depending on the research question. Data cleaning consisted of first a semantic data cleaning phase and then a domain-specific data cleaning phase. The semantic data-cleaning phase involved removing all breeds that were not dog breeds. Data entry problems are endemic in human clinical records systems, and the veterinary context is no different. Therefore, an individual could set the animal field to canine accidentally and then select a non-dog breed (e.g., ‘ferret’).

During the domain-specific data cleaning phase, we removed birth dates that occurred after the admit date since an individual or a pet cannot be admitted prior to their own birth. We also excluded patients where their birth date differs across visit encounters and where their weight was beyond 500 pounds as this signals a data entry error. Our final dataset consisted of 84,405 canines. We iteratively refined our algorithm until we were satisfied that the data were adequately cleaned. We investigated the relationship between dog breed and age, and the relationship between dog breed, weight and age. If outliers were identified, then we revisited the cleaning process to make further changes. We also assessed each dog breed’s association with disease and compared with the literature.

Association Analyses to Validate Data Accuracy: We performed several association analyses to validate our clinical data repository of 84,405 canines. For data validation purposes, we compared the results from our datasets to
those in the literature for disease–breed associations. We investigate the relationship between certain dog breeds and three specific diseases – mitral valve disease, atrial fibrillation and osteosarcoma [12].

**Empirical Validation Using Permutation Analysis:** We used permutation analysis to compare our results with those obtained from the literature. For each disease, we developed a random cohort of patients that was the same size as the case population. For example, there were 717 mitral valve patients. Therefore, we set 717 random patients as having the disease. We then performed breed association analysis adjusting the p-values for multiple hypotheses using the FDR metric (less stringent), using 1000 random samples. We computed the precision and recall by comparing the findings from this random analysis with the gold standard.

### 3. Results

We started with a dataset containing 84,580 canines. After the cleaning process was complete, we obtained a dataset of 84,405 canines (99.8% retained). For data validation purposes, we compared the results from our datasets to those in the literature for disease–breed associations for three diseases. Precision ranged from 60.0% to 83.3% (avg. 74.2%) and recall ranged from 31.6% to 83.3% (avg. 53.3%) depending on the disease of interest. Our method (blue line, Figure 1) outperformed random permutation results demonstrating that we were able to replicate known disease–breed associations.

**Novel Disease Associations:** We are the first study to report dog breeds that are protected against developing osteosarcoma, namely Yorkshire Terriers (OR=0.0, CI: 0.0, 0.37) and Shih Tzus (OR=0.0, CI: 0.0, 0.47). We are also the first to report at risk associations for osteosarcoma and Anatolian shepherds (OR=45.90, CI: 4.97, 208.26) and bulldogs (OR=10.90, CI: 4.60, 22.15). We also reported novel associations between dogs breeds protected against developing mitral valve disease, including the Pug (OR=0.24, CI=0.05, 0.70) and the American Pit Bull Terrier (OR=0.03, CI: <0.001, 0.15).

### 4. Discussion

Our study describes a two-phase method to construct a clinical data repository using canine data obtained during routine clinical care at a veterinary hospital while maximizing data retention (99.8% retained). Our final population contained 84,405 dogs treated between 2000 and 2017 from 194 distinct dog breeds. We observed the expected breed associations with mitral valve disease, atrial fibrillation, and osteosarcoma (P<0.05). We achieved high precision (average 74.2% across three diseases) and recall (53.3% across three diseases). Furthermore, we were able to uncover novel disease-breeds associations that warrant further investigation.

**References**

1. Boland MR, Driik E, Kraus M, Gelzer A. Cardiovascular Disease Risk Varies by Birth Month in Canines. Scientific Reports. 2018;8(7130):10.1038/s41598-018-25199-w