Comprehensive representation of health-related phenotypes in one million dogs using topic modelling of electronic health records

Peter-John Mäntylä Noble
rtnorle@liverpool.ac.uk

University of Liverpool

Sean Oliver Farrell
Durham University

Noura Al-Moubayed
Durham University

Alan David Radford
University of Liverpool

Method Article

Keywords: Veterinary clinical records, topic modelling, phenotype discovery

Posted Date: July 15th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4621991/v1

Additional Declarations: No competing interests reported.

Comprehensive representation of health-related phenotypes in one million dogs using topic modelling of electronic health records

Peter-John Mäntylä Noble[1*], Sean Oliver Farrel[2],
Noura Al-Moubayed[2], Alan David Radford[3]

[1*]Small Animal Teaching Hospital, Institute of Veterinary and Ecological Sciences, University of Liverpool, Leahurst Campus, Chester High Road, Neston, CH64 7TE, Wirral, UK.
[2]Department of Computer Science, Durham University, Upper Mountjoy Campus, Stockton Road, Durham, DH1 3LE, UK.
[3]Institute of Veterinary and Ecological Sciences, University of Liverpool, Leahurst Campus, Chester High Road, Neston, CH64 7TE, Wirral, UK.

*Corresponding author(s). E-mail(s): rtnorle@liverpool.ac.uk;
Contributing authors: sean.farrell2@durham.ac.uk;
noura.al-moubayed@durham.ac.uk; alanrad@liverpool.ac.uk;

Abstract

Historically, veterinary studies screening for breed, age and sex predisposition to disease have relied on collating small-scale studies of clinical datasets. The availability of larger datasets through groups such as the Small Animal Veterinary Surveillance Network (SAVSNET) promise access to information regarding wide range of clinical presentations at scale, however, methodological limitations surrounding the extraction of specific disease information or screening for disease predispositions result in a substantial reduction in the number of animals studied. These studies often address very focused hypotheses - only leveraging a small fraction of the intrinsic value of the data at any one time. Here, we implemented an unsupervised machine learning methodology, creating a representation of a large volume of clinical notes collected by SAVSNET from veterinary practices across the UK. We capture breed, age and sex predisposition and offer statistical and temporal possibilities across various clinically important presentations. We utilise BERTopic, a topic-modelling tool based on Bidirectional Encoder Representations using Transformers (BERT) architecture, which surfaces known phenotypes, such as breed predispositions to hypoadrenocorticism, diabetes mellitus and mitral valve disease, and potential novel patterns of disease phenotypes. This scalable and granular modelling technique facilitates the rapid interrogation of large clinical datasets, enabling the identification of broad phenotypic diversity within the population and the early detection of temporal changes indicative of emerging infectious or environmental diseases.

Keywords: Veterinary clinical records, topic modelling, phenotype discovery

1 Introduction

In the veterinary health literature, the identification of which breeds (ages or sexes) are affected by a specific disease or which range of diseases affect a specific breed has typically been studied using comparatively small cohorts of animals with a specific condition, for example, animal predisposition to mammary cancer [36] or of breeds of animals where records could be screened for wide-ranging diseases, for instance, diseases affecting French Bulldogs or Boxers [23, 25]. Such studies typically include hundreds to small numbers of thousands of individual animals. Similarly, modeling the occurrence of disease with age is not easily achieved at scale for many breeds at one time. Ultimately, consensus about disease in dogs is usually reached through a review of multiple publications, often incorporating comparatively small groups of animals.

Recently, extremely large datasets have become available through groups such as Vet Compass [22] and the Small Animal Veterinary Surveillance Network (SAVSNET, [31]). These hold significant volumes of canine electronic health records, which in turn will embody key patterns of phenotypes displayed by dogs as represented by the clinical narratives contained therein. While these systems allow for some disease-coding by attending clinicians, these suffer from either simple single-choice coding

(SAVSNET) or complex coding choices that are associated with poor compliance, and for all such coding systems, the reliability of the attending clinician to enter this data is very variable [19, 11].

The veterinary informatics sector struggles to fund work on a scale that would allow wide-ranging manual record annotation, and some form of automation of record annotation at scale might alleviate this challenge. A potential candidate for this might be machine learning(ML), which has been used to create automated record classification systems for both text and image data. However to date, these are usually supervised methods requiring both prior knowledge of the phenotypes likely to be seen and reasonably large datasets of expert-annotated records in order to train these systems [20, 9], all of which require substantive investment of time on a case by case basis. Additionally, the process by which machine learning systems generate classifications can often be opaque. This can lead to systematic errors due to the model identifying unexpected features in the data necessitating the development of systems such as LIME and SHAP to generate explanations for machine-learning outputs [29, 3]. An ideal system would involve unsupervised annotation of records according to disease phenotypes based on intrinsic characteristics of the notes. The aim would be that explainable results would be intrinsic to the method.

We have previously shown that latent Dirichlet allocation (LDA) topic modelling allowed the identification of a specific gastroenteric disease in clinical notes from dogs that would have allowed early detection of a disease outbreak [21]. LDA is based on a bag-of-words statistical method that does not differentiate word meaning according to context, potentially limiting the ability of the technique to differentiate topics. A more modern approach incorporates a combination of neural language modelling using bidirectional encoder representations using transformers (BERT) [7] to create document embeddings with subsequent dimensionality reduction using UMAP and clustering into documents with common word-composition based topics using hdbscan. This combination of techniques is implemented in the BERTopic package [10]. In addition to creating topic models, BERTopic is able to represent topic evolution across time (dynamic topic modelling) or other classifications (e.g. breed) inherent to the data under study.

Whilst novel in veterinary studies, topic modelling has been evaluated for information extraction from human electronic health records [30] and social-care notes with clinical content [34] highlighting a clear opportunity to leverage its use for veterinary data.

Here, we use BERTopic to create topic models based on a large subset of SAVSNET clinical narratives. We evaluate the patterns of phenotypes revealed within BERTopic-generated topics, comparing how topics for disease change with time and in specific breed, age, and sex cohorts compare to known disease occurrence and how the models perform in revealing new clinical insights.

2 Methods

2.1 SAVSNET datasets

SAVSNET collects data from approximately 580 veterinary premised across the UK. After each patient consultation, data are passed to SAVSNET and comprises the species, breed, age, and sex of the patients, along with the clinical narrative recorded by the attending veterinarian. The current study used a random sample of one million clinical records, each record coming from a unique dog and comprising the deidentified clinical narrative, the breed, age, sex and the date of consultation.

2.2 Topic models

Narratives from the one million record datasets were used to train a BERTopic model [10]. Briefly, BERTopic embeds documents using the sentence transformer all-MiniLM-L6-v2'[13] followed by reduction of dimensionality using UMAP [18] and subsequent clustering using HDBScan [17]. Clusters are then analysed using term frequency/inverse document frequency (TF/IDF) to identify word weighting for each topic [27]. A trial and error method was used to adjust the UMAP and HDBSCAN parameters to generate a usable model with minimal generation of duplicated and vacuous topics (A1, A2). The topics were assessed by a clinically active academic (Noble) to attribute clinical relevance. The distribution of topics by age, breed class, or consultation month was created using the topics_per_class method.

Models were trained using a Ryzen-9 12 core CPU PC with 64Gb of DDR-4 RAM equipped with an Nvidia A4000 GPU (16Gb internal memory).

2.3 Data analysis

Using this method, each clinical narrative was assigned a probability distribution for containing any topic and here, consultations were classified according to the most probable topic. Topic word-contents were reviewed using the BERTopic visualize_barchart method. Plotly [32] was used to present a filterable line plot of proportions of narratives labeled with given topics broken down by age and date. These were normalised to peak or mean values for each time series to allow comparison. For breed and sex-related data, odds ratios along with 95% confidence intervals were calculated for the proportions of topic-labelled narratives against suitable references and plotted as tree plots using Plotly.

Ethical approval declarations Owners contributing data to the SAVSNET system have the option to opt out of doing so, and the project has University of Liverpool ethics committee approval (RETH001081).

3 Results

The SAVSNET database contained 5,467,034 narratives from dog consultations. With the hardware in use, BERTopic would overrun memory limits in the GPU when used to train the model with more than 1,000,000 records. When the number of possible topics was unconstrained, BERTopic generated a model with over 900 topics, many of which were vacuous (sequences of words with no clear clinical correlate) or very similar in word composition to other topics (eg multiple combinations describing a vaccination event). As a consequence, BERTopic was run, limited to producing a maximum of 200 topics which qualitatively appeared to represent diverse clinical presentations with less duplication (HDBScan and BERTopic parameters are shown in supplementary material tables A1, A2). The top 15 most common topic representations (15 keywords for each topic) are shown in supplementary material (figure. SA1). Despite the count reduction, 53 topics still contained words relating to vaccination or booster.

3.1 Breed distribution

SAVSNET data comprised information about 217 unique breeds. Using the topics_per_class method of BERTopic allowed the distribution of topics for any breed to be identified. A wide range of potential breed predispositions could be identified. An example of 4 topics representing different endocrine diseases is shown in figure 1. Here increased occurrence of diabetes topic is seen in Samoyeds, Huskies and West Highland White Terriers whereas the description of thyroid disease is seen more commonly in standard poodles, dobermans and Scottish terriers. Hypoadrenocorticism-related narratives were seen in the standard poodle, Bearded Collies, Labradoodles and German short-haired pointers (figure.1). Similarly, breed associations for clinical syndromes involving the the cardio-respiratory system could be evaluated with an example dataset shown in figure 2. Example topics representing upper respiratory signs (sneezing, reverse sneezing), cardiac findings (heart murmur) and cough are shown and illustrate potential predispositions (e.g. murmur in Cavalier King Charles spaniel and Boxer and Chihuahua, sneeze in Pug and Chihuahua) and reduced instances (e.g. murmur in Pugs, cough in French Bulldog).

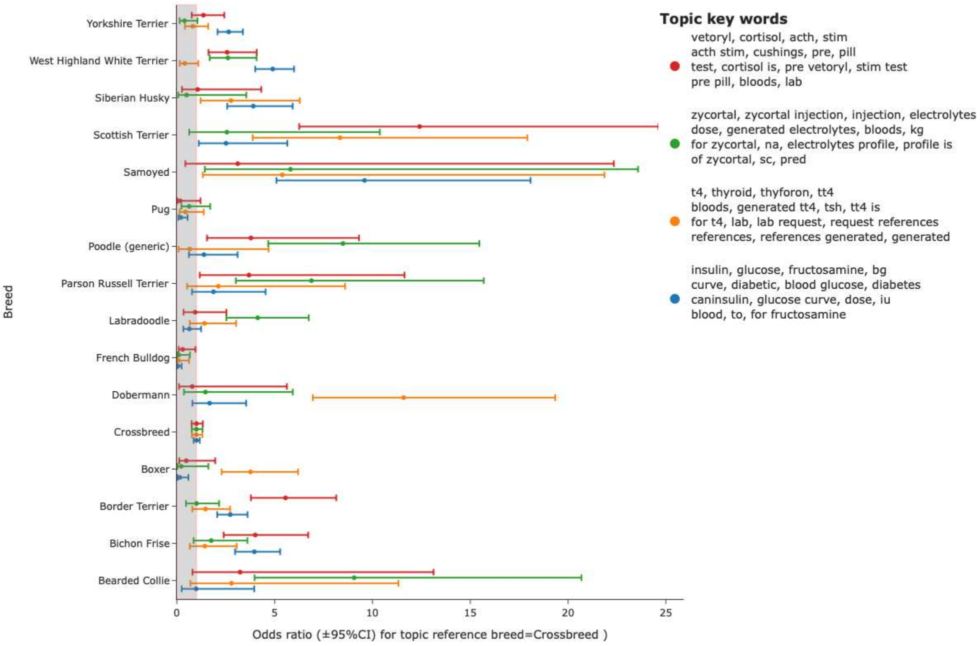

Fig. 1 Breed distribution of endocrine disease topics. Topics representing animals probably affected by diabetes(blue), hyperadrenocorticism(red), hypoadrenocorticisma(green) and hypothyroidism(orange) were readily identified by topic wording. Odds ratio and 95% confidence interval were plotted for a selection of breeds with potential predispositions to individual endocrinopathies.

3.2 Age distribution

SAVSNET data comprised records from animals aged 0-18 years old. Topics illustrated a wide range of age-related variation such that topics relating to vaccination and

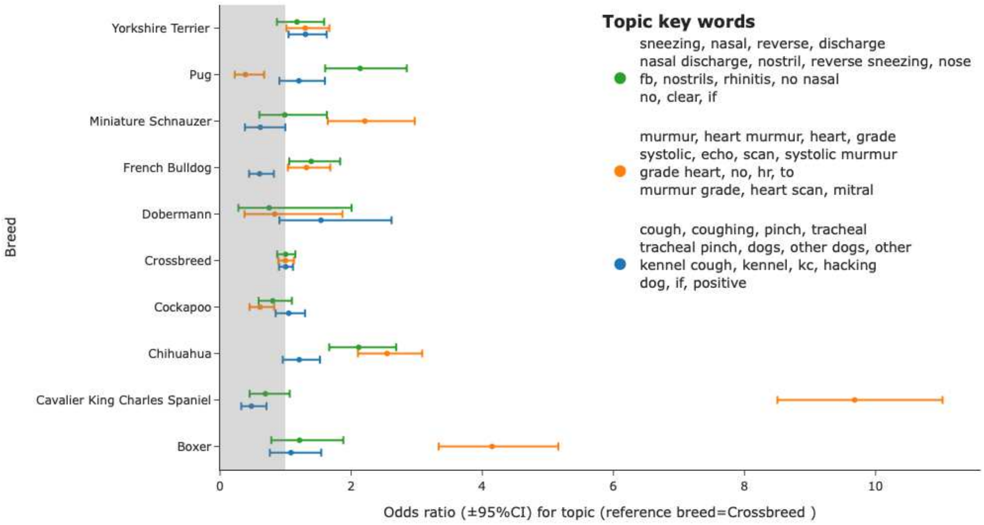

Fig. 2 Breed distribution of cough (blue),sneeze(green) and murmur (orange) topics. Here, the distribution for topics reflecting respiratory or cardiac clinical signs are shown for a selection of breed. Odds ratio and 95% confidence intervals for breed having consultations matched by the given topic

puppy health checks occurred in younger animals. Topics relating to masses occurred more commonly in middle aged dogs and topics relating to vestibular disease, and euthanasia occurred in older dogs (figure 3).

3.3 Sex distribution

Differences in topic distributions between male and female dogs were evaluated and demonstrated some variation. Odds ratios for male and female neutering were strongly (and appropriately)segregated for the relevant sexes such that the only topic containing the word spay ('spay', 'season', 'pre', 'for spay', 'pre spay', 'spay check', 'op', 'pre op', 'last season', 'lap', 'spey', 'vulva', 'check', 'mammary', 'lap spay') had an

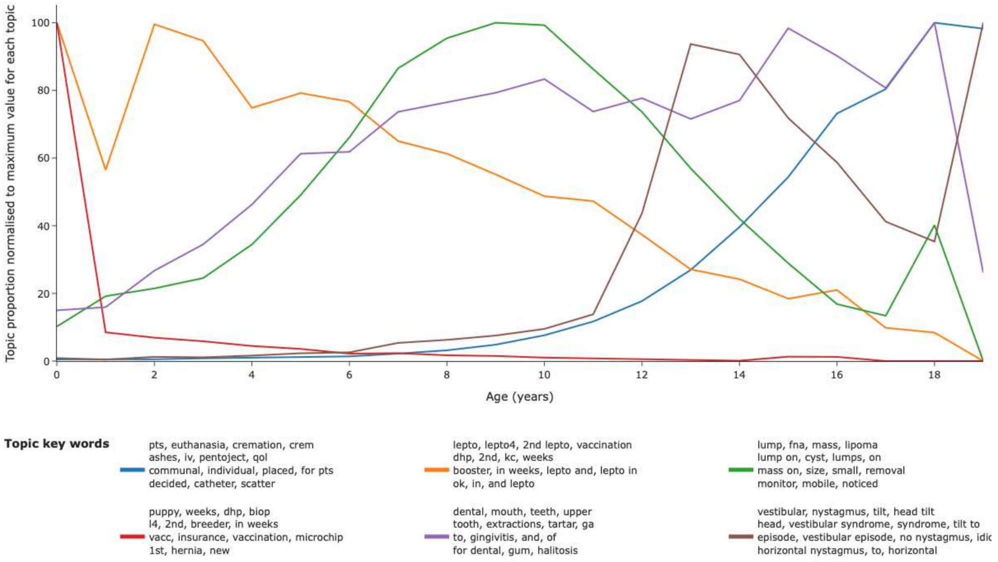

Fig. 3 Examples of topics with distinctive age distributions. Age distribution for a group of topics that represent early-life (vaccination and puppy check), mid-life (lumps and masses) and late-life (Vestibular disease and Euthanasia) presentations. Each topic is normalised to its maximum proportion

odds ratio of 0.01 in males compared to females and castration ('testicle', 'castration', 'castrate', 'testicles', 'pre', 'implant', 'suprelorin', 'both testicles', 'op', 'pre op', 'for castration', 'descended', 'check', 'for castrate', 'scrotum') topic had an odds ration of 13.7 in males compared to females . Additionally, odds ratios for muzzling (a likely marker of aggression), seizures, coughing, sneezing and skin disease topics were higher for males while odds ratio for urinary tract infection and mammary disease were lower in males (figure 4).

3.4 Outbreak patterns

Analysis of topic occurrence with time allowed for the detection of increases in clinical syndromes on a national level (using this model). Here, a topic relating to gastrointestinal signs revealed a seasonal increase in these signs in winter with a pronounced

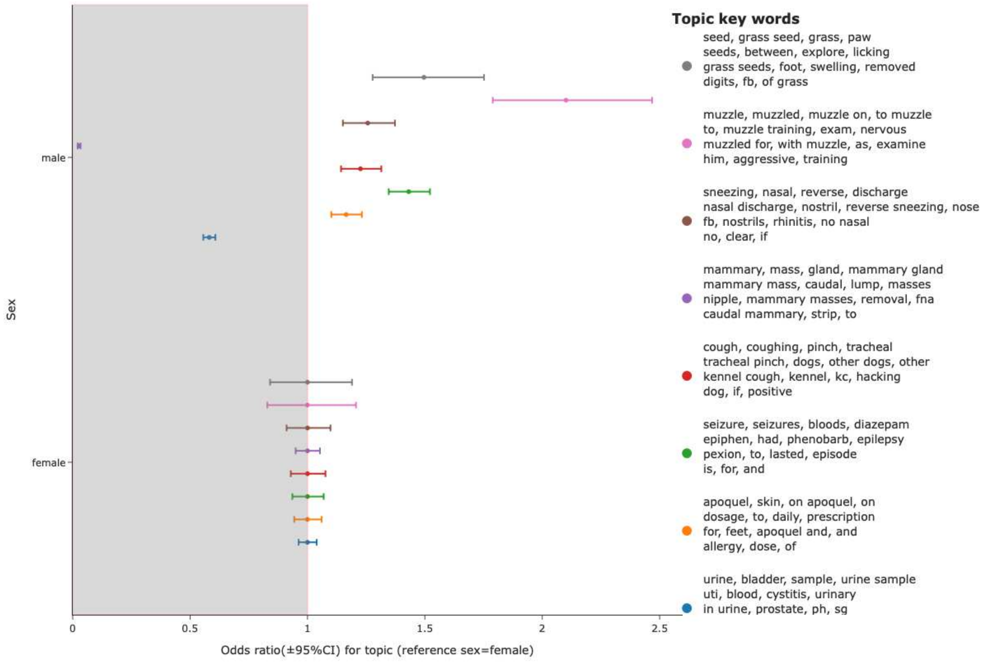

Fig. 4 Topic distribution for male and female dogs. Selected topics are shown where odds ratios suggest diffences between male and female dogs

spike in activity of this topic in winter 2020. A marked increase in activity of a topic relating to respiratory signs was seen in Autumn 2021 (figure5).

3.5 Syndrome seasonality

Given the ease of screening the patterns of topics with time, it was straightforward to identify clinical syndromes with marked seasonality. Thus topics relating to grass seed foreign bodies, firework anxiety and removal of ticks from the patient showed marked, repeatable seasonality (figure 6).

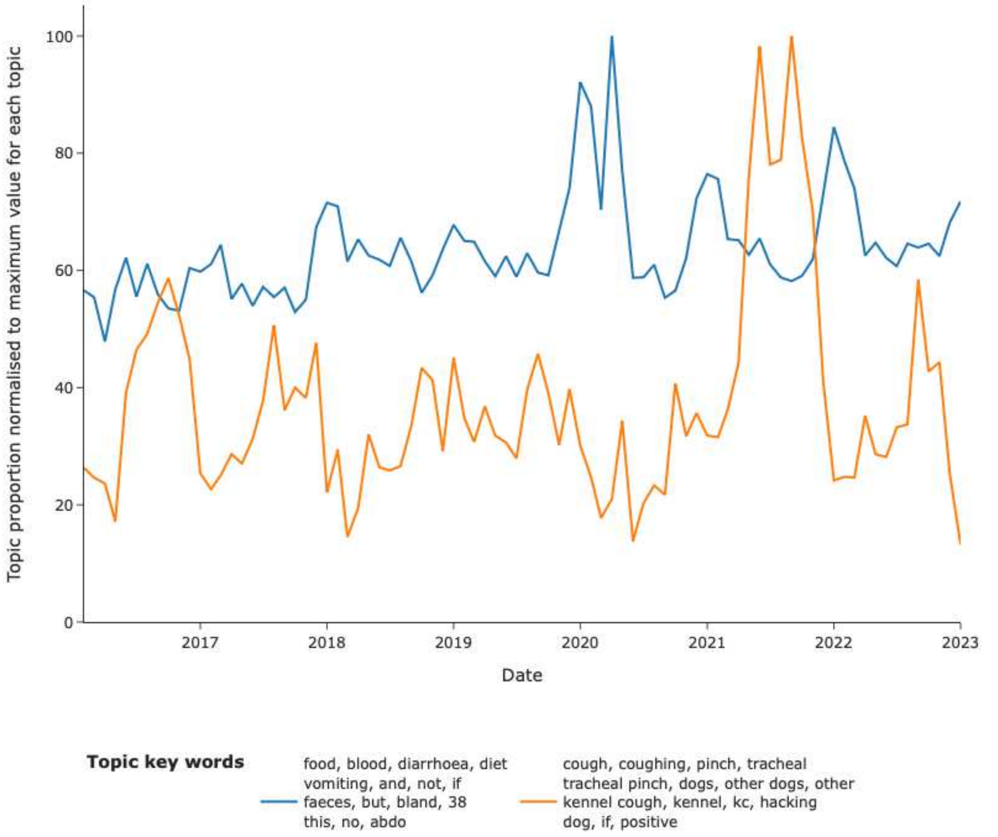

Fig. 5 Examples of topics highlighting potential disease outbreaks. The distribution of topics with time are shown for a topic most-likely representing gastroenteric disease and one representing respiratory disease.The former reflects a national outbreak of gastroenteric disease in spring 2020, the latter, a national increase in respiratory disease occurring in autumn 2021

3.6 Temporal trends

The key words for each topic often indicated that use of a specific drug was a core feature in that topic, allowing an evaluation of trends in use of those drugs with time. Example of these trends are seen in figure 7 where consultations describing meloxicam usage decrease in proportion with time where the proportion of consultations

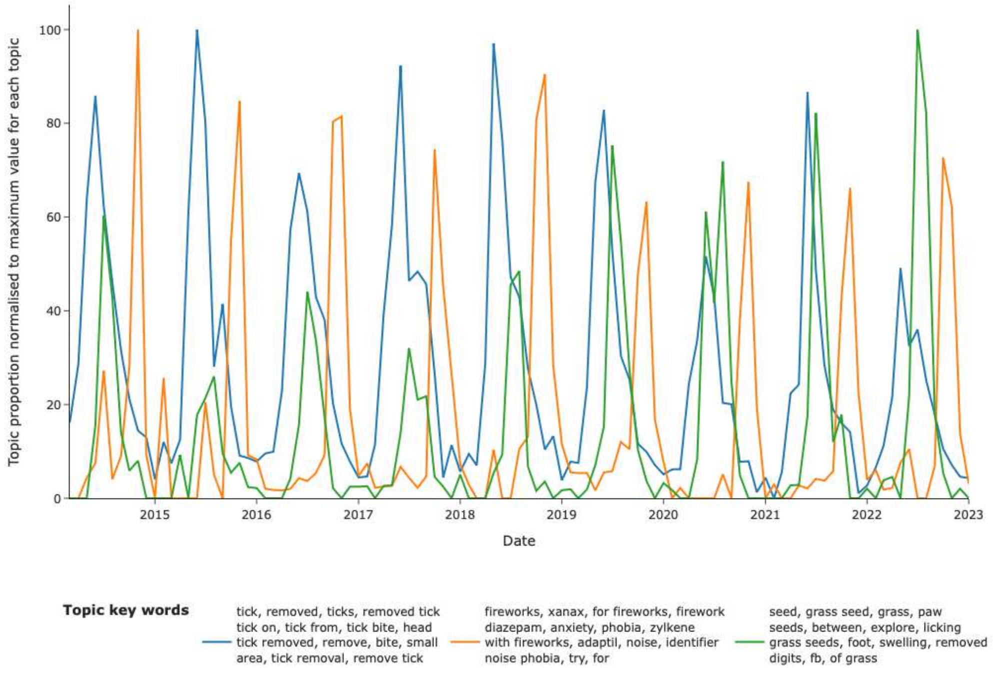

Fig. 6 Examples of topics that illustrate detection of seasonal patterns , in this case, firework anxiety(orange), tick infestation(blue) and grass-seed trauma(green)

describing use of occlacitinib increased steadily with time and bedinvetmab steeply after 2020.

3.7 Effects of lockdown

At the point of the initial lockdown for COVID in April 2020, the number of veterinary consultations decreased significantly [33] however veterinary visits continued and certain topic-labelled narratives formed a higher proportion of consults. These were identifiable as increased description of specific syndromes such as vestibular disease, torn nails, and consultations resulting in euthanasia of the dog. Other topics relating to routine health care did not occur at an increased proportion (figure.8).

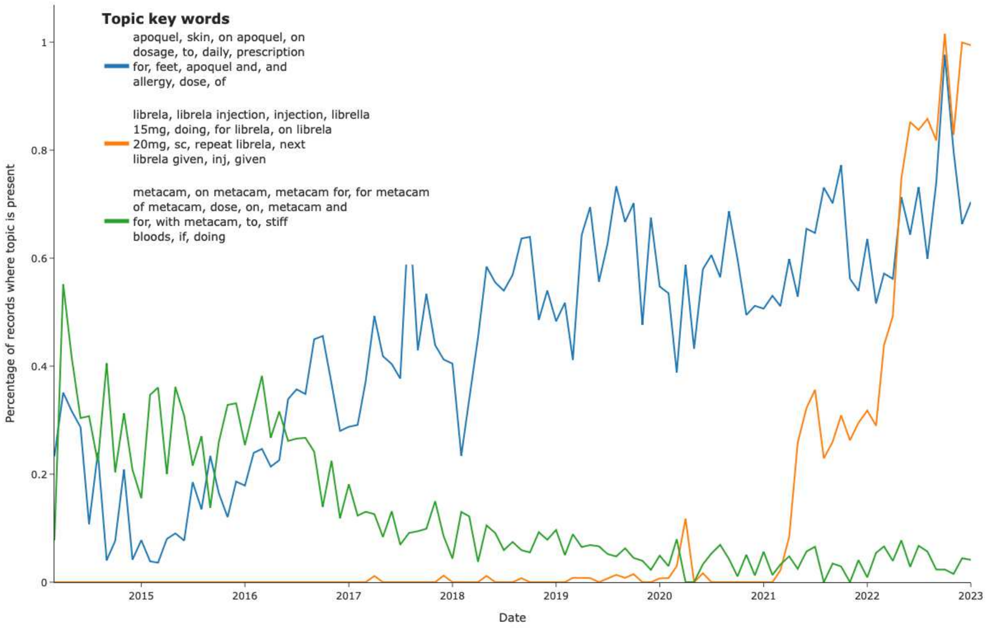

Fig. 7 Examples of topic that show long term trends with time. In this case consultations discussing use of drugs Occlacitinib (Apoquel®), meloxicam and bedinvetmab (Librela®).

4 Discussion

To realise the potential of big data, it is clear that annotation of records for specific features can not be performed manually at the million-record scale due to lack of time, compliance, limitation of the feature set that can be assessed and accuracy of annotations. We have shown that a neural language model method can be used to build a representation of clinical syndromes affecting 1 million patients. The topic model presented a rich overview of clinical features in the data, allowing a review of the prevalence of clinical syndromes across breed, age, sex and with time. As an unsupervised method, topic modelling allows the clinical phenotypes inherent to the data to

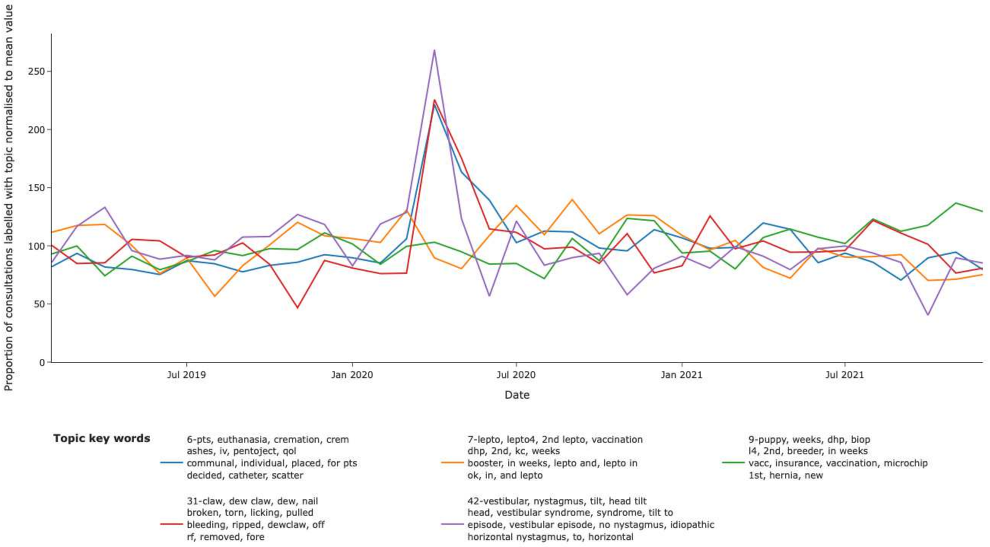

Fig. 8 Clinical priorities during lockdown. During lockdown, certain topics became less common and others were over-represented as a proportion of consultations visits. These included visits for painful, distressing or terminal disease. The proportions of routine health care visits did not increase

surface without stipulating what those phenotypes should be. Consequently, a rich set of features can be exposed, and a representative sample of which are highlighted here.

Breed predisposition to disease is always a complex challenge, even when big datasets are available, because the clinical phenotype under study may require manual annotation. The number of records that can be annotated leads to a significant reduction in group sizes when dividing down to individual breeds, of which there are many 100s in dogs. Thus, compromises are made in studies whereby only limited breed sets can be evaluated. With automated labelling, far more records are annotated, and more representatives from each breed can be evaluated. Additionally, Topic modelling immediately incorporates numerous phenotypes in a single study.

Here, we evaluated records annotated with topics relating to four common endocrinopathies. In the case of diabetes mellitus, breed predispositions noted in previous big-data studies were clearly identifiable for Border terriers, West Highland White Terriers and Yorkshire Terriers [12]. In that study, case numbers were limited due to a methodology requiring keyword search and manual reading. The wider sampling facilitated through topic modelling as described here also allowed the identification of breeds with disease predispositions, such as Scottish terrier, Siberian Husky, Samoyed [14] and Bichon Friese, additionally highlighting Boxers, Pugs and French bulldogs as under-represented breeds.

Hypothyroidism was more common in Dobermans, Scottish terriers, Boxer and Samoyed in line with a recent review [24]. Hypoadrenocorticism was seen in Beared collies, Parson Russel Terriers, Poodles and Labradoodles and Hyperadrenocorticism in West Highland White terrier, Border terrier, Bichon Friese and Scottish Terriers.

Similar reviews of the phenotype were possible, such as the evaluation of the distribution of topics relating to heart murmur and upper and lower respiratory tract disease. In this case, heart murmurs were discussed more often in Cavalier King Charles Spaniels (a long recognised predisposition [6]) and less often in Pugs, whereas in this group of breeds, sneezing was more common in Pugs and Chihuahuas and less common in French Bulldogs.

For each consultation in this dataset, full patient data (breed, age, sex, neutering status and geographical location) was available, and a comprehensive logistic regression analysis would have been possible but outside the scope or scale of this study. Here, the key feature of this study is the range of phenotypes that can be explored. For example, where previous studies have been able to capture information about diabetes mellitus, the approach demonstrated here concurrently captured data about three other endocrinopathies with no extra investment of time. To the author's knowledge, a comprehensive single study identifying wide ranging disease predispositions in this manner has not been performed.

The same model, when examined using the topics_per_class method using age as a class, allowed a rapid evaluation of the distribution of given topics with age revealing a variety of patterns, some of which were quite predictable (puppy vaccination in young dogs, booster vaccination with slow decline according to age, euthanasia more common in older animals, dental disease steadily increasing with age) with others less intuitive at first glance (lump/mass discussion in middle age, vestibular disease increasing extremely steeply after 10 years of age). This methodology is ready to combine classes of age and breed in order to highlight how individual breeds express specific phenotypes with age.

When topics were analysed per class based on patient sex, numerous examples of topics were more common in one sex or the other. Unsurprisingly, consultations relating to male neutering and female neutering and mammary disease were strongly biased, but additionally, key known predispositions were immediately surfaced, including seizures in male dogs [8], grass seed injury in male dogs [2], urinary tract disease in female dogs [4]. Additionally, consults describing dog aggression were more common in male dogs, in line with previous studies that indicate male dogs are more prone to human-oriented aggression [16]. Critically, a number of female dog narratives were labelled with the topic with keywords including testicles and castration. These frequently related to descriptions of undescended testicles in puppies brought in with their dam. In some cases these illustrated rare errors where the original data (from the practice management system of the contributing practice) contained the wrong sex highlighting an opportunity to feed back to our contributors on the accuracy of their clinical notes.

Utilising Bertopic's topic_over_time method offered the opportunity to evaluate various phenotypic features temporally in three ways, namely seasonal disease, unexpected changes in proportion of labelled narratives (potential outbreaks) and long term trends. Here, we demonstrate that phenotypes with known seasonal variation were immediately identified illustrating, known patterns of tick infestation, grass seed foreign bodies [2, 35, 1] and fireworks anxiety peaking around bonfire night. Critically this data can be reviewed over time to evaluate the impact of climate change on timing of these effects and in addition, better understanding of the timing of ectoparasite activity will help to inform more focused use of ectoparasiticides which are known to contaminate local environments [26].

The outbreak of gastrointestinal disease seen in 2020 [28] was readily identified with subsequent seasonal peaks which have since been identified [5]. Interestingly, the temporal dataset also identified a substantial change in the proportion narratives labelled with a topic describing kennel cough-like signs (acute respiratory disease) during 2021. To the authors' knowledge, this has not been demonstrated in any other reports and warrants further investigation. We demonstrated three examples of long term trends in drug usage (as described in clinical notes) suggesting a decrease in the use of a specific meloxicam-containing product (Metacam®), which may reflect the licensing of other products containing the same drug alongside the emergence of a number of COX-2 selective antagonists in recent years. A steady increase in described use of oclacitinib (Apoquel®, a drug used to manage allergic skin disease) was also seen, and a dramatic increase in the description of the use of bedinvetmab (Librella®, a drug used to manage arthritis), which coincided with its release in 2021.

This temporal approach also allowed for the evaluation of the impact of COVID-19 lock-downs. Thus, syndromes associated with pain (torn nails), distress (vestibular disease) and terminal illness and euthanasia represented a larger proportion of visits. These would all be associated with acute distress for the patient and their owners leading owners to pursue immediate care where perhaps vaccination and routine health care were considered acceptable to delay [15].

The sentence-embedding model used has a constrained input length of 512 tokens which means that a fraction (approximately 0.2%) of records will have been truncated prior to embedding and subsequent clustering which may have led to some loss of detail and range of topics detected by the system. In future studies, models with larger input lengths will be usable. The study's unsupervised topic modelling on a large veterinary clinical dataset offers a broad view of clinical syndromes among one million patients, however, the constraints set on cluster size to avoid vacuous or duplicated topics prevent rarer topics from being exposed, which may lead to important but very uncommon syndromes being missed. While topics are often easily interpreted from their key-words, their remains a requirement to audit the underlying narratives where critical conclusions are being drawn. In future work, studies will include setting thresholds for topic probability when attributing a topic to a given consultation in order to improve labelling accuracy. Despite these constraints, the study provides a foundational exploration of clinical phenotypes, offering potential avenues for future investigations with access to full patient histories for in-depth analyses.

5 Conclusion

Unsupervised Bertopic topic modelling, when applied to a corpus of veterinary clinical notes, surfaces a diverse array of clinically relevant phenotypes which can be used to expose breed, age and sex predispositions to disease and highlight seasonal and outbreak variations in the occurrence of disease in a single experiment. Additionally, this approach allows the visualisation of trends in the appearance of clinical signs and treatment modalities and changes in treatment priorities during lockdown. This methodology leverages a freely available neural language model but is of particular value in this setting because of the availability of the large SAVSNET dataset with national coverage of clinical records. The model can be used to classify narratives unseen during the initial training, and the full array of topics by breed, age, sex, and date is available to view at https://public.tableau.com/app/profile/savsnet.at.liverpool/viz/Onemilliondogshealthdata/Dashboard2

Availability of data and materials. The datasets generated and/or analysed during the current study are not publicly available but are available from the corresponding author on reasonable request subject to ethical approval. A sample of such narratives is available at https://www.liverpool.ac.uk/media/livacuk/savsnet/SAVSNET,sample,vet,data.xlsx.

Competing interests. While this work was not directly funded by any specific body, SAVSNET does undertake some commercial work for a variety of companies in order to maintain the projects sustainability.

Funding. This work was partly supported by a Dogs Trust Canine Welfare Grant. Sean Farrell is funded through BBSRC.

Authors' contributions. PJMN performed the data retrieval, topic modelling, data analysis and the majority of the paper writing. SOF advised on language model usage and reviewed the manuscript. NA helped in project design and review of the manuscript. AR helped with project design, data collection and review of the manuscript. All authors read and approved the final manuscript.

Acknowledgments. Preliminary scoping experiments were performed by Muhammad Faisal Jawed. This work would not have been possible without the contribution of practicing vets across the UK contributing to the SAVSNET project (especially the

CVS group practices) and without the help and support of the SAVSNET core team comprising Bethaney Brant, Steve Smyth and Gina Pinchbeck.

Author information. PJMN works as a companion animal clinician in the University of Liverpool Small Animal Teaching Hospital and is a co-investigator on the SAVSNET project where he supervises clinical direction, AI and software development for the project. SF is a BBSRC PhD student working in Durham using language models to examine drivers for antibiotic use in companion animals. NA is an associate professor in the Department of Computer Science at Durham University focusing on explainable machine learning, natural language processing, and optimisation. ADR is a professor of veterinary informatics and as a co-founder of SAVSNET manages its day to day running with the co-investigators.

Appendix A Model parameters

parameter	value
min_cluster_size	50
metric	'euclidean'
cluster_selection_method	'eom'
prediction_data	True
core_dist_n_jobs	8

Table A1 HDBScan parameters

parameter	value
n_gram_range	(1,2)
embedding_model	petBert
language	"english"
top_n_words	15
min_topic_size	200
hdbscan_model	hdbscan_model
nr_topics	200
low_memory	True
verbose	True
calculate_probabilities	True

Table A2 Parameters sent to BERTopic

References

[1] Elena Arsevska, Tomislav Hengl, David A. Singleton, Peter John M. Noble, Cyril Caminade, Obiora A. Eneanya, Philip H. Jones, Jolyon M. Medlock, Kayleigh M. Hansford, Carmelo Bonannella, and Alan D. Radford. Risk factors for tick attachment in companion animals in great britain: a spatiotemporal analysis covering 2014-2021. *Parasites & vectors*, 17, 12 2024. ISSN 1756-3305. doi: 10.1186/S13071-023-06094-4. URL https://pubmed.ncbi.nlm.nih.gov/38254168/.

[2] Bethaney J Brant, David A Singleton, P J M Noble, and Alan D Radford. Seasonality and risk factors for grass seed foreign bodies in dogs. *Preventive veterinary medicine*, 197:105499, dec 2021. ISSN 1873-1716. doi: 10.1016/j.prevetmed.2021.105499. URL http://www.ncbi.nlm.nih.gov/pubmed/34583207.

[3] James Burton, Sean Farrell, Peter-John Mäntylä Noble, and Noura Al Moubayed. Explainable text-tabular models for predicting mortality risk in companion animals. *Scientific Reports 2024 14:1*, 14:1–12, 6 2024. ISSN 2045-2322. doi: 10.1038/s41598-024-64551-1. URL https://www.nature.com/articles/s41598-024-64551-1.

[4] Julie K. Byron. Urinary tract infection. *Veterinary Clinics of North America: Small Animal Practice*, 49:211–221, 3 2019. ISSN 0195-5616. doi: 10.1016/J.CVSM.2018.11.005.

[5] Edward Cunningham-Oakes, Jack Pilgrim, Alistair C. Darby, Charlotte Appleton, Chris Jewell, Barry Rowlingson, Carmen Tamayo Cuartero, Richard Newton, Fernando Sánchez-Vizcaíno, Ivo Salgueiro Fins, Bethaney Brant, Shirley Smith, Rebekah Penrice-Randal, Simon R. Clegg, Ashley P.E. Roberts, Stefan H. Millson, Gina L. Pinchbeck, P. J.M. Noble, and Alan D. Radford. Emerging variants of canine enteric coronavirus associated with outbreaks of gastroenteric disease - volume 30, number 6—june 2024 - emerging infectious diseases journal - cdc. *Emerging infectious diseases*, 30:1240–1244, 6 2024. ISSN 10806059. doi: 10.3201/

EID3006.231184. URL https://wwwnc.cdc.gov/eid/article/30/6/23-1184_article.

[6] D. K. Detweiler and D. F. Patterson. The prevalence and types of cardiovascular disease in dogs. *Annals of the New York Academy of Sciences*, 127:481–516, 9 1965. ISSN 1749-6632. doi: 10.1111/J.1749-6632.1965.TB49421.X. URL https://nyaspubs.onlinelibrary.wiley.com/doi/10.1111/j.1749-6632.1965.tb49421.x.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1:4171–4186, 10 2018. URL http://arxiv.org/abs/1810.04805.

[8] Alexander Erlen, Heidrun Potschka, Holger A. Volk, Carola Sauter-Louis, and Dan G. O'Neill. Seizure occurrence in dogs under primary veterinary care in the UK: prevalence and risk factors. *Journal of Veterinary Internal Medicine*, 32(5):1665, sep 2018. ISSN 19391676. doi: 10.1111/JVIM.15290. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6189390/.

[9] Sean Farrell, Charlotte Appleton, Peter John Mäntylä Noble, and Noura Al Moubayed. PetBERT: automated ICD-11 syndromic disease coding for outbreak detection in first opinion veterinary electronic health records. *Scientific Reports 2023 13:1*, 13(1):1–14, 10 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-45155-7. URL https://www.nature.com/articles/s41598-023-45155-7.

[10] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[11] P. A. Hall and N. R. Lemoine. Comparison of manual data coding errors in two hospitals. *Journal of clinical pathology*, 39(6):622–626, 1986. ISSN 0021-9746. doi: 10.1136/JCP.39.6.622. URL https://pubmed.ncbi.nlm.nih.gov/3722414/.

[12] Angela M. Heeley, Dan G. O'Neill, Lucy J. Davison, David B. Church, Ellie K. Corless, and Dave C. Brodbelt. Diabetes mellitus in dogs attending UK primary-care practices: frequency, risk factors and survival. *Canine Medicine and Genetics 2020 7:1*, 7(1):1–19, jun 2020. ISSN 2662-9380. doi: 10.1186/S40575-020-00087-7. URL https://cgejournal.biomedcentral.com/articles/10.1186/s40575-020-00087-7.

[13] Hugging Face. all-minilm-l6-v2 sentence transformers, June 2024. URL https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2. [Accessed 2024-06-17].

[14] Susan E. Kimmel, Cynthia R. Ward, Paula S. Henthorn, and Rebecka S. Hess. Familial insulin-dependent diabetes mellitus in Samoyed dogs. *Journal of the American Animal Hospital Association*, 38(3):235–238, 2002. ISSN 0587-2871. doi: 10.5326/0380235. URL https://pubmed.ncbi.nlm.nih.gov/12022409/.

[15] Rebecca Littlehales, P. J.M. Noble, David A. Singleton, Gina L. Pinchbeck, and Alan D. Radford. Impact of Covid-19 on veterinary care. *Veterinary Record*, 186 (19):650–651, jun 2020. ISSN 20427670. doi: 10.1136/VR.M2495.

[16] M. Lord, B. A. Loftus, E. J. Blackwell, and R. A. Casey. Risk factors for human-directed aggression in a referral level clinical population. *Veterinary Record*, 181:44–44, 7 2017. ISSN 2042-7670. doi: 10.1136/VR.103638. URL https://bvajournals.onlinelibrary.wiley.com/doi/10.1136/vr.103638.

[17] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2:205, 3 2017. ISSN 2475-9066. doi: 10.21105/JOSS.00205. URL https://joss.theoj.org/papers/10.21105/joss.00205.

[18] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3:861, 9 2018. ISSN 2475-9066. doi: 10.21105/JOSS.00861. URL https://joss.theoj.org/papers/10.21105/joss.00861.

[19] Jose Antonio Miñarro-Giménez, Catalina Martínez-Costa, Daniel Karlsson, Stefan Schulz, and Kirstine Rosenbeck Gøeg. Qualitative analysis of manual annotations of clinical text with SNOMED CT. *PLoS ONE*, 13(12), dec 2018. ISSN 19326203. doi: 10.1371/journal.pone.0209547.

[20] Allen Nie, Ashley Zehnder, Rodney L. Page, Yuhui Zhang, Arturo Lopez Pineda, Manuel A. Rivas, Carlos D. Bustamante, and James Zou. Deeptag: inferring diagnoses from veterinary clinical notes. *npj Digital Medicine 2018 1:1*, 1:1–8, 10 2018. ISSN 2398-6352. doi: 10.1038/s41746-018-0067-8. URL https://www.nature.com/articles/s41746-018-0067-8.

[21] Peter-John Mäntylä Noble, Charlotte Appleton, Alan David Radford, and Goran Nenadic. Using topic modelling for unsupervised annotation of electronic health records to identify an outbreak of disease in UK dogs. *PLOS ONE*, 16(12):e0260402, dec 2021. ISSN 1932-6203. doi: 10.1371/JOURNAL.PONE.0260402. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0260402.

[22] Dan G O'Neill, David B Church, Paul D McGreevy, Peter C Thomson, and Dave C Brodbelt. Approaches to canine health surveillance. *Canine Genetics and Epidemiology*, 1(1):2, 2014. ISSN 2052-6687. doi: 10.1186/2052-6687-1-2.

[23] Dan G. O'Neill, Lauren Baral, David B. Church, Dave C. Brodbelt, and Rowena M. A. Packer. Demography and disorders of the French Bulldog population under primary veterinary care in the UK in 2013. *Canine Genetics and Epidemiology 2018 5:1*, 5(1):1–12, may 2018. ISSN 2052-6687. doi: 10.1186/S40575-018-0057-9. URL https://cgejournal.biomedcentral.com/articles/10.1186/s40575-018-0057-9.

[24] Dan G O'neill, Janine Su, Pheng Khoo, Dave C Brodbelt, David B Church, Camilla Pegram, and Rebecca F Geddes. Frequency, breed predispositions and other demographic risk factors for diagnosis of hypothyroidism in dogs under primary veterinary care in the UK. *Canine Medicine and Genetics 2022 9:1*,

9(1):1–14, oct 2022. ISSN 2662-9380. doi: 10.1186/S40575-022-00123-8. URL https://cgejournal.biomedcentral.com/articles/10.1186/s40575-022-00123-8.

[25] Dan G. O'Neill, Alison M. Skipper, Kate Barrett, David B. Church, Rowena M. A. Packer, and Dave C. Brodbelt. Demography, common disorders and mortality of Boxer dogs under primary veterinary care in the UK. *Canine Medicine and Genetics*, 10(1):6, jun 2023. ISSN 2662-9380. doi: 10.1186/S40575-023-00129-W. URL https://cgejournal.biomedcentral.com/articles/10.1186/s40575-023-00129-w.

[26] Rosemary Perkins and Dave Goulson. To flea or not to flea: survey of uk companion animal ectoparasiticide usage and activities affecting pathways to the environment. *PeerJ*, 11, 2023. ISSN 2167-8359. doi: 10.7717/PEERJ.15561. URL https://pubmed.ncbi.nlm.nih.gov/37554336/.

[27] Shahzad Qaiser, Universiti Utara, Malaysia Sintok, Malaysia Kedah, Ali Ramsha, and Text Analytics. Text mining: Use of tf-idf to examine the relevance of words to documents. *International Journal of Computer Applications*, 181:25–29, 7 2018. doi: 10.5120/IJCA2018917395.

[28] Alan D. Radford, David A. Singleton, Chris Jewell, Charlotte Appleton, Barry Rowlingson, Alison C. Hale, Carmen Tamayo Cuartero, Richard Newton, Fernando Sánchez-Vizcaíno, Danielle Greenberg, Beth Brant, Eleanor G. Bentley, James P. Stewart, Shirley Smith, Sam Haldenby, P. J.M. Noble, and Gina L. Pinchbeck. Outbreak of Severe Vomiting in Dogs Associated with a Canine Enteric Coronavirus, United Kingdom. *Emerging infectious diseases*, 27(2):517–528, feb 2021. ISSN 1080-6059. doi: 10.3201/EID2702.202452. URL https://pubmed.ncbi.nlm.nih.gov/33496240/.

[29] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 13-17-Augu, pages 1135–1144. Association for Computing Machinery, 8

2016. ISBN 9781450342322. doi: 10.1145/2939672.2939778.

[30] Emil Rijcken, Uzay Kaymak, Floortje Scheepers, Pablo Mosteiro, Kalliopi Zervanou, and Marco Spruit. Topic modeling for interpretable text classification from ehrs. *Frontiers in big data*, 5, 5 2022. ISSN 2624-909X. doi: 10.3389/FDATA.2022.846930. URL https://pubmed.ncbi.nlm.nih.gov/35600326/.

[31] Fernando Sánchez-Vizcaíno, Peter John M. Noble, Phil H. Jones, Tarek Menacere, Iain Buchan, Suzanna Reynolds, Susan Dawson, Rosalind M. Gaskell, Sally Everitt, and Alan D. Radford. Demographics of dogs, cats, and rabbits attending veterinary practices in Great Britain as recorded in their electronic health records. *BMC Veterinary Research*, 13(1):218, jul 2017. ISSN 17466148. doi: 10.1186/s12917-017-1138-9.

[32] Carson Sievert. *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman and Hall/CRC, 2020. ISBN 9781138331457. URL https://plotly-r.com.

[33] David A. Singleton, P. J. Noble, Beth Brant, Gina L. Pinchbeck, and Alan D. Radford. Social distancing impact on companion animal practice. *Veterinary Record*, 186(18):607–608, jun 2020. ISSN 2042-7670. doi: 10.1136/VR.M2271. URL https://onlinelibrary.wiley.com/doi/full/10.1136/vr.m2271.

[34] Shenghuan Sun, Travis Zack, Christopher Y.K. Williams, Madhumita Sushil, and Atul J. Butte. Topic modeling on clinical social work notes for exploring social determinants of health factors. *JAMIA Open*, 7, 4 2024. ISSN 25742531. doi: 10.1093/JAMIAOPEN/OOAD112. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10788143/.

[35] J. S.P. Tulloch, L. McGinley, F. Sánchez-Vizcaíno, J. M. Medlock, and A. D. Radford. The passive surveillance of ticks using companion animal electronic health records. *Epidemiology and infection*, 145(10):2020–2029, jul 2017. ISSN 1469-4409. doi: 10.1017/S0950268817000826. URL https://pubmed.ncbi.nlm.nih.

gov/28462753/.

[36] Danielle Varney, Dan O'Neill, Maeve O'Neill, David Church, Anneliese Stell, Sam Beck, Matthew J. Smalley, and David Brodbelt. Epidemiology of mammary tumours in bitches under veterinary care in the UK in 2016. *Veterinary Record*, page e3054, may 2023. ISSN 2042-7670. doi: 10.1002/VETR.3054. URL https://bvajournals.onlinelibrary.wiley.com/doi/10.1002/vetr.3054.

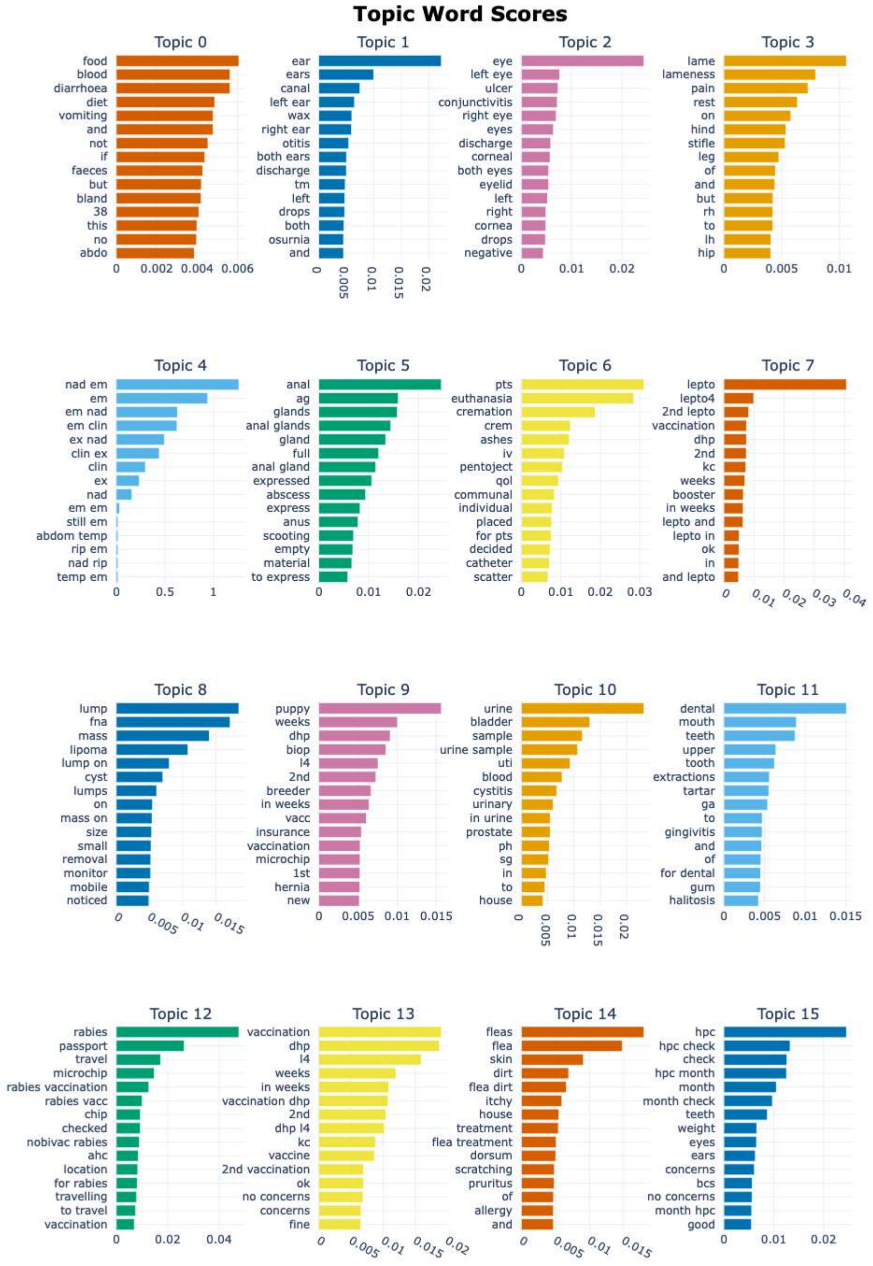

Fig. A1 Topic representations inferred from clinical records, each plot shows the topic number and the 15 most important words contributing to that topic along with the probability weighting for each word in the topic. Word patterns are often explanatory e.g. fleas, flea, skin, dirt, itchy reflecting skin disease in presence of fleas; ear,canal, wax, otitis reflecting ear disease.