Preventive Veterinary Medicine xxx (xxxx) xxx-xxx

Contents lists available at ScienceDirect

Preventive Veterinary Medicine

journal homepage: www.elsevier.com

Validation of text-mining and content analysis techniques using data collected from veterinary practice management software systems in the UK

Julie S. Jones-Diette [a, *], Rachel S. Dean [a, 1], Malcolm Cobb [b], Marnie L. Brennan [a]

[a] Centre for Evidence-based Veterinary Medicine, School of Veterinary Medicine & Science, University of Nottingham, Sutton Bonington Campus, LE12 5RD, UK
[b] School of Veterinary Medicine & Science, University of Nottingham, Sutton Bonington Campus, LE12 5RD, UK

ARTICLE INFO

Keywords:
Text mining
Content analysis
Veterinary practice
Practice based research

ABSTRACT

Electronic patient records from practice management software systems have been used extensively in medicine for the investigation of clinical problems leading to the creation of decision support frameworks. To date, technologies that have been utilised for this purpose such as text mining and content analysis have not been employed significantly in veterinary medicine.

The aim of this research was to pilot the use of content analysis and text-mining software for the synthesis and analysis of information extracted from veterinary electronic patient records. The purpose of the work was to be able to validate this approach for future employment across a number of practices for the purposes of practice based research. The approach utilised content analysis (Prosuite) and text mining (WordStat) software to aggregate the extracted text. Text mining tools such as Keyword in Context (KWIC) and Keyword Retrieval (KR) were employed to identify specific occurrences of data across the records. Two different datasets were interrogated, a bespoke test dataset that had been set up specifically for the purpose of the research, and a functioning veterinary clinic dataset that had been extracted from one veterinary practice.

Across both datasets, the KWIC analysis was found to have a high level of accuracy with the search resulting in a sensitivity of between 85.3–100%, a specificity of between 99.1–99.7%, a positive predictive value between 93.5–95.8% and a negative predictive value between 97.7–100%. The KR search, based on machine learning, was utilised for the clinic-based dataset and was found to perform slightly better than the KWIC analysis.

This study is the first to demonstrate the application of content analysis and text mining software for validation purposes across a number of different datasets for the purpose of search and recall of specific information across electronic patient records. This has not been demonstrated previously for small animal veterinary epidemiological research for the purposes of large scale analysis for practice-based research. Extension of this work to investigate more complex diseases across larger populations is required to fully explore the use of this approach in veterinary practice.

1. Introduction

A patient record is the product of a consultation between a patient and a healthcare provider. In the case of veterinary care the information is generated from an interaction between an animal, an owner and a veterinary surgeon and is recorded to document this encounter. There are, however, other uses for this type of information including decision support, epidemiology research and quality assurance. Patient information focused on the health issues discussed that are held within an electronic patient record (EPR) has been shown to be a unique source of data for human population based research (Kane et al., 2017; Stewart et al., 2017; Hersh, 2003). The aggregation of patient data is also of great value to veterinary practice-based research to examine population and environmental influences on health including disease prevalence and incidence within the vet visiting pet population (Lund et al., 1999; Faunt et al., 2007; Radford et al., 2010a,b, Garcia-Constantino et al., 2012; Jones et al., 2014; O'Neill et al., 2014, 2015; O'Neill et al., 2017; Anholt et al., 2013, 2014). Data of this type can however be difficult to access (Hripcsak et al., 1995) with much of the

* Corresponding author. Present address: Medical Detection Dogs, 3 Millfield Greenway Business Park, Winslow Road, Great Horwood, Milton Keynes, MK17 0NP, UK.
Email addresses: Jones-Diette@outlook.com (J.S. Jones-Diette); rachel.dean@vetpartners.co.uk (R.S. Dean); marnie.brennan@nottingham.ac.uk (M.L. Brennan)
1 Present address: VetPartners Ltd, Leeman House, Station Business Park, Holgate Park Drive, York YO26 4GB, UK.

https://doi.org/10.1016/j.prevetmed.2019.02.015
Received 30 July 2018; Received in revised form 26 February 2019; Accepted 27 February 2019
Available online xxx
0167-5877/

information captured in veterinary practice in the form of clinical narrative or 'free text'. It therefore makes data analysis challenging.

Clinical notes are usually written or dictated quickly as time during the consultation is short. Often this information is referred to as "locked" in the clinical narrative due to the difficulties of identifying the relevant data from the free text (Hripcsak et al., 1995). The information may be in a shorthand style and may include acronyms, misspellings or grammatical errors. An alternative to the use of free text analysis is the use of coded information, such as the read codes or ICD codes used across primary and secondary healthcare in the UK (Holt et al., 2008; Hippisley-Cox and Stables, 2011). Despite efforts by some research groups to promote a veterinary clinical coding system (VeNom coding group, 2018), the current lack of an agreed standardised language that is to be used in veterinary practice means this is currently not an option for individuals in veterinary medicine. Furthermore, research from the human health field has suggested the coding of clinical information is often too simplified to provide a rich and complex representation of both patient and disease needed for research (Jollis et al., 1993). It has been reported that there is often disagreement between the data in coded portions and free-text portions of the human medical record (Stein et al., 2000) and in the veterinary context (Robinson et al., 2015). These examples highlight some of the challenges involved with the use of routinely captured patient data for research (Benchimol et al., 2015).

The use of emerging informatics techniques such as automated content analysis or text mining may offer an answer to the problems of interpreting this type of information and some veterinary researchers have applied such techniques in an attempt to unlock the free text collected during the clinician-patient consultation (Lam et al., 2007; Garcia-Constantino et al., 2012; Anholt et al., 2013). Content Analysis software is often employed for the rapid analysis of the contents of records where the information contained is textual information, rather than numeric. Content analysis usually describes an automated systematic search and processing of the textual content of a large number of records (Lam et al., 2007); in this case the consultation notes captured in an electronic patient record (EPR) and extracted from a veterinary practice management system (PMS). Content analysis software can be used to rapidly scan and accurate categorise a large data set of patient records into a smaller subsets such as just consultations involving a feline patient with a diagnosis of a lower urinary tract infection. The process is similar to a simple word search in a word document but the search process can search across hundreds of separate documents simultaneously. Text Mining describes a more detailed second stage of analysis, after categorisation, to identify a more focussed subset such as cats who have had a vaccination using specific words or phrases (vaccination or booster) to search within the record to pick out patients with similar clinical signs, to create a patient cohort for further investigation. The usefulness of the analytical method is in the speed at which the software can scan hundreds and thousands of consultation notes within seconds to perform this categorisation, much faster than human ability and with greater accuracy, greatly reducing the man-hours needed to run the analysis.

Content analysis is often combined with text mining as a process of extracting knowledge from large datasets containing multiple records or documents of information simultaneously such as survey responses, interview transcripts or clinical narrative for analysis of the unstructured text (Shortcliffe and Blois, 2003; Hersh, 2003; Chen et al., 2005; Meystre et al., 2008). The value and quality of information extracted from veterinary clinical records for research has yet to be fully validated (Jones-Diette et al., 2017). Nevertheless this type of methodology including large scale processing of information, text mining and content analysis can assist with the rapid identification and categorisation of patient records and can provide a rapid and systematic synthesis of information from these records for identifying patient cohorts. This suggests that methods often used in medical informatics could provide a powerful tool for evidence-based veterinary medicine (Lam et al., 2007; Duz et al., 2017).

This study describes the investigation and validation of a method of automated processing of data extraction, content analysis and text mining techniques applied to a veterinary EPR and PMS system for the purpose of evidence-based research.

The aims of the study were;

(i) To pilot the use of a content analysis software program to categorise and synthesise the information extracted from a test veterinary patient record system.
(ii) To validate the use of a content analysis software program to simultaneously categorise and synthesise the information from multiple records extracted from a large database of veterinary patient records within a veterinary clinic system.
(iii) To validate the use of text mining tools to rapidly identify a patient cohort of vaccinated animals from a large database of patient consultation records extracted over an 8 week period from a veterinary clinic system.

2. Materials and methods

In this study we investigated the suitability of the text mining function of an analytical software package for the identification and separation of words or phrases from within the free text (clinical notes) portion of a test and a clinic based PMS system and veterinary EPR. The purpose was to quickly identify a small number of patients where a clinical condition common to those individuals was recorded in the EPR from within a much larger extracted dataset of many different patient records, to create a subet or patient cohort. The software search and retrieve facility was tested against a manual search of the same dataset by MRCVS registered veterinary volunteers in duplicate, termed 'gold standard' for the purposes of this study. The performance of a content analysis software platform (Prosuite[2]) was investigated. The software included a content analysis and text mining program (WordStat) to aggregate the extracted text and text mining tools to identify specific information. The diagram in Fig. 1, presents a summary of the order of data analysis.

2.1. Data extraction

The method of data extraction for automated content analysis is explained in full in a previous publication (Jones-Diette et al., 2016). Briefly, this was achieved using a bespoke XML schema (Clinical Evidence XML Schema generated by the XML consortium facilitated by the Society for Practicing Veterinary Surgeons) integrated into the clinic's PMS (www.vet-one.com) by the software provider (Gemhadar software Ltd). The XML schema was designed by the author team (JJD, RD, MB) with the assistance of the PMS system developer (Ken Coates, MD Gemhadar Ltd) to extract the data within selected fields of the patient record, such as animal ID number and date of birth. Twenty one fields of patient information were selected for extraction from each of the patient records (Table 1). Clients visiting the clinic were provided with literature describing the work and the option to opt out of the study should they wish; no owners opted out across the course of the study.

A single PMS was utilised for the research (Vet-One Veterinary Practice Management Software)[3] and two separate datasets of patient information were investigated;

[2] https://provalisresearch.com/products/prosuite/.
[3] ©2017 Gemhadar software Ltd

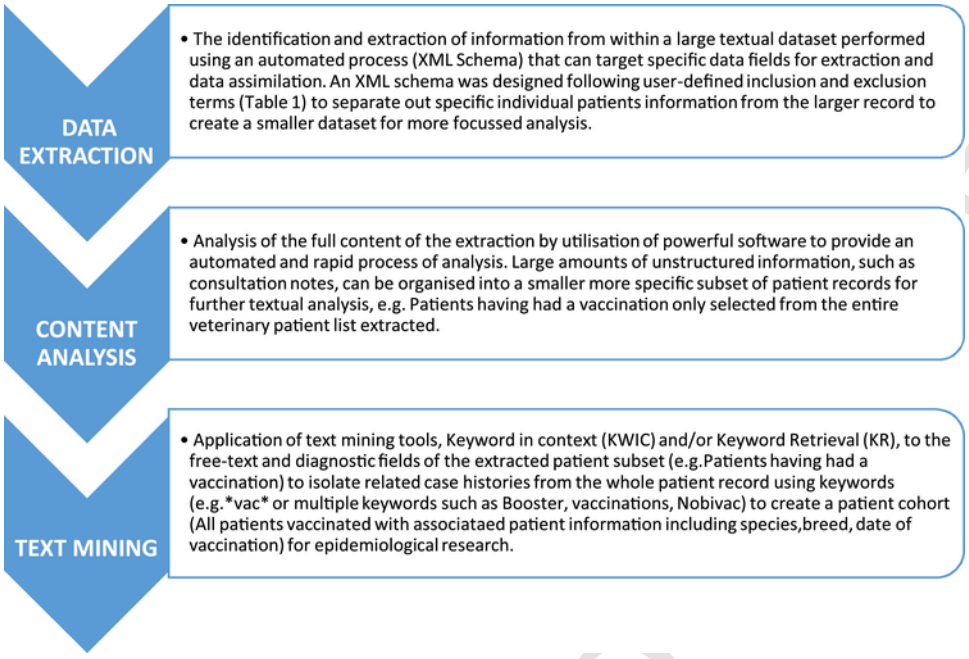

Fig. 1. A diagram summarising the order of data analysis for electronic patient record extraction.

Table 1
Clinical Evidence XML Schema. The 21 fields of patient data selected for extraction.

Data	Field description
Animal fields	1. Practice ID (numerical)
	2. Animal ID (numerical)
	3. Species
	4. Breed
	5. Gender and Neuter Status
	6. Notable Conditions (e.g. allergies)
	7. Remarks (e.g. aggressive)
	8. Deceased (Yes/No)
	9. Dangerous (Yes/No)
	10. Insured (Yes/No)
	11. Date of Birth
	12. Body Weight
	13. Body Weight units (e.g. kg)
	14. Last Weight Date
	15. Registration Date (at the practice)
Consultation information fields	16. Date (of entry)
	17. Time (of entry)
	18. Entered By ID (person who entered the data-numerical identification)
	19. Text Entry (free text for consultation and health notes, insurance details, test results)
	20. Diagnosis (practice specific codes or treatments (including trade name, drug name, drug dose and length of course of treatment) and prescriptions
	21. VeNom Code (from VeNom coding group)

1 A test PMS populated with mock patient data created by the primary author (JJD) for pilot analysis (from here onwards known as the 'Test system').
2 A clinically active PMS containing patient records extracted from a veterinary practice termed 'Clinic system'.

The Test system was created to pilot the functionality of the software. It was populated with mock patient information and included a number of keywords added into the Text Entry and Diagnosis field by the author (JJD) to test the ability of the software to search for and successfully identify records containing keywords as follows; 'CEVM', 'JJD', 'Vaccination' and 'Julie Jones-Diette'. The total number of records created for the test system was n = 326 and of these a subset of n = 23 records included the named keywords.

Table 2
Results of the KWIC search of extracted data from the Test system using' CEVM' as the search term within the Text Entry fields.

Manual count			
KWIC search for 'CEVM'	'CEVM' Present in record	'CEVM' Not present in record	Total
Term found	23	1	24
Term not found	0	302	302
Total	23	303	326

Sensitivity = 100% (95% CI 85.6%–100%)
Specificity = 99.7% (95% CI 98.1%–99.9%).

The Clinic system was a working veterinary practice patient record system containing the records for all small animal consultations seen during normal working hours over an 8 week trial period. The data set contained a total of 2519 records. The records contained all information ever recorded for the patient including the most recent consultation. The Clinic dataset was created by the transfer of patient data, this was actioned by the senior clinic veterinary surgeon. Once actioned the system extracted and de-identified the records and then the veterinarian sent the de-identified volunteer patient records to the research team for all daily consultations on a weekly basis over an 8 week period; the assessor (JJD) was blind to the full record content. Once extracted, the information was downloaded as an XML file and then forwarded to the assessor (JJD) for analysis.

2.2. Content analysis

To prepare the data for analysis the patient records from both systems (Test and Clinic) were extracted in full and transferred into the software for content analysis. The information held within the Animal ID field, Text Entry field and Diagnosis field were selected for the focus of the analysis. Once the data were transferred, the content analysis platform was launched. The software can search thousands of records

simultaneously. The process is similar to an automated word search function but the search can be performed on thousands of records simultaneously and completed in a matter of seconds. Word and phrase frequency analysis was performed on the information fields in the records and data found and categorised by the content analysis software using the search and retrieve function.

2.3. Text mining

Two methods of text mining were applied using the text mining software, the 'Keyword in context' (KWIC) function and 'Keyword Retrieval' (KR) function.

The KWIC search method allows detection of all instances of a single term occurring within the selected reports and identifies both the number of times the word appears (frequency), along with the context within which the word appears including the sentence structure where the word was found. For the Test system, as it was only a pilot trial, it was decided that the single term 'CEVM' would be selected for the KWIC search simply to test the system was working.

For the Clinic system a more complex keyword identifications was requested of the software. All consultations within the vet practice where an animal had received a vaccination were chosen for the KWIC validation of the clinic system due to its likely frequent occurrence and also because it presents a challenge as the terminology used may vary depending on which veterinarian completed the consultation record. The single term '*vac*', using a wildcard search method was employed, with the asterisk either side of the truncated term ensuring that all variations of the word would be identified regardless which term the veterinarian selected (e.g. vacc, vaccination, vaccines, Nobivac etc.). All records were mined for the presence of this term.

The KR search method creates a list of terms that can be identified within the search, rather than just one. The frequency analysis function creates a list of all lexicon (stand-alone) terms within the free-text, aside from the grammatical linking terms, and includes a qualitative report of how many times each term appears (frequency analysis). This function can then learn to suggest terms of interest, identified from the text, to select and create a smaller specific list to utilise in the search. Within the Clinic system the list of keywords selected were as follows: Booster, vacc, vaccination, Svaccination, Nobivac, Lambivac. The term Svaccination was suggested by the software and subsequently selected from the available list of terms as the addition of 'S' to the word was found to be a practice specific code.

The number of patient records in both the Test and Clinic systems that were identified (e.g. those containing 'CEVM' in the Test system, and those containing records pertaining to vaccination in the Clinic system) were compared to those actually present in the individual datasets via a manual search of the full printed record. The records retrieved using the text mining software were then cross referenced to that recorded within the practice management system using a print out of the practice record validated by veterinary volunteers who read through the printed records and reported the frequency of patients found (gold standard).

A 10% sample (n = 252) of the records was selected using a blinded process of randomisation using the visit ID number and utilising a Microsoft Excel[4] random number generator for manual comparison. These records consisted of textual information and were extracted from either the consultation or invoice notes records; no financial details were included. For the observational comparison the 252 visit records were printed out in full from the PMS by the practice and given to a team of 4 veterinary surgeons who volunteered to review the records manually in duplicate. The veterinary surgeons worked independently but the work was performed in duplicate with two veterinarians separately reviewing any one record. The veterinarians were blinded to the results of text mining. Upon completing the exercise the findings were compared by an independent veterinarian and no disagreements were found. Text mining assessment was performed by the primary author (JJD) blinded to the results of the observational study and then compared to the findings of the veterinarians and the comparison reported (Tables 3 and 4). Strict instructions were given to the veterinary volunteers about what was required prior to the commencement of the assessment. This was considered the gold standard method of data search and retrieval against which the content analysis software was measured.

2.4. Statistical analysis

The results from the manual and automated searches were assessed by calculating sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) using the equations found in Petrie and Watson (2006) and 95% confidence intervals (CI) using the methods for proportional values greater than 95% (Wilson, 1927).

3. Results

3.1. KWIC search – test system

Analysis found a high level of accuracy with the search using the' CEVM' term, retrieving 24 occurrences of the keyword within 23 patient records out of the total number of 326 records. One patient record contained the word' CEVM' twice, and a manual examination of the 326 patient records found there were 23 patient records containing the term' CEVM' or' cevm'. This resulted in a high sensitivity of 100%, a positive predictive value of 95.8% and a negative predictive value of 100% (Table 2).

3.2. KWIC search – clinic system

The final number of patient records correctly identified as having terms relating to vaccination by the KWIC search was 29 animals, with 5 records where terminology around vaccination was not identified (Table 3). The KWIC search using the search term '*vac*' found 31 occurrences of variations of '*vac*' within patient records. On two occasions a single patient record had two occurrences of '*vac*' within the same patient record so were counted twice resulting in two false positives (Table 4). The manual count identified 34 records where a vaccination was recorded in the patient's record or words relating to vaccination appeared in the invoice. The search identified correctly 216/218 records where terms around '*vac*' did not appear (Table 3).

Field Key: Visit ID, Animal ID, Variable describes the field where the term was found NS1_ TEXT refers to the clinical notes and NS1_DIAGNO refers to the clinical code recorded within the diagnosis field.

Table 3
Results of the KWIC search of extracted data from the Clinic system using *vac* as the search term.

Manual count (Gold Standard)			
KWIC search *vac*	'*vac*' Present in record	'*vac*' Not present in record	Total
Term found	29	2	31
Term not found	5	216	221
Total	34	218	252

Sensitivity = 85.3% (95% CI 69.8%–93.5%).
Specificity = 99.1% (95% CI 96.7%–99.7%).

[4] Microsoft Corporation, Redmond, Washington, United States.

Table 4
Sentences of text identified by Clinic system KWIC search (n = 31) and duplicate records highlighted where the search term occurred twice (Animal ID 16666 and Animal ID 9217).

Visit ID	Animal	Variable	Clinical Notes pre-keyword	Keyword	Clinical Notes post-keyword
13664	16640	TEXT ENTRY	1 × milbernax dog 5kg + tablets –(1 from 35027); 1 table repeat 3 months 1x	Svaccination	dog booster Procyon 7
12649	7564	TEXT ENTRY	5 each × needles 21 Gauge × 5/8 each; quote each (5 from presale) 1 x 50 ml x	LAMBIVAC	50 ml BN: L007WB03 EXP:09.2013:Use as directed by weight
13038	9967	DIAGNO		vacc	
13076	15788	TEXT ENTRY	1 x well pet blood test 1 x	Svaccination	dog booster Procyon 7 Lepto/Pi
13127	16641	DIAGNO		vacc	
13270	14710	TEXT ENTRY	1 x	Svaccination	dog booster Procyon 7 Lepto/Pi
13275	13704	TEXT ENTRY	1 x	Svaccination	dog booster Procyon 7 Lepto/Pi
13338	16730	TEXT ENTRY	1 chip x TRACER CHIP ; (1 from presale) 2 x Milbemax dog 0.5 – 10 Kg tablets (2 from 35263); 2 tablets repeat 3 months 1 x	Svaccination	dog booster Procyon 7 Lepto/Pi
13370	159	DIAGNO		vacc	
13756	13905	TEXT ENTRY	1 x	SVaccination	2nd FOC
13415	12819	TEXT ENTRY		NObivac	FeLV CVRP
13879	13997	TEXT ENTRY	0.8 x	svaccination	Rabbit myx; (0.9 from PreSale)
14129	16666	TEXT ENTRY	1 x	Svaccination	Cat booster Q FeLV1 x sVaccination cat booster CRV
14129	16666	TEXT ENTRY	1 x Svaccination cat booster Q FeLV 1 x	SVaccination	cat booster CRV
14389	17570	TEXT ENTRY	1 x Advocate 40 mg; 4 KG free puppy; (1 from PreSale); On skin repeat monthly 1 x ADAPTIL COLLAR PUPPY / SMALL 45CM C66420C; (1 From PreSale) 1 x	SVaccination	dog booster Procyon 7 puppy course pay at 1st cons
14471	7495	TEXT ENTRY	1 x	Svaccination	Horse F 2nd
14484	15820	TEXT ENTRY		NObivac	A089A01/A025B01
14549	16664	TEXT ENTRY		Nobivac	PIL A025A01/A089A01
14539	17579	TEXT ENTRY	1 x Advocate 100 mg 4 – 10Kg free puppy ; (1 from PreSale) 1 x	SVaccination	dog Procyon 7 2nd
14776	12768	TEXT ENTRY	1 x	Svaccination	dog booster Procyon 7 1 x ADVOCATE SP/ON DOG LARGE 19-25KG 6 PIP 250 – (1 from KP07FRD); on skin repeat monthly 1 x MILBEMAX REMINDER 3 MONTHS; (1 from PreSale) 3 x Milbemax dog 5Kg + Tablets; (3 from 35121) Give 1 tablet as a si
14657	12922	TEXT ENTRY	1 x	Svaccination	dog booster Procyon 7 Lepto/Pi
14869	9217	TEXT ENTRY	1 x	SVaccination	Kennel cough 1 x Svaccination dog booster Procyon 7 Lepto/Pi
14869	9217	TEXT ENTRY	1 x Svaccination Kennel Cough 1 x	Svaccination	dog booster Procyon 7 Lepto/Pi
14951	11932	TEXT ENTRY	1 x	Svaccination	Horse F 3rd 1 x Svisit 1
14943	12401	TEXT ENTRY	1 x	SVaccination	dog Procyon 7 1st Amnesty
15041	17598	DIAGNO		vacc	
15043	17244	TEXT ENTRY	1 x	SVaccination	2nd FOC
15044	17245	TEXT ENTRY	1 x	SVaccination	2nd FOC
15069	16220	TEXT ENTRY		Nobivac	PILA026A01/A087A01
15075	3030	TEXT ENTRY	4 x Milbemax dog 5Kg + tablets (4 from 35639); 2 tablets per dog repeat 3 months 1 x	Svaccination	dog booster Procyon 7
15083	14911	TEXT ENTRY	1 x Milbemax reminder 3 months (1 from PreSale); 1 x Milbemax dog chewey tablets 5 Kg + (1 from PreSale); 1 tablet repeat 3 months 1 x well pet blood test 1 x	Svaccination	dog booster Procyon 7 Lepto/Pi 1 X Svisit with nurse

3.3. KR search - clinic system

The search identified 29 out of a possible 34 patient records where a vaccination had been recorded in the animals' EPR, with 5 records not detected. The 5 records that were missed were found to have no mention of the term 'vaccination' or even 'vac' in their EPR, and were only identified at the manual search stage due to a record of a vaccine batch number added by the veterinary surgeon into the animals EPR. The search also correctly identified 218 out of a possible 218 records where the animals had no mention of a vaccination in their EPR (Table 5). The results therefore were found to have a slightly higher positive

Table 5
Results of the KR search of extracted data from the Clinic system using BOOSTER, LAMBIVAC, NOBIVAC, SVACCINATION, VACCINATION and VACC as the search terms.

Manual count (Gold Standard)			
Keyword Retrieval	Vaccination terms present	Vaccination terms not present	Total
Term found	29	0	29
Term not found	5	218	223
Total	34	218	252

Sensitivity = 85.3% (95% CI 69.8%–93.5%).
Specificity = 100% (95% CI 98.2%–100%).

predictive value than the KWIC search to identify patient records with one of the keywords present.

4. Discussion

The Wordstat program was found to be an excellent resource for the content analysis and text mining of extracted EPRs in veterinary practice and resulted in the correct extraction of most records containing the sought after terms. This is the first time this system has been validated across different systems using records from small animal practice for identifying patient cohorts for veterinary practice based research and demonstrates the value of this technique for veterinary practice-based research.

4.1. Keyword in context

The KWIC search of the Test system allowed a useful test of search accuracy when applied to extracted veterinary records. The duplicate occurrence in a single sentence of the search term (vaccination) created a duplicate find of this animal. This indicates the technique has a high level of sensitivity and specificity but caution must be taken when using such data for calculating the prevalence of conditions and when selecting the denominators used for this calculation. The Clinic system offered the opportunity to test the KWIC method on real patient data which contained natural variation in terminology. The results of this analysis showed a good level of accuracy in relation to minimising false positives and to some extent, false negative results. However, this was affected somewhat by the number of cases missed owing to the recording of 5 vaccinations by batch code alone on the part of the Clinic veterinary team. The use of a batch code to record a vaccination in the EPR highlights the need for greater care in recording of information but also allows for development of the analysis method described here. For example, although the batch code was used to record the administration of the vaccine maintaining inventory records, the invoice may have indicated a vaccination was given. The invoice data was not extracted for the current project at the request of the vets in the practice. Therefore by checking both the free text and the invoice records, both forms of data could have provided even more accurate extraction results.

Prevalence and incidence estimates in clinical epidemiology have a degree of trade-off between sensitivity and specificity, and it may be considered less of a concern if false positives are identified than missing records that contain the term of interest (Fletcher and Fletcher, 2005). For the purpose of practice-based research such as this it is important true positives for disease are identified (Petrie and Watson, 2006), particularly if the purpose is to assist with treatment decisions. However, predictive values are affected by prevalence (Wong and Lim, 2011) and the low prevalence found resulted in good negative predictive values at the expense of positive predictive values. It would be beneficial to see how the techniques used here perform under conditions of high prevalence in comparison. Ultimately, the purpose of the test and how the results will be used should be the deciding factor when judging the importance of the results from this type of analysis.

4.2. Keyword retrieval

The KR technique had high search sensitivity and improved on the sensitivity achieved in the KWIC analysis. This was due in part to the option to select search terms from a list created by the WordStat program self-learning and providing a list of vaccination terms or codes used by the veterinary surgeon's themselves within the EPR.

All records for the Clinic system were selected for inclusion using a random sampling method and the randomisation selected a single patient record rather than all details of a single visit in the EPR. Each patient selected at random had a record of the clinical history termed a consultation record and a record of treatments that were charged for (excluding monetary values) but not both. It is possible that the 5 records missed during the search could have been identified if the full dataset had been used instead of a sample, or both the treatment record and consultation notes used. This is a limitation of the study design but not the methodology. A sample of records were selected for detailed analysis rather than the full 2519 records because of the large time requirement for manual verification. Identification of the performance of the text mining techniques over a larger dataset is required, and ideally using data from a large number of different practices.

The veterinary surgeons working in the practice used for this study have their own list of clinical codes which they refer to and use within their notes. However the majority of patient records (80%) were coded as 'consultation' which meant most of the clinical data was recorded within the Text Entry field. In addition, for billing purposes the KR results highlighted that the practice used a separate coding system to charge for 'services' placing an S ahead of the item to be charged (e.g. Svaccination, Sconsultation etc.). It is also common for busy vets in practice to use a type of short hand for certain terms such as dt for Diet, op for Operation, dx for diagnosis, to name a few. As a standard veterinary terminology is not currently available for use within the profession (Wilcke et al., 2000), it is difficult to anticipate the many different terms vets may use to record clinical signs or even diagnoses. The WordStat KR method was able to overcome some of these complications by providing the user with a selection of search terms, identified by the content analysis function of the software, to select from the text extracted producing a very accurate result. Although all possible cases could not be identified, a very small percentage was missed (2%), which is a better performance than found in similar work previously published (Lam et al., 2007) and is equivalent to what has been achieved more recently using records from equine patients (Duz et al., 2017).

This study suggests the automated method of disease reporting and patient record aggregation is a powerful tool for evidence-based veterinary medicine and research as hypothesised by many veterinary researchers (Lund et al., 1999; Wilcke et al., 2000; Moore et al., 2005, 2007; Johnson et al., 2011; Santamaria and Zimmerman, 2011). The software can provide large scale analysis in a very short time period and far exceeds the ability of human search and retrieve capability within the same timeframe. The results of the study suggest the program would be an excellent resource for practice-based research using many PMS systems and extracted EPRs.

5. Conclusions

This is the first time that a method of data extraction has been described that could be applied across a number of different practice management systems. Once extracted, data from various practice management systems could be assimilated and analysed utilising the method of validation described. The work presented a validation of the

extraction and analysis method to ensure the extraction of data was complete and accurate and presents tools that could be used in veterinary practice based research for the identification and extraction of patient cohorts. The methods presented here suggest that the content analysis and text mining software used, particularly the keyword retrieval function, provided a high level of precision for search and recall of patient records sharing common clinical information, which has not been demonstrated previously for small animal veterinary epidemiologic research. However, populations with more complex cases of disease with a more unpredictable terminology may provide more of a challenge and further large scale work is required to fully explore the application of this methodology in small animal practice.

Conflict of interest

None declared; Julie Jones-Diette is currently a UK recommended consultant and trainer for Provalis, although she was not retained in this capacity at the time the research was conducted.

Acknowledgements

This work was supported by an unrestricted grant (grant number RK3801) from Elanco Animal Health and The University of Nottingham as work conducted within the Centre for Evidence-based Veterinary Medicine (CEVM). The study design, analysis, interpretation of the results, decision to publish and writing of the manuscript were undertaken independently of all funders of the CEVM.

References

Anholt, R., Berezowski, J., Maclean, K.L., Russell, M., Jamal, I., Stephen, C., 2013. The application of medical informatics to the veterinary management programs at companion animal practices in Alberta, Canada: a case study. Prev. Vet. Med. 113, https://doi.org/10.1016/j.prevetmed.2013.11.005.

Anholt, R., Berezowski, J., Jamal, I., Ribble, C., Stephen, C., 2014. Mining free-text medical records for companion animal enteric syndrome surveillance. Prev. Vet. Med. 113, https://doi.org/10.1016/j.prevetmed.2014.01.017.

Benchimol, E.I., Smeeth, L., Guttmann, A., Harron, K., Moher, D., Petersen, I., Sørensen, H.T., von Elm, E., Langan, S.M., Committee, R.W., 2015. The reporting of studies conducted using Observational Routinely-collected health data (RECORD) statement. PLoS Med. 12, e1001885.

Chen, H., Fuller, S.S., Friedman, C., Hersh, W., 2005. Knowledge Management, Data Mining, and Text Mining in Medical Informatics. Med Inform.. Springer, 3–33.

Duz, M., Marshall, J.F., Parkin, T., 2017. Validation of an improved computer-assisted technique for mining free-text electronic medical records. JMIR Med. Inform. 5 (2), e17. https://doi.org/10.2196/medinform.7123.

Faunt, K., Lund, E., Novak, W., 2007. The power of practice: harnessing patient outcomes for clinical decision making. Vet. Clin. North. Amer-Small Anim. Pract. 37, 521.

Fletcher, R.H., Fletcher, S.W., 2005. Clinical Epidemiology: the Essentials, 4th ed. Lippincott Williams & Wilkins, Baltimore, USA.

Garcia-Constantino, M., Coenen, F., Noble, P.J., Radford, A., Setzkorn, C., 2012. A semi-automated approach to building text summarisation classifiers. In: Perner, P. (Ed.), Machine Learning and Data Mining in Pattern Recognition.. Springer, Berlin Heidelberg, pp. 495–509.

Hersh, W.R., 2003. Information retrieval: a health and biomedical perspective. Springer Science & Business Media.

Hippisley-Cox and Stables, 2011. J. Hippisley-Cox, D. Stables, QRESEARCH an Ethical High Quality General Practice Database for Research, The University of Nottingham and Egton Medical Informations Systems, 2011.

Holt et al., 2008 T.A. Holt, D. Stables, J. Hippisley-Cox, S. O'Hanlon, A. Majeed, Identifying undiagnosed diabetes: cross-sectional survey of 3.6 million patients' electronic records, Br. J. Gen. Pract. 58 (2008) 5.

Hripcsak, G., Friedman, C., Alderson, P.O., DuMouchel, W., Johnson, S.B., Clayton, P.D., 1995. Unlocking clinical data from narrative reports: a study of natural language processing. Ann. Int. Med. 122, 681–688.

Johnson, L.M., Ames, T.R., Jacko, J.A., Watson, L.A., 2011. The informatics imperative in veterinary medicine: collaboration across disciplines. J. Vet. Med. Educ. 38, 5–9.

Jollis, J.G., Ancukiewicz, M., DeLong, E.R., Pryor, D.B., Muhlbaier, L.H., Mark, D.B., 1993. Discordance of databases designed for claims payment versus clinical information systems: implications for outcomes research. Ann. Int. Med. 119, 844–850.

Jones, P.H., Dawson, S., Gaskell, R.M., Coyne, K.P., Tierney, A., Setzkorn, C., Radford, A., Noble, P.J., 2014. Surveillance of diarrhoea in small animal practice through the Small Animal Veterinary Surveillance Network (SAVSNET). Vet. J. 201, 412–418.

Jones-Diette, J.S., Brennan, M.L., Cobb, M., Doit, H., Dean, R.S., 2016. A method for extracting electronic patient record data from practice management software systems used in veterinary practice. BMC Vet. Res. 12, 239.

Jones-Diette, J.S., Robinson, N.J., Cobb, M., Brennan, M.L., Dean, R.S., 2017. Accuracy of the electronic patient record in a first opinion veterinary practice. Prev. Vet. Med. 148, 121–126.

Kane, R., Howell, D., Smith, A., Crouch, S., Burton, C., Roman, E., 2017. Emergency admission and survival from aggressive non-Hodgkin lymphoma: a report from the UK's population-based haematological malignancy research network. Eur. J. Cancer 78, 53–60.

Lam, K., Parkin, T., Riggs, C., Morgan, K., 2007. Use of free text clinical records in identifying syndromes and analysing health data. Vet. Rec. 161, 547–551.

Lund, E.M., Armstrong, P.J., Kirk, C.A., Kolar, L.M., Klausner, J.S., 1999. Health status and population characteristics of dogs and cats examined at private veterinary practices in the United States. J. Am. Vet. Med. Assoc. 214, 1336–1341.

Meystre, S.M., Savova, G.K., Kipper-Schuler, K.C., Hurdle, J.F., 2008. Extracting information from textual documents in the electronic health record: a review of recent research. IMIA Med. Inform. 128–144.

Moore, G.E., Guptill, L.F., Ward, M.P., Glickman, N.W., Faunt, K.K., Lewis, H.B., Glickman, L.T., 2005. Adverse events diagnosed within three days of vaccine administration in dogs. J. Amer. Vet. Med. Assoc. 227, 1102–1108.

Moore, G.E., DeSantis-Kerr, A.C., Guptill, L.F., Glickman, N.W., Lewis, H.B., Glickman, L.T., 2007. Adverse events after vaccine administration in cats: 2,560 cases (2002-2005). J. Amer. Vet. Med. Assoc. 231, 94–100.

O'Neill, D.G., Church, D.B., McGreevy, P.D., Thomson, P.C., Brodbelt, D.C., 2014. Prevalence of disorders recorded in cats attending primary-care veterinary practices in England. Vet. J. 202, 286–291.

O'Neill, D.G., Church, D.B., McGreevy, P.D., Thomson, P.C., Brodbelt, D.C., 2015. Longevity and mortality of cats attending primary care veterinary practices in England. J. Feline Med. Surg. Open Rep. 17, 125–133.

O'Neill, D.G., Riddell, A., Church, D.B., Owen, L., Brodbelt, D.C., Hall, J.L., 2017. Urinary incontinence in bitches under primary veterinary care in England: prevalence and risk factors. J. Small Anim. Pract. 58, 685–693.

Petrie, A., Watson, P., 2006. Statistics for Veterinary and Animal Science. Blackwell Publishing, London.

Radford, A., Tierney, A., Coyne, K.P., Gaskell, R.M., Noble, P.J., Dawson, S., Setzkorn, C., Jones, P.H., Buchan, X.X., Newton, J.R., Bryan, J.G., 2010. Developing a network for small animal disease surveillance. Vet. Rec. 167, 472–474.

Radford, A., Noble, P.J., Coyne, K.P., Gaskell, R.M., Jones, P.H., Bryan, J.G., Setzkorn, C., Tierney, A., Dawson, S., 2010. Antibacterial prescribing patterns in small animal veterinary practice identified through SAVSNET: the Small Animal Veterinary Surveillance Network. Vet. Rec. 169, 310.

Robinson, N.J., Brennan, M.L., Cobb, M., Dean, R.S., 2015. Agreement between veterinary patient data collected from different sources. Vet. J. 205, 104–106.

Santamaria, S.L., Zimmerman, K.L., 2011. Uses of informatics to solve real world problems in veterinary medicine. J. Vet. Med. Educ. 38, 103–109.

Shortcliffe, E.H., Blois, M.S., 2003. The Computer Meets Medicine and Biology: Emergence of a Discipline. Springer-Verlag, New York.

Stein, H.D., Nadkarni, P., Erdos, J., Miller, P.L., 2000. Exploring the degree of concordance of coded and textual data in answering clinical queries from a clinical data repository. J. Am. Med. Inform. Assoc. 7, 42–54.

Stewart, D., Han, L., Doran, T., McCambridge, J., 2017. Alcohol consumption and all-cause mortality: an analysis of general practice database records for patients with long-term conditions. J. Epidemiol. Commun. Healthjech-2017-209241.

VeNoM coding group, 2008. Clinical Coding Systems for Veterinary Practice. Available at: http://www.venomcoding.org/VeNom/Welcome.html. (Accessed 7th June 2018).

Wilcke, J., Hahn, A., Case, J.T., 2000. Status of ani2mal health information standards in the United States. International Conference of Animal Health Information Specialists.

Wilson, E.B., 1927. Probable inference, the law of succession, and statistical inference. J. Comput. Graph. Stat. 22, 209–212.

Wong, H.B., Lim, G.H., 2011. Measures of diagnostic accuracy: sensitivity, specificity, PPV and NPV. Statistics 20, 316–318.