Real-time classifiers from free-text for continuous surveillance of small animal disease

Thesis submitted in accordance with the requirements
of the University of Liverpool
for the degree of

Doctor in Philosophy

By

Jenny Newman

March 2018

For my friends, two legged and four

Acknowledgement

The author would like to express gratitude to friends and colleagues for support and inspiration over the past decade, not least when I chose to embark on a PhD in a subject I'd never studied, and to all who have contributed to the collation of the SAVSNET dataset. This work was funded and made possible by the Health e-Research Centre, a hub of the Farr Institute of Health Informatics Research.

Abstract

Real-time classifiers from free-text for continuous surveillance of small animal disease

Jenny Newman

A wealth of information of epidemiological importance is held within unstructured narrative clinical records. Text mining provides computational techniques for extracting usable information from the language used to communicate between humans, including the spoken and written word. The aim of this work was to develop text-mining methodologies capable of rendering the large volume of information within veterinary clinical narratives accessible for research and surveillance purposes.

The free-text records collated within the dataset of the Small Animal Veterinary Surveillance Network formed the development material and target of this work. The efficacy of pre-existent clinician-assigned coding applied to the dataset was evaluated and the nature of notation and vocabulary used in documenting consultations was explored and described. Consultation records were pre-processed to improve human and software readability, and software was developed to redact incidental identifiers present within the free-text. An automated system able to classify for the presence of clinical signs, utilising only information present within the free-text record, was developed with the aim that it would facilitate timely detection of spatio-temporal trends in clinical signs.

Clinician-assigned main reason for visit coding provided a poor summary of the large quantity of information exchanged during a veterinary consultation and the nature of the coding and questionnaire triggering further obfuscated information. Delineation of the previously undocumented veterinary clinical sublanguage identified common themes and their manner of documentation, this was key to the development of programmatic methods. A rule-based classifier using logically-chosen dictionaries, sequential processing and data-masking redacted identifiers while maintaining research usability of records.

Highly sensitive and specific free-text classification was achieved by applying classifiers for individual clinical signs within a context-sensitive scaffold, this permitted or prohibited matching dependent on the clinical context in which a clinical sign was documented. The mean sensitivity achieved within an unseen test dataset was 98.17 (74.47, 99.9)% and mean specificity 99.94 (77.1, 100.0)%. When used in combination to identify animals with any of a combination of gastrointestinal clinical signs, the sensitivity achieved was 99.44% (95% CI: 98.57, 99.78)% and specificity 99.74 (95% CI: 99.62, 99.83). This work illustrates the importance, utility and promise of free-text classification of clinical records and provides a framework within which this is possible whilst respecting the confidentiality of client and clinician.

Declaration

I hereby certify that this dissertation constitutes my own product, that where the language of others is set forth, quotation marks so indicate, and that appropriate credit is given where I have used the language, ideas, expressions or writings of another.

I declare that the dissertation describes original work that has not previously been presented for the award of any other degree of any institution.

Dr Jenny Newman

Table of Contents

List of Tables

Chapter Three

Chapter Four

Chapter Five

Chapter Six

Chapter Seven

List of Figures

Chapter One

Chapter Two

Chapter Three

Chapter Four

Chapter Five

Chapter Six

Chapter Seven

Glossary

Classifier	An algorithm that assigns data to specified categories dependent on the presence or absence of given features of that data.
Clinical sign	An observable abnormality of structure or function
Concordance	The context in which a word occurs
Corpus	A large structured collection of text
De-identification	The process used to prevent an individual's identity from being connected with information. This is not a synonym of anonymous which refers to data where identifiers were not collected or were not retained and cannot be retrieved.
Diatypic variation	Within domain language variation in response to the situation it is being used in and relationships involved in the communication.
Discourse community	Individuals with a shared purpose who communicate with a specific group of words and phrases to convey information and feedback.
Domain	The topic of reference and intent of communication.
Entity	Linguistic constructs representing objects or concepts within natural language.
Homonym	A word with the same spelling or pronunciation but two or more different meanings.
Lexical diversity	The ratio of different unique words to the total number of words within a piece of text, a measure of lexical richness.
Lexis	The words and phrases of a language.

Overloading	The existence of multiple meanings for a given word (polysemy). Overloading generates homonyms.
Paraphrastic reduction	Intuitive changes to the structure of a sentence that occur by eliminating information that is redundant in context whilst maintaining its information content.
Polysemy	The existence of multiple meanings for a given word. (overloading).
Safety-netting	Strategies used by a clinician in controlling risk by advising circumstances in which a patient should be reviewed, for example when a patient with an apparently minor illness develops clinical signs of serious illness.
Semantic	The meaning of language.
Signalment	The age, breed and sex of an animal.
Syndrome	A coexistent collection of symptoms and or signs.
Syntax	The grammatical rules of a language.
Telegraphic phraseology	Information conveyed by the minimum word sequence, as a result of the omission of words occurring at high frequency.
Tokenisation	The process of deconstructing text into functional units.

Abbreviations

ANOVA	Analysis of variance (statistical test)
CDC	Center for Disease Control and Prevention
CI	Confidence interval
CRGV	Canine cutaneous and renal glomerular vasculopathy
EHR	Electronic health record
FN	False negative
FP	False positive
HL7	Health level 7 messages
ICD	International Classification of Diseases
NLP	Natural language processing
NLTK	Python's natural language toolkit
NPV	Negative predictive value
OIE	Office International des Epizooties (World Organisation for Animal Health)
ONS	Office for National Statistics
PPV	Positive predictive value
RCVS	Royal College of Veterinary Surgeons
SAVSNET	Small Animal Veterinary Surveillance Network
SNOMED	Systematized Nomenclature of Medicine
TN	True negative
TP	True positive
UMLS	Unified Medical Language System
VeNom	Veterinary nomenclature
VetCompass	Veterinary Companion Animal Surveillance System
WHO	World Health Organisation

Chapter One Introduction & review of the literature

1.1 Introduction

The work described in this thesis aimed to develop methodologies for extracting clinical features from the free-text narrative records of small animal veterinary consultations for application in syndromic surveillance (Figure 1.1.a).

Chapters one and two set the scene with an overview of the research landscape with regard pertinent aspects of surveillance, text-mining and the electronic health record, the Small Animal Veterinary Surveillance Network (SAVSNET) dataset and the Python programming language.

Chapter three evaluates the current clinician-assigned coding utilised by SAVSNET, demonstrating its limitations and thus the promise of an adaptive system capable of classifying every consultation record for the presence or absence of multiple clinical features, based on routinely documented information.

The raw material for classifier development, consultation records collated within the SAVSNET dataset, were documented for the purpose of maintaining the animals' clinical records. The language used in documenting veterinary consultations has been little studied, chapter four describes an exploration of the nature of the small animal veterinary clinical sublanguage and tools developed in the process of exploration. As would be anticipated, these consultation records contained a substantial volume of clinician and owner identifying information. Chapter five describes the methodology and efficacy of a system designed to redact information that risked compromising confidentiality within unstructured veterinary clinical narratives, whilst attempting to preserve research valuable information.

These steps paved the way for the development of methods of classifying consultations reliant on clinical and contextual features of the free-text record and extracting numeric physiological parameters where they had been documented, this methodology is described in chapter six. An alternative text-mining strategy for classification of consultations is described in chapter seven for its potential application where an emergent syndrome has been identified and the material available for identification of the language used is too sparse for extensive exploration of the context of documentation.

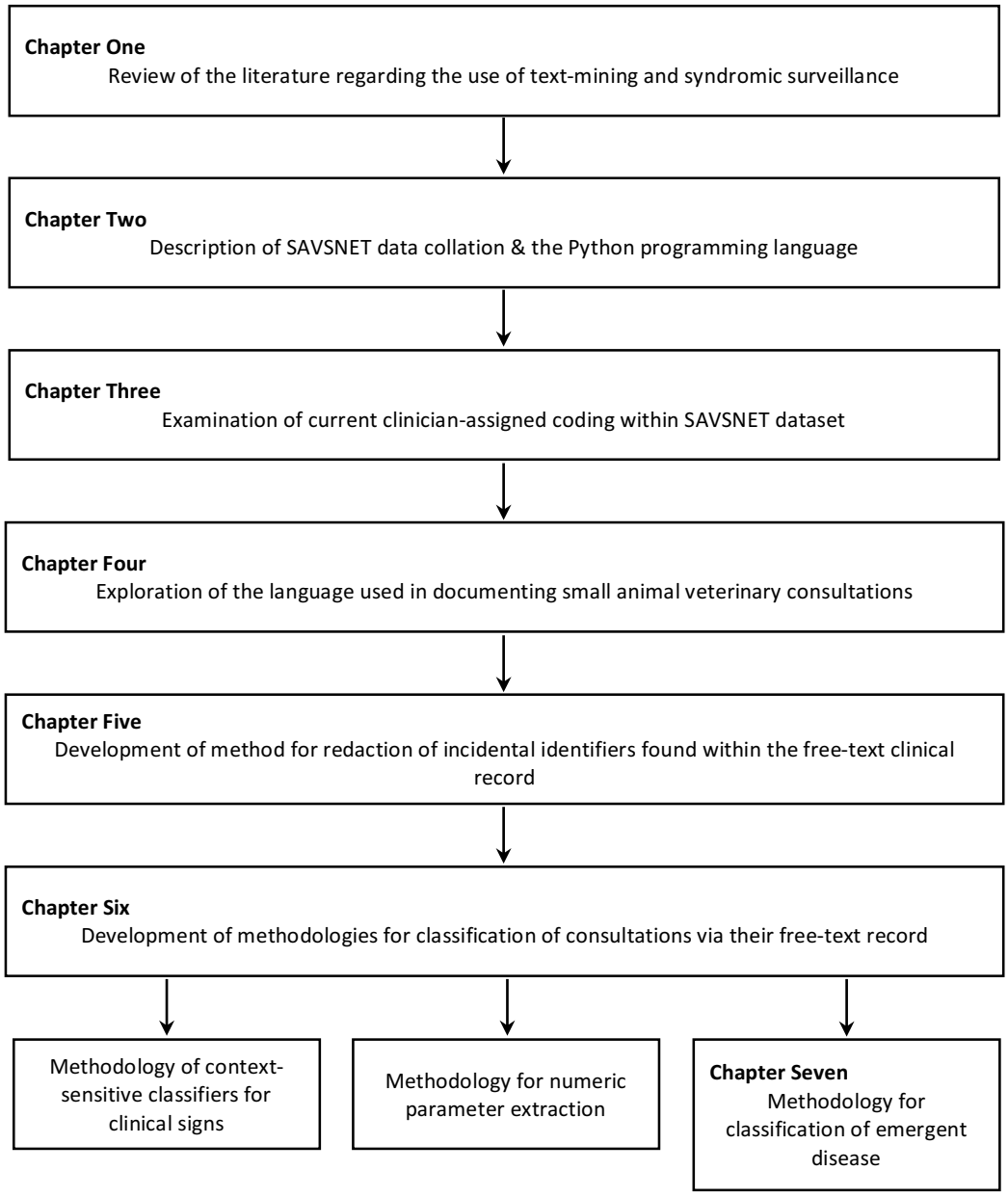

Figure 1.1.a: Outline of the work described in this thesis.

1.2 Surveillance

Surveillance, the systematic ongoing collection, collation, and analysis of data and the timely dissemination of information (Langmuir 1963), derives from the French words *veiller* (*to watch*, from the Latin *vigilare*) and *sur* (*over*) (Algeo, Barnhart, and Steinmetz 1989). The notion of collecting and analysing health data can be traced to the time of Hippocrates (460-370 B.C.) (Eylenbosch and Noah 1988). Systematic collection of mortality data began in the early 1500s, with tallies of those dying from the plague being kept in London, UK (Pearl 1930). However, data collection without dissemination for action is not surveillance, it was not until a century later that John Graunt (1620-1674), a haberdasher by trade, analysed mortality data and quantified patterns of disease, demonstrating that population data could be used to study the causative pathways of disease (Graunt and Petty 1662).

The modern concept of surveillance can be attributed to the work of William Farr (1807- 1883) (Langmuir 1976). Following establishment of the General Register Office in 1836, and introduction of universal death registration in England and Wales the following year (Galbraith 1992), Farr became its first Compiler of Abstract. Between 1838 and 1879 Farr collated and analysed statistics, disseminating his findings to both the authorities and general public, creating a fledgling surveillance system (Galbraith 1992; Thacker and Berkelman 1992; Langmuir 1976).

Surveillance may be divided into 'passive' and 'active' dependent on the manner of data collection (Teutsch and Churchill 2000). Passive surveillance involves predefined definitions and expectations of reporting, such that on seeing a clinical case meeting the case definition, a protocol is followed and the required information reported. This type of system is passive on the part of the data collator and active on the part of those seeing cases, and is thus reliant on the engagement of a third party. An early example of this was legislation introduced in 1741 in Rhode Island, United States, requiring inn keepers to report patrons exhibiting signs of contagious disease (Thacker and Berkelman 1988); more common today would be a clinician reporting disease, such as with UK notifiable disease legislation (Public Health England 2010).

In contrast, active surveillance is initiated by those maintaining the surveillance system, and is usually triggered by receipt of an indication of aberrance from the

usual pattern of disease, as for example the contact-tracing and heightened surveillance response triggered if a clinician were to report several cases of a, usually unseen within their population, notifiable or novel disease. Thus, it is a system active on the part of the data collator and passive on the part of the clinician.

The nature of surveillance required is dependent on the aims of the specific surveillance programme, the population, and the condition under surveillance. For early warning systems, the path from reporting to response must be rapid, whereas for control of an endemic disease greater finesse and confirmatory steps would be required (World Health Organization. Dept. of Epidemic and Pandemic Alert and Response 1999).

1.2.1 Syndromic surveillance

Syndromic surveillance, also known as public health surveillance, is the continuous, systematic collection, analysis and interpretation of health-related data needed for the planning, implementation, and evaluation of public health practice (WHO Global Observatory for eHealth 2006). A syndrome is a coexistent collection of symptoms and/or signs (Porta et al. 2014) that can be used for surveillance purposes without requiring laboratory data (Triple S Project 2011).

Surveillance of this nature utilises large volumes of health and other data to monitor a range of indicators for the presence of illness, in order to identify disease outbreaks and provide prompt feedback to public health and strategic bodies (CDC Evaluation Working Group on Public Health Surveillance Systems For Early Detection of Outbreaks 2004). Large integrated syndromic surveillance systems depend on multiple data sources, for example ESSENCE II integrates human behavioural and clinical data with veterinary clinical data from military and civilian sources (Lombardo et al. 2003). Syndromic surveillance is a powerful tool in protecting public health (Paterson and Durrheim 2013), facilitating early detection and response to infectious disease outbreaks and other environmental challenges (Travers et al. 2013).

1.2.1.1 Clinical data sources for syndromic surveillance

Diverse data sources are exploited in syndromic surveillance, these aim to optimise timely collation of available information, facilitating observance of signal

variation outside background levels, creating the potential for identifying emergence of illness before cases are being diagnosed with a recognised disease (Paterson and Durrheim 2013) or indeed before the emergence of a new disease is recognised (Dupuy et al. 2013). Surveillance methods are broadly divided into two groups; repurposing of clinical and behavioural data, and active reporting of observations from sentinel sites (Paterson and Durrheim 2013).

Emergency department electronic health records are extensively used in syndromic surveillance. The primary focus of Emergency Department based systems is often detection of variation in presentations of signs attributable to infectious disease, for example presentations with influenza-like-illness. Josseran et al (2006) monitored routinely recorded ICD-10 (World Health Organization 2010) coding for influenza and demonstrated the similar trends in these codes to sentinel sites and mortality data. Ansaldi et al. (2008) demonstrated the time advantage of syndromic surveillance with their system monitoring five coded syndromes alerting 2.5 days prior to established sentinel-based surveillance systems. Westheimer et al (2012) examined correlation between an algorithmic syndrome assignment based on the Emergency Department chief complaint field and laboratory-confirmed influenza and respiratory syncytial virus. Only 15% of the laboratory confirmed influenza cases were categorised as influenza-like-illness whilst 56% were categorised as fever/flu. The influenza-like-illness did however have the highest specificity at 90%.

Body temperature provides a readily measurable parameter influenced by infective illness and with a narrow normal range. In a study designed to examine the relative availability of self-reported fever and measured body temperature in Emergency Department records, a documented raised body temperature was found in 37% of those for whom self-reported fever was included in the chief complaint, conversely self-reported fever featured in the chief complaint of 60% of the records with a documented raised temperature (Kass-Hout et al. 2012). An alternative approach to identifying a population trend in body temperature used the numeric value of body temperature as recorded by data loggers at triage (Bordonaro et al. 2016). This latter method showed promise in correlating with peaks of influenza cases, and the authors suggested the data capture

method could be expanded to non-clinical settings to improve timely aberrance detection.

Other broader systems monitor emergency department presentations of multiple or any type (Wu et al. 2008; Gerbier et al. 2011; University of Pittsburgh, 2016; Ansaldi et al. 2008; Lall et al. 2017; Ziemann et al. 2014). Ambulance despatch records have been used as an alternative source of information regarding urgent care seeking in the UK (Todkill et al. 2017), Australia (Coory, Kelly, and Tippett 2009) and Europe (Ziemann et al. 2014).

The impact of other environmental factors on presentations to emergency departments has also been captured, for example meteorological extremes (Kite-Powell and Livengood 2006; Eastwood et al. 2008; Rappold et al. 2011) and large gatherings (Carrico and Goss 2005; Kajita et al. 2017; Elliot et al. 2012; Pogreba-Brown et al. 2013). Similarly, presentations with non- or indirect environmental aetiologies have been targeted, for example mental health related presentations, such as self-harm and substance misuse (Kuramoto-Crawford, Spies, and Davies-Cole 2017; Goldman-Mellor et al. 2017; Liljeqvist et al. 2014; Lall et al. 2017; Vilain et al. 2017) cardiovascular disease (Mathes, Ito, and Matte 2011) and injuries (Seil et al. 2015; Dinh et al. 2015).

Records of less urgent care have also been utilised in syndromic surveillance systems. Flamand et al. (2008) described a system for capturing the reasons for out of hours calls to a primary care service in France. This system used reason for encounter coding, the International Classification of Primary Care 2nd edition codes (ICPC-2) (World Organisation of Family Doctors 2018), and was able to capture changes in patterns of ambulatory presentations, such as respiratory and heat-related illness in real time.

In the UK, coded reasons for attendance at sentinel primary care providers are captured using Read coding (Health Social Care Information Centre 2011) and trends disseminated on a weekly basis (Public Health England 2015). In Ireland, sentinel General Practices code electronic health records for influenza-like-illness and telephone calls to out-of-hours providers are queried using free-text searches (Brabazon et al. 2010). Features captured during telephone calls to health advice lines (Baker et al. 2003; Dobson 2007; Kavanagh et al. 2012) and

sales of over-the-counter influenza remedies (Todd et al. 2014) offer a further source of information regarding health events in the community.

1.2.1.2 Social media & syndromic surveillance

Capture of social media trends provides real-time information among populations who may not have attended a health care facility, this augments information available from clinical data (Seo and Shin 2017). Little work has been published using this modality for surveillance of animal health. A recent study by Robertson and Yee. (2016) examined the use of Twitter (www.twitter.com) key word trends in relation to avian influenza, with the hope of improving situational awareness by capturing information regarding the infection in wild and domestic bird populations, which can act as a zoonotic reservoir. Peaks in activity identified by their system were related to real-world events, however the limitations created by less than 1 in 20 tweets being geocoded and the non-uniform use of Twitter internationally were evident in their findings.

Surveillance of influenza and influenza-like-illness via Twitter has been explored extensively with regard to the human population (Collier, Son, and Nguyen 2011; Kagashe, Yan, and Suheryani 2017; Sharpe et al. 2016; Deiner et al. 2016; C. Allen et al. 2016; Shin et al. 2016). Change point analysis (Taylor 2000) was used to compare trends in signals for influenza-like-illness captured via Google (www.google.com) search terms, Twitter and Wikipedia (www.wikipedia.com) page views with clinical data curated by Centers for Disease Control and Prevention (CDC) (Sharpe et al. 2016). This work found that the signal in Google searches compared best to the trend in the CDC collated data, with a sensitivity of 92% and positive predictive value of 85%. In comparison, the signal within Twitter data had a sensitivity of 50% and Wikipedia page views 33%, PPV was similarly low at 43% and 40% respectively.

Similar comparisons, using Spearman Rank correlation, for correlation between signals for conjunctival eye disease and influenza in Twitter, Google and electronic health records found that the correlation was associated with not only the data from which the signal was derived but the signal being studied (Deiner et al. 2016). In their work studying trends in data from the United States The term 'pink eye' in Google searches had better correlation with the trend in clinical data for all cause conjunctivitis (Spearman Rank coefficient (ρ) 0.68,

95% CI, 0.52 to 0.78, p < .001) than the same term in Twitter data (ρ 0.38, 95% CI, 0.16 to 0.56, p < .001). However, it was the terms 'eye allergy' (ρ 0.44, 95% CI, 0.24 to 0.60, p < .001) and 'eye drops' (ρ 0.47, 95% CI, 0.27 to 0.62; p < .001) that best correlated Google data with clinical diagnoses of allergic conjunctivitis. The correlation of the seasonal trend may have been impaired here however by comparing social media and search engine data geo-located to the United States as a whole to clinical data from California.

Surveillance via social media has also been approached in the veterinary field, Ding et al. (2015) detected and categorised information regarding medication use from veterinary discussion fora. Utilisation of domain specific linguistic and semantic features achieved a precision of at least 75% for all but one category and recall ranging from 54% to 86%.

1.2.2 Animal health surveillance

The World Organisation for Animal Health (OIE), formed in 1924 as the Office International des Epizooties, is an intergovernmental organisation with responsibility for improving animal health. Each of its 181 member countries have a responsibility to report animal disease detected within its borders, the OIE takes responsibility for dissemination regarding the geographical disease situation and current scientific evidence for disease control, to permit appropriate preventive and remedial action by member states (World Organisation For Animal Health 2018). In the UK the Animal and Plant Health Agency (APHA 2018) works on behalf of the English, Scottish and Welsh governments and is responsible for identification and control of endemic and exotic diseases in animals, plants and bees.

1.2.2.1 Production data

Within the production animal population, systems of surveillance utilise data sources integral to husbandry practices and veterinary care, but largely unavailable for monitoring of companion animal and human populations. The emergence of Bluetongue Virus in 2007 led to work exploring the statistical comparison of observed to expected milk production as an indicator of dairy herd health and an early warning of emergent disease in France (Madouasse et al. 2013) the Netherlands and Belgium (Veldhuis et al. 2016). Similarly, reproductive indicators, including rates of calving and abortion have been used

as proxy indicators of herd health. Exposure to the emergent Bluetongue Virus and its inactivated vaccine were found to be associated with decreased fertility (Marceau et al. 2014; Nusinovici et al. 2014; Nusinovici, Madouasse, and Fourichon 2016).

1.2.2.2 Mortality data

Within Europe, compulsory identification of all bovine animals and registration in national databases, in accordance with European Parliament legislation (European Parliament and of the Council 2000) has created large cattle registries, including mortality data for fallen stock. This data is hindered by its lack of timeliness, but provides a wealth of information and a reliable denominator population whose application in syndromic surveillance is being explored (Torres et al. 2015; Struchen et al. 2015; Perrin et al. 2012).

1.2.2.3 Abattoir data

Abattoir meat inspection data may also be a source of information for use in surveillance. This was investigated by a team in the UK who found that, in slaughtered pigs, correlation between the meat inspection data, laboratory systems and the targeted surveillance currently in place, was moderate for conditions with high prevalence, but poor for low prevalence conditions. The team concluded that at population level the meat inspection data may have a role as a component of a multi-faceted surveillance system, but it was not adequate for use at producer level (Correia-Gomes et al. 2016). Where an animal carcass is taken to slaughter and subsequently found to carry evidence of disease the carcass may be condemned as a whole or in part, this provides opportunity for surveillance at the point of slaughter (Alton et al. 2010; Thomas-Bachli et al. 2014; Vial and Reist 2015).

1.2.2.4 Laboratory data

Laboratory data is hindered as a data source for timely surveillance due to its poor population coverage, absence of a known denominator, and inherent delay in comparison to clinical data recorded at the time of consultation. However, laboratory output is largely constrained and digitised with centralisation, in comparison to the dispersed clinical premises, these are advantageous to its use within surveillance systems (Danan et al. 2010; Dórea et al. 2014).

1.2.2.5 Syndromic & other methods of small animal surveillance

Surveillance of animal health has typically utilised information from accumulated diagnostic reports of notifiable diseases or clinician and laboratory coded data (Wu et al. 2008; Dórea and Vial 2016; Dórea, Sanchez, and Revie 2011; Dupuy et al. 2013). This has notably more commonly involved farmed animals rather than small or companion animals, such as the dog and cat (Dórea and Vial 2016; Dupuy et al. 2013). True animal health syndromic surveillance is in its infancy (Dupuy et al. 2013). Recent initiatives have however moved forwards towards the development of robust systems to meet the demands of the International Health regulations (Ziemann et al. 2014).

Anholt et al. (2014; 2015) used a proprietary text-mining package, WordStat, to identify enteric syndrome within the electronic health records of companion animals visiting twelve veterinary practices in Calgary, Alberta. Their classifier achieved a sensitivity of 87.6% (95%CI: 80.4-92.9%) and a specificity of 99.3% (95%CI: 98.9-99.6%) and the signal it generated demonstrated statistically significant clustering of cases. The ability to identify risk factors for detected cases of enteric syndrome was hampered by insufficient available, epidemiologically relevant, information.

The Small Animal Veterinary Surveillance Network (SAVSNET), initially established as a joint venture between the University of Liverpool and the British Small Animal Veterinary Association (BSAVA), has been collating veterinary consultation records in real-time since 2008 (Radford et al. 2010; University of Liverpool 2017). This growing dataset holds a wealth of information regarding the UK's small animal population, its health and veterinary management.

The SAVSNET dataset has been used to create a quarterly surveillance report since 2015. These reports have utilised laboratory (Sánchez-Vizcaíno et al. 2016) and secondary clinical data (Sánchez-Vizcaíno et al. 2015), i.e. data captured by the attending clinician indicating the main reason for an animal's visit within a constrained selection of options, at the end of each consultation, via the SAVSNET interface within the clinic's practice management system.

The Veterinary Companion Animal Surveillance System (VetCompass) is a collaboration between the Royal Veterinary College in the UK and the University of Sydney in Australia. The focus of VetCompass is on improvement of animal

welfare via epidemiological study to identify prevalence data and risk factors for companion animal disease (VetCompass 2017; McGreevy et al. 2017). Consultations within the VetCompass dataset are associated with VeNom (Brodbelt 2012) classification tags, these have been used extensively in epidemiological studies (see section 1.3.6) and have the potential for use in syndromic surveillance.

An initiative for surveillance of disease in dogs and cats in the Veneto region of Italy, SVETPET was established in 2015. This system utilises a web interface for active online data entry by veterinary surgeons, the information captured includes the individual animal's identification, signalment and husbandry in addition to elements of their clinical history and diagnoses encoded in a standardised nomenclature adapted from the International Classification of Diseases (ICD 10) (World Health Organization 2010; Martini et al. 2017)

The electronic heath records of Banfield veterinary hospitals in the US were used as proof of principle for a syndromic surveillance system utilising coded fields (Kass et al. 2016). Following workshops with academic experts, syndromic components were selected on the basis of their recognition in the event of foodborne disease outbreaks. The signs monitored were anorexia, diarrhoea, lethargy, seizures, urolithiasis and vomiting, and laboratory findings; elevated serum calcium, alanine aminotransferase or creatinine and Salmonella-positive faecal sample. Two simulated outbreaks fed into the system were promptly recognised by its aberrance detection algorithm. The electronic health records on which this system was based are extensively coded, for both clinical and billing purposes, as the hospitals are part of, and governed by, a large corporate veterinary care provider.

There is also a role for syndromic surveillance in the audit of healthcare provision and risks. Multi-centre prospective longitudinal studies successfully demonstrated the efficacy of syndromic surveillance in surveying for nosocomial infection in hospitalised horses (Ruple-Czerniak et al. 2014) and small animals in a critical care setting (Ruple-Czerniak et al. 2013).

1.3 Electronic Health Record

1.3.1 History of clinical records

From the times of ancient Egypt documented clinical case histories were used as aids to didactic teaching (Al-Awqati 2006), gaining prominence at the Hippocratic School a millennium later (Reiser 1991a). By the early 19th century paper medical records were being kept in the teaching hospitals of Western Europe (Hess 2010) and, with the development of statistical methodology these records had gained a further role in observational hypothesis testing (Reiser 1991b). It was not until the late 19th century however that the importance of unified clinical records for the purpose of health care itself was commonly recognised (Siegler 2010). Their primary purpose today is the recording and communication of clinical findings and management. In veterinary and human medicine the keeping of accurate, legible and comprehensive clinical records is a matter of professional conduct (Royal College of Veterinary Surgeons 2014; General Medical Council 2013).

1.3.2 Adoption of the Electronic Health Record

Recognised as an essential tool for health care since the early 1990s (Dick and Steen 1991), the electronic health record (EHR) is intended to support efficient, high-quality integrated health care, improving management and co-ordination of care (Samal et al. 2011). Adoption of the EHR in human health care has seen considerable investment and incentivisation over recent decades; notably in the United States the American Recovery and Reinvestment Act 2009 (House of Representatives 111th Congress 2009) highlighted inadequacy of current medical record keeping procedures and provided support for a national system of electronic health records.

In responding to a cross-sectional survey undertaken by Schoen et al. in 2009, 46% of human primary care clinicians in the United States reported that they used an electronic health record (Schoen et al. 2009). In the ten developed countries studied at that time only Canadian clinicians reported a lower level of EHR use at 37%. A similar survey undertaken in 2012 found that the use of EHRs had remained stable for most of the studied countries, but increased by half in the US and Canada, to 69% and 56% respectively (Schoen et al. 2012). Even these increased 2012 figures fall considerably below uptake in the UK with 97% reporting using an EHR, Norway (98%), New Zealand (97%), the Netherlands (98%) and Australia (92%). Switzerland was not included in the

initial study but reported the lowest level of EHR use in the 2012 study at 41% of clinicians.

Although reliable estimates are not published, it is likely that the majority of pet animals in developed countries now have an electronic health record. This assertion is supported by Robinson and Hooker's finding that in the United Kingdom, in 2006, 94% of respondents to a survey of all veterinary surgeons registered with the Royal College of Veterinary Surgeons used a computer system for client records (D. Robinson and Hooker 2006).

1.3.3 Clinical advantages of the electronic health record

At the patient level, electronic health records facilitate care that is safer, more responsive to patient needs, and more efficient (OECD 2010). Electronic prescribing and decision support within the electronic health record are associated with a reduction in medication errors (Sidorov 2006). Studies have however failed to consistently show improved quality of care with EHRs in chronic disease management (Baer et al. 2013).

Digital records, within a well-designed infrastructure, increase accessibility to information between clinicians, with legible digital records available across specialties within an organisation. The EHR permits automated generation of reports describing care received from data entered elsewhere within the record. A survey of primary care clinicians and found that in the UK 38% were able to electronically exchange patient summaries and test results with doctors outside their own practice, this ranged from 14% in Canada to 55% in New Zealand (Schoen et al. 2012).

In their 2010 report, the Organisation for Economic Co-operation and Development (OECD 2010) suggested that the implementation of information and communication technologies in clinical practices can result in care that is safer, and more responsive to patients' needs and, at the same time, more efficient, however the structure of that technology is likely to be critical (Hyppönen et al. 2014).

The EHR is typically composed of a mixture of constrained and unconstrained fields. Patient descriptors and contact information are generally found within designated, although not necessarily constrained, fields. Unconstrained fields

may include structured text and unstructured natural language. Referral letters, laboratory and imaging reports, themselves often largely structured text (Meystre, Savova, Kipper-Schuler, and Hurdle 2008b), may form part of the electronic record dependent on the practice management system in use. The move away from paper notes, in favour of electronic health records, has facilitated the ready retrieval and analysis of the information held within designated fields of the clinical record.

1.3.4 Clinical challenges of the electronic health record

In their infancy there were numerous perceived barriers to the acceptance of computerised record systems (C. J. McDonald 1997; Trace et al. 1993). Low quality or incomplete record entry impairs the quality of the record at both patient and population level, it is paramount that acceptance is gained within the consultation room, ward and emergency department (Gilbert 1998; Walsh 2004).

The design of an electronic record system, and especially its user interface, is fundamental to both acceptance by the clinician and capture of the maximal and most interpretable data (Hyppönen et al. 2014; Van Ginneken 2002). Structuring is perceived to support clinical care processes and the collation of high quality data to enable the development of evidence-based best practice and epidemiological data (Hyppönen et al. 2014). However the balance of risks and benefits of capturing free-text vs structured data is delicate, with structured data requiring appropriate and intuitive structure design for the case mix of a practice and impacting on the workstyle of the clinician (Van Ginneken 2002).

1.3.5 The electronic health record in epidemiological research

Wide adoption of electronic health record systems facilitates population level research from a readily available dataset (Casey et al. 2016). Analysis of information held within coded segments facilitates linkage to other datasets, for example geographic and socio-demographic data via the postcode (Sánchez-Vizcaíno et al. 2017) permitting comparisons of disease burden within subsets of a population. Examination of management practices is also possible with the ability to identify biases within regular practice (Schrader and Lewis 2013).

The EHR lends itself well to case series, case-control and longitudinal studies. One of the challenges of cohort studies, based on the longitudinal information

captured by the EHR, is that periods of missing data may result from the patient, be they human or animal, having left the care of the treating clinician. Where this occurs, the patient is no longer under observation within the study without that being known to the study team. Alternatively, missing data may result when the patient has simply not attended their clinician during the follow up period. The latter potentially introducing confounding by the co-impact of factors on health, healthcare provision and healthcare-seeking behaviours (Tudor Hart 1971; Haroon, Barbosa, and Saunders 2011; M. C. Arcaya, Arcaya, and Subramanian 2015).

For observational studies reliant on coded data, identification of cases requires caution and the development of a validated case-selection algorithm. This is exemplified by the work of Hsu et al. who in seeking to identify cases of chronic rhinosinusitis within the EHR, found that International Classification of Disease Ninth revision (ICD9) codes used in isolation had low and disparate positive predictive values(PPV). The ICD9 code for nasal polyps (471.x) had a PPV of 85% whereas that for chronic sinusitis (473.x) had a PPV of only 34%. Case selection using a combination of information present within the EHR was able to achieve a PPV of 91% (Hsu et al. 2014).

1.3.6 Small animal epidemiological research utilising the EHR

Cross-sectional studies have been used to identify the prevalence and risk factors for disease in the vet-visiting small animal population. For example, the SAVSNET dataset was used to study factors associated with dogs and cats presenting with diarrhoea, and its management. This study used a combination of primary clinical and signalment data collected from electronic health records and a questionnaire presented within the SAVSNET interface (P. H. Jones et al. 2014).

The VetCompass dataset has been used to study gastric dilation-volvulus (O'Neill, Case, et al. 2017), hyperadrenocorticism (O'Neill, Scudder, et al. 2016), patellar luxation (O'Neill, Meeson, et al. 2016) and urinary incontinence (O'Neill, Riddell, et al. 2017) in dogs and diabetes mellitus in cats (O'Neill, Gostelow, et al. 2016) and hyperthyroidism in cats (Stephens et al. 2014). These studies utilised a combination of VeNom coded data, prescribing records and manual information extraction from the free-text record, with a nested case-control element to evaluate risk factors. Similarly, a retrospective case-control study

was used to examine temporal trends in chocolate consumption and risk factors in the vet-visiting dog population captured within the SAVSNET dataset (Noble et al. 2017).

Associations between signalment, aspects of preventative healthcare and sociodemographic factors have been described using the SAVSNET dataset (Sánchez-Vizcaíno et al. 2017). The datasets of VetCompass and SAVSNET have been examined in relation to patterns of antimicrobial use (Radford et al. 2011; Singleton et al. 2017; Burke et al. 2017; Buckland et al. 2016). Retrospective cohort studies have been used to identify demographic and spatial associations with road traffic trauma and mortality in cats (J. L. McDonald et al. 2017) and their longevity (O'Neill et al. 2015) and comparison of longevity and the ageing process in male and female dogs (Hoffman et al. 2017).

1.3.7 The clinical narrative

Narrative clinical data is generated from several sources; directly entered during a consultation, as communication between care providers or summarisation of findings following investigations. The clinical consultation is an interactive and exploratory patient-centred process (Neighbour 2004). It typically includes exchange of an array of information from presenting complaint through history and examination, differential diagnoses, excluded diagnoses, to prescriptions, referrals and safety-netting (Everitt et al. 2013).

The language used to document the consultation potentially holds within it a wealth of information, of value at both patient and population levels (Walsh 2004; Kay and Purves 1996; Greenhalgh 1999). Clinical narratives written as a result of an interaction between health care provider and patient describe the patient, husbandry of the animal and any features of the owner's social setting impacting their healthcare, or in the case of human healthcare social and environmental setting, present and past medical histories, the outcomes of investigations, pathological diagnoses and examination findings. Clinician thought processes and communication between healthcare providers are also often captured.

Narrative records created contemporaneously, in the clinic or at the bedside, are frequently ungrammatical composed of telegraphic phraseology, abbreviations and acronyms, with minimum number of words required to convey meaning.

These records are primarily intended as a record of what has occurred during the patient-healthcare provider interaction. In contrast, narrative specifically intended for communication, are likely to be purposely structured for clarity (Meystre, Savova, Kipper-Schuler, and Hurdle 2008b).

Medical natural language, that used to communicate and document clinical information between health care professionals is recognised to constitute a sublanguage (D. A. Campbell and Johnson 2001) (a sublanguage is an ancillary language with its own terms and expressions that is used by a particular group when talking about a given subject with a defined purpose). This may increase our ability to extract information from clinical text, both by human reading and computational methods, but in so doing creates difficulties in applying specifically developed computational techniques outside their clinical domain and conversely in applying tools developed for non-medical language to clinical text.

Much important information within the electronic health record is unconstrained text, either as narrative or short text fields. This poses barriers to automated interpretation of the health record at both an individual and population level. Attempts to standardize the structure and contents of medical records is seldom successful, hindered by both the diverse nature of medicine and medical establishments, compliance and workflow demands (Yli-Hietanen et al. 2009).

1.4 Natural language processing

Turing described the challenge of creating computers able to comprehend or generate natural language as the ultimate test of machine intelligence (Turing 1950). Natural language refers to language that has evolved with inherent fluctuations in vocabulary and syntax, the language used to communicate between individuals in contrast to artificial programmatic language used by computers. Natural language processing (NLP) aims to build computational models able to decipher the meaning of natural language. Text-mining is closely related to natural language processing, describing the process of discovering and extracting knowledge from unstructured data, there is much overlap and the distinction between the two terms is not clear cut. Once Information has been extracted, it can be analysed by comparable means to coded data, seeking trends and associations amongst the extracted data (H. Liu and Friedman 2004; Hearst 1999).

1.4.1 Origins of natural language processing & text-mining

The roots of natural language processing involved high profile, heavily funded, automatic translation from one natural language to another in the 1940s (K. S. Jones 1994). Using word-for-word translation, coupled with a degree of syntactic analysis, the true sophistication of natural language and the difficulty of programmatically replicating the intuitive language understanding of humans became apparent (Lenat 1995). Consequently, researchers at that time focused on tasks where the semantics were clear and could be explicitly encoded, however the strategies developed performed poorly in real-world texts.

Concurrently other teams were, out of necessity, summarising information from real-world text, complete with the idiosyncrasies, errors and colloquialisms that abound, and it is from this that text-mining, with its focus on real-world data, grew and differentiated from natural language processing (Witten 2005). Natural language processing itself subsequently developed to facilitate domain-specific algorithms enabling the type of deep processing previously beyond its scope.

Regular expressions, codified representation of character ranges and patterns used to create a search pattern, form a valuable component of today's text-mining techniques (Section 2.3 for further description). The basis of regular expressions was first explored in 1943 works by McCulloch and Pitts (McCulloch and Pitts 1943). Unlikely computer science pioneers, a logician and a neuroscientist, McCulloch and Pitts inadvertently had a major influence on modern day computer science, when they investigated how the human brain produced complex patterns via it’s neural interconnections (McCulloch and Pitts 1943).

Chomsky’s theoretical analysis of language grammars (Chomsky 1956) led, via Backus-Naur Form notation, to ‘context-free grammar’ (Chomsky 1959), these in turn formed the basis of regular expressions used in text search phraseology (Jäger and Rogers 2012). The syntax of regular expressions was defined by Kleene in 1956 (Kleene 1956). Thomson's grep (global regular expression print) utility to print all lines from a document that contain a specific sequence of characters, was the first to support regular expression functionality (Thompson 1968); this is still, some 50 years later, an integral component of Unix and Linux-based operating systems.

1.4.2 Evaluation of the efficacy of information extraction

The efficacy of information extraction by text-mining systems is dependent on the quality and breadth of free text available to the system, and the ability of classifiers to accurately extract information (Johnson and Friedman 1996). Various measures are used to quantify the efficacy of a text-mining algorithm, these are largely akin to the measures of diagnostic test reliability, although expressed and named differently on occasion (Figure 1.4.a). All measures require the use of a gold standard test for the presence of the concept being classified, in practice this may be expert opinion, manual classification or an alternative coding system.

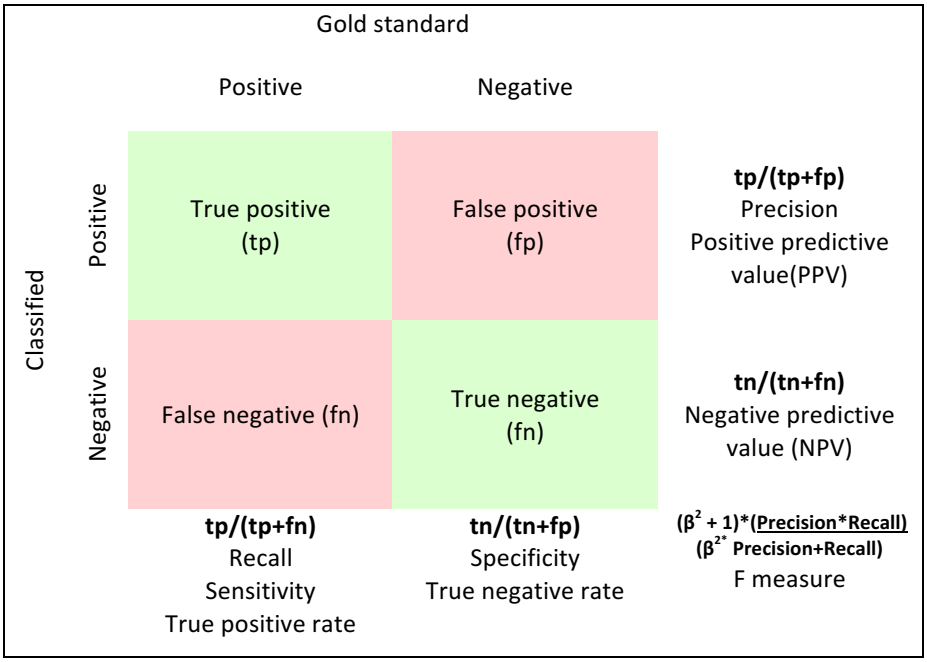

Figure 1.4.a: Relationship between classification outcomes and measures of efficacy

Precision is equivalent to positive predictive value, the proportion of test positives that are actually positive:

$$precision = \frac{true\ positives}{all\ test\ positives}$$

The prevalence of the concept and the ability of the classifier to correctly identify negative records determine precision, thus in isolation precision is not an adequate independent measure of test efficacy. The manner of calculation predicates that as the prevalence of a concept decreases the need for high specificity increases, in this way the precision is useful as a measure of what the test adds to the pre-test probability. Where multiple classifiers directed at the

same concept are compared in a population (of documents or patients) comparison of the precision of each classifier carries more meaning. Recall is equivalent to sensitivity, also known as the true positive rate, calculated as the proportion of actual positives that are correctly identified, it is thus a measure of the ability of the classifier to find the concept it is looking for:

$$recall = \frac{true\ positives}{(true\ positives\ +\ false\ negatives)}$$

Where false positives and false negatives have similar real-world cost, accuracy, the proportion of tests that are correct, may be an adequate measure of classification.

$$accuracy = \frac{(true\ positives\ +\ true\ negatives)}{(\text{total number tested})}$$

If there is an uneven distribution of the concept (rarity or commonality) the F measure is likely to provide a more useful quantification of the efficacy of classification. The F measure calculates a weighted harmonic mean of precision and recall, combining the ability to identify actual positives and actual negatives. The general equation for *F* is defined as:

$$F = \frac{(\beta^2 + 1)(\text{Precision} \times \text{Recall})}{(\beta^2 \cdot \text{Precision} + \text{Recall})}$$

A β value of 1 gives equal weight to the ability to detect the presence and absence of the concept, this is the F_1 measure:

$$F_1 = \frac{2(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})}$$

Concepts common to epidemiological study, the specificity and negative predictive value of a test are often not described when evaluating text-mining efficacy. Specificity describes the ability of a text-mining algorithm to identify where a feature is absent and negative predictive value the proportion of test negatives that are really negative:

$$specificty = \frac{true\ negatives}{(true\ negatives\ +\ false\ positives)}$$

$$negative\ predictive\ value = \frac{true\ negatives}{all\ test\ negatives}$$

The common lack of citing specificity and negative predictive value of text-mining algorithms results from validation examining the pieces of text identified by the algorithm, and determining which are true and which false positive matches, as a result it is not uncommon to find efficacy described only in terms of precision. Determination of specificity and negative predictive value requires the manual coding of a dataset, or use of a different gold standard, to identify which pieces of data have a feature and which not, and then applying the text-mining algorithm and appraising its efficacy.

1.4.3 Text-mining & the electronic health record

The wide adoption of electronic health records (Schoen et al. 2012; D. Robinson and Hooker 2006) and growing demands for high quality and evidence-based health care (Schmidt 2007; Evidence-Based Medicine Working Group 1992) and means of surveillance (Ziemann et al. 2014; Dupuy et al. 2013) are powerful drivers to the development of techniques for automated extraction of information from the narrative components of health records.

With the vast quantity of information held within electronic health record systems, attempts have been made to summarise free text fields into more readily accessible structured fields. Liu and Friedman generated a system which combined natural language processing and eXtensible Markup Language (XML) for summarisation and navigation of the growing narrative fields with the conceptualised structured information intended to enable physicians to efficiently access information across several clinical encounters of a given patient (H. Liu, Teller, and Friedman 2004). Other work has focused on creating or enhancing the clinical problem list, which forms a summary of the active clinical issues of a patient. The aim of this work was to increase completeness, accuracy and timeliness of a component of the EHR which was becoming redundant due to shortcomings in these areas (Meystre and Haug 2006; Meystre and Haug 2005). Using a finite problem list in their proof of principle work, Meystre and Haug achieved a sensitivity of 90% for detecting the problems for which it was designed, this was superior to human reading, although humans achieved greater specificity than natural language processing.

Text-mining can also be used to identify cases matching specified criteria from their narrative record. A trauma centre in Salt Lake City in the United States addressed the time consuming and labour intensive task of maintaining a trauma registry by instituting an algorithm to classify the nature of trauma patient injuries from their free-text records (Day et al. 2007). This also has a valuable role in research, with text-mining offering a mechanism for case identification from information held within free-text records. This was utilised to identify cases presenting following chocolate ingestion within the SAVSNET dataset prior to performing a retrospective case-control study (Noble et al. 2017).

Patient safety is the focus of many text-mining efforts, with adverse event detection a common topic. Wang et al. (X. Wang et al. 2009) used the Medical Language Extraction and Encoding System (MedLee) (Friedman 2012) as a knowledge map to identify co-occurrence of specific clinical event and drug entities. This adverse drug reaction extraction system achieved a recall (sensitivity) of 75% and precision (PPV) of 31%, equivalent to F_1 measure of 0.44. Rule based and hybrid decision tree and rule based systems were developed by Sohn et al. (2011) and achieved an *F* measure of 0.8 and 0.75, respectively for the rule-based and hybrid system, in identifying adverse drug reactions and the drugs implicated.

Other work has focused on the identification of vaccine adverse events (Botsis et al. 2011; Hazlehurst et al. 2005; Botsis et al. 2012; Hazlehurst, Naleway, and Mullooly 2009). Botsis and colleagues later evaluated their system's efficacy in detecting post vaccination Guillain Barre Syndrome (Botsis, Woo, and Ball 2013), they concluded that the presence in diagnostic criteria of elements not found in the clinical narrative being examined impaired sensitivity compared to a system that included investigation results. However, the system designed for vaccine adverse reactions achieved a specificity of 95% compared to 88% for an algorithm which included laboratory results.

1.4.3.1 Entity extraction

Text is replete with structured data confluent within unstructured strings, this includes non-domain specific items such as telephone numbers, street addresses and postcodes, and domain specific entities such as, in the veterinary sublanguage, microchip numbers, passport numbers and vaccine identifiers. Where documents are short and contain many structured items within

a constrained format a template can be recognised, to both the human reader and computational methods.

Entities are linguistic constructs representing objects or concepts within natural language, although a single vocabulary item an entity may be represented by a multi word phrase, as for example the concept of *torrential rain* which may be documented as a single word *downpour* or phrase *raining cats and dogs.* Dictionary-based techniques for identifying entities have a place, however as the sole mechanism for identification in real-world texts this is unlikely to be effective for many common entities which have multiple legitimate variants and may also be document in colloquial and non-standard abbreviated form.

For some entities, reliable identification can be achieved by simple pattern recognition. For example United Kingdom postcodes are restricted to a small range of specific patterns (UK Parliament 2017b), which can readily be encompassed in a sequence of characters and coding symbols, a regular expression, for example Figure 1.3.b illustrates the regular expression used to match UK postcodes taken from Table 5.3.e. Where this is not feasible, rule-based mechanisms are used, with for example recognition of the pattern [Title][Name] as one of several patterns denoting an individual's name.

```
#not preceded by an alphanumeric character
(?<!\w)

#limited range of combinations of letters followed by number(s)
((([A-PR-UWYZ][0-9])|([A-PR-UWYZ][0-9][0-9])|
A-PR-UWYZ][A-HK-Y][0-9])|([A-PR-UWYZ][A-HK-Y][0-9][0-9])|
([A-PR-UWYZ][0-9][A-Z])|([A-PR-UWYZ][A-HK-Y][0-9][ABEHMNPRVWXY]))

#one or more spaces
\s+

#a number followed by two letters from a limited range
([0-9][ABD-HJLNP-UW-Z]{2})

#or
|

#GIR 0AA the non geographic postcode previously used by Girobank
#and currently by Santander
GIR\s*0AA)
```

Figure 1.4.b: Example of a simple pattern matching regular expression, this matches UK postcodes. Purple text following the #symbol is explanatory comment.

Once identified, entities may be retrieved, utilised in abstraction, redacted or substituted for other information or placeholders. Context and syntax can aid in this process, however in many fields the expected syntactical rules are not observed, inconsistent capitalisation of names for example (Witten 2005).

1.4.3.2 Information extraction

Information extraction refers to text-mining with resultant abstraction of predefined types of data; concepts, entities and events alongside their relationships and attributes from free-text (Small and Medsker 2013). Information extraction techniques scan text for information pertinent to the subject of interest, enabling automated coding of narrative data. In the veterinary field this may be information regarding clinical signs, diagnoses, owner or clinician thoughts, drugs administered, and relationships between events (Hobbs 2002). Techniques vary from pattern recognition with regular expressions denoting the relative order of specified characters or character groups and symbols, to rule based structures and statistical machine-learning techniques, information extraction extracts structured information from unstructured free-text (Witten 2005; Appelt and Israel 1999).

Hirschman et al described one of the earliest experimental applications for information extraction in health care (Grishman and Hirschman 1978; Hirschman et al. 1981), with their progenitor to the Linguistic String Project-Medical Language Processor (LSP-MLP)(Lyman et al. 1989). Discharge summaries were parsed, recognising English grammar with subject-verb-object constructs and additional grammar peculiar to medical notations, such as dosage and telegraphic sentences, and then the distribution of words analysed to construct a template of the semantic classes present. There was for example a sign-symptoms class and a body part class, with additional modifiers such as where the sentence was a negative statement (*no back pain*), temporal relationship classes and fields to capture numeric values. The mean processing time was 22 seconds, with a mean accuracy in comparison to human reading across 33 classes of 91%, it must be noted that the evaluation only cited findings on processing three discharge summaries, and did not report sensitivity or specificity.

The Linguistic String Project work illustrates the typical routine of an information extraction algorithm, identifying the text that contains information, the

relationships between pieces of information and extracting the information into a tailored template (Witten 2005). This has typically been illustrated and examined with reference to news items or military manoeuvres with a *what who when where* type template (Grishman and Sundheim 1996), this is not dissimilar to the template needed for clinical information extraction which can be distilled to *what[sign/symptom/procedure] who[patient/other] when[absolute/relative] where[body part]*. Whereas a human extracting information from a document is likely to have access to, or prior knowledge of, other sources of corroborating evidence, computational information extraction attempts to produce a completed extraction template based on the information in an individual document (Appelt and Israel 1999). A typical information extraction pipeline would begin with entity extraction and establishing relationships between the entities extracted, in turn reliant on syntactic parsing and overcoming co-reference ambiguity (Witten 2005).

Forty years on from the early steps into information extraction from clinical documents, the current work still bears considerable resemblance to those rudimentary systems. A recent review of publications regarding clinical information extraction applications between 2009 and 2016 identified 263 relevant studies (Y. Wang et al. 2018). Of these 16.3% (n=43) used radiology reports as their data source, 9.9% (n=26) used discharged summaries, 8.4% (n=22) pathology reports 5.7% (n=15) secondary care progress notes (in or out patient) and 1.1% (n=3) primary care free-text records. The latter being the nearest human medicine equivalent to the veterinary consultation records collated within the SAVSNET dataset. Of the studies identified, 65% (n=171) utilised rule-based information extraction systems, commonly a series of regular expressions used to match patterns in specified syntactic or semantic locations within the document. The multiple rules of clinical rule-based systems are developed either via manual knowledge engineering utilising domain experts, this was the case in 45.6% (n=78) of the studies identified by Wang et al., or as in the case of 30% (n=53) of the studies, exploiting medical knowledge databases, such as the Unified Medical Language System (UMLS) Metathesaurus (National Library of Medicine 2009) or the medical terminology reference Systematized Nomenclature of Medicine Clinical terms (SNOMED CT) (National Library of Medicine 2017b). The remaining 23.4% (n=40) of the rule-based studies used a combination of the two methods for rule development.

Of the recent clinical information extraction studies utilising machine-learning methods, the most common method reported was Support Vector Machine (SVM); supervised machine-learning models that utilise a binary tagged training set of documents to build a model able to assign new examples to the correct binary category. SVM was used in 28.3% (n=26) of the machine-learning based applications, whilst logistic regression, a statistical probabilistic regression model with a binary dependent variable, for example the presence or absence of a clinical sign of interest, was used in 12% (n=11) and conditional random field (CRF), probabilistic statistical modelling methods able to account for context to a degree, in 9.8% (n=9). Decision trees, Naive Bayes classifiers and Random Forest methods were used in 8.7%, 6.5% and 4.3% of applications respectively (Y. Wang et al. 2018).

A recent study by Jackson et al. (R. G. Jackson et al. 2017) encapsulated the valuable functionality of information extraction in expansive clinical sign or symptom information extraction from free-text. The prototype study used natural language processing for information extraction to capture fifty key symptoms of severe mental illness from discharge summaries at a large UK mental healthcare provider, in turn intended to facilitate the use of the discharge summaries in epidemiological research. The team reported successfully classifying for the presence of 46 of the 50 symptoms, with a median F_1 measure of 0.88. In addition to the efficacy of the prototype system, this work highlighted that mental health symptoms were less bound to diagnoses than expected. Findings in the control group also suggested a considerable burden of mental health symptoms in patients not receiving mental healthcare. Overall this work highlighted the added value that the ability to rapidly and reliably extract information from a large volume of clinical documents confers.

1.4.3.3 Advantages of text-mining over clinical coding

The ability to assign keywords to clinical text, coding, is advantageous as it reduces the volume of stored data and standardises consultation records, improving our ability to make comparisons between patients or health care providers, facilitates quality control and assessment of compliance with practice standards or protocols, furthermore coded data is more readily accessible for research purposes (Van Der Zwaan, Sang, and De Rijke 2007; Franz et al. 2000; Ribeiro-Neto, Laender, and De Lima 2001).

Attempts to constrain free-text, for example through menu-driven data entry or pre-defined mandatory terms, especially within the specialised and complex clinical narrative, risk the introduction of errors, omission and delay (Hall and Lemoine 1986; Penz, Wilcox, and Hurdle 2007). In a healthcare system this in turn introduces risk to patients. Penz and colleagues assessed the ability of a natural language processing surveillance system to identify adverse events secondary to central venous catheter placement within the Veterans Affairs Computerized Patient Record System (VACORS). Sensitivity and specificity was 0.72 and 0.8 respectively. A key finding highlighting the importance of the narrative to the complete clinical record was the absence of coding indicative of catheter placement in 89% of cases where there was narrative evidence of placement (Penz, Wilcox, and Hurdle 2007), discrepancies in the clinical coding and information available within the narrative record was also illustrated by work by Assareh et al. (Assareh et al. 2016) which found that half (51%) of those with hypertension and a quarter (27%) of those with HIV infection were not coded to have it in subsequent admissions.

When automating coding of the electronic health record within the constraints of the limited processing power available in many healthcare settings, speed comes at the cost of sophistication (Delbecque and Zweigenbaum 2007). However, agreement between human experts in coding can be low (Van Der Zwaan, Sang, and De Rijke 2007; Hall and Lemoine 1986) and indeed much manual coding is not performed by clinical experts (Nouraei et al. 2016). Hall compared error rates in manually assigned clinical coding in two British hospitals, they concluded that much of the error introduction could be attributed to reliance on memory and failure to consult the reference manual (Hall and Lemoine 1986).

1.4.3.4 Challenges in text-mining clinical free-text

The nature of clinical text poses a number of challenges to attempts at interpretation by natural language processing. Usual rules of grammar are often disregarded in deference to telegraphic phrasing, generating ungrammatical text that does not fit the norms expected of free-text from other sources. Abbreviations and acronyms abound within the medical lexicon, many of these are non-standard local or individual shorthand, this is frequently further complicated by individual sequences having multiple meanings (H. Liu, Lussier,

and Friedman 2001). For example, in the veterinary clinical free-text the following two sentences carry the same meaning, sentence i) is standard English, which may be found in some clinical records and sentence ii) is a typically telegraphic abbreviated notation of the same as may be found in the veterinary clinical sublanguage:

i) *Re-examined Fred, he still has diarrhoea, he is not vomiting but has lost his appetite, review again tomorrow if he is not much better*

ii) Rex fred, *still d+v- a- rev tom inmb*

and when Fred has recovered, sentences iii and iv have the same meaning*:*

iii) Everything fine, Fred is now bright alert and responsive, he is drinking, eating, urinating, and defaecating normally.

iv) *aok, bar d+u+d+e+*

The taking of a medical history follows a general trend; efforts are often made to provide structure to the free-text reflecting the common facets of consultations. However where formal templates have been instituted these are commonly used idiosyncratically, negating their benefit to text-mining techniques (Hyppönen et al. 2014). Clinicians and electronic health record systems frequently insert sections of non-natural language text, laboratory results or time stamps for example, within free-text fields, thereby rendering parsing and recognition of sentence structure problematic. Over-loaded acronyms and abbreviations, misspellings, local and personal colloquialisms are commonly used within the clinical narrative, in addition to the atypical grammatical structures these pose challenges to the successful use of many natural language processing tools developed for standard text in developing classifiers of the clinical narrative (Chapman 2006). Notably, spelling correction algorithms are of limited value (Chapman et al. 2005), as described below, and tools used to examine the phrases concordant to an index word, such as the Natural Language Toolkit's concordance method (NLTK Project 2015), tend to be designed for correctly spelled real words and not abbreviated words or those containing non-word characters.

1.4.3.4.1 Misspelling

Likely as a result of the need to record a large volume of information in a limited period of time, medical free-text includes a large number of mis-spelled words,

(Shapiro 2004; Hersh and Campbell 1997) including mis-spelled acronyms. This poses a particular challenge as automated recognition, of both spelling errors and the intended word, is hindered by the lexical complexity of the medical text, despite the high rate of spelling errors, due to the atypical word frequency distribution and specialised vocabulary of clinical narrative (Turchin, Chu, and Shubina 2007; Kukich 1992; Chapman et al. 2005).

Ruch et al (2003) noted that the incidence of misspellings within clinical narrative is around 10%, considerably higher than that found in other corpora. In addition to a classical spellchecker, they used a morpho-syntactic disambiguation tool and named-entity recognition to augment the context-independent classical spell-checker in selecting the best candidate correct word. This achieved a reduction from a spelling correction error rate of over 20% to approximately 3%. Tolentino et al. (2007) successfully improved the specificity of spelling correction in surveillance reports by expanding abbreviations and creating a spelling correction tool trained on the lexicon of the UMLS metathesaurus (National Library of Medicine 2009) and WordNet (Princeton University 2018), a lexical database of standard English.

1.4.3.4.2 Sentence interpretation

The telegraphic, poorly structured, pseudo grammatical nature of language commonly found in clinical narrative poses challenges to the parsing of sentence structure. Sentence splitting and appropriate word tokenisation require understanding of the syntactic nature of the target domain (Tomanek, Wermter, and Hahn 2007), prior tokenisation risks redacting information (Witten 2005). The use of dependency grammar, where each word is permitted only a single attachment has been suggested to be superior to traditional parsing techniques (D. A. Campbell and Johnson 2002).

Where part of speech tagging is used to aid sentence interpretation, appropriately identifying the applicable part of speech is crucial to minimising error in interpretation (D. A. Campbell and Johnson 2001). Coden et al. (2005) suggest the difficulties of applying part of speech taggers trained on common usage language to clinical narrative could be overcome by enhancing the lexicon with domain specific words or annotated domain specific text. Their success was however limited because clinical text does not consist only of different words, but an atypical grammar and sentence construct, i.e. clinical

narrative has its own sublanguage (Coden et al. 2005; Pakhomov, Coden, and Chute 2006).

Liu et al. (2007) found that re-training a part of speech tagger with an annotated domain specific corpus improved performance above that achieved with the addition of domain specific words to the tagger's lexicon. They also established that 30% of words used in a corpus of pathology reports were not known to a non-domain specific part of speech tagger, highlighting the difficulty in applying generalist tools to a specialist lexicon. Conversely, a statistical tagger trained on general texts had far superior performance to a rule based system when applied to clinical text (Hahn and Wermter 2004).

1.4.3.4.3 Redundancy

Language contains significant redundancy, with many terms for the same entity. For example, the electronic health record (EHR) is also known as the Computerised Patient Record (CPR), Electronic Medical Record (EMR) and Electronic Patient Record (EPR) and refers to records created by and maintained by medical personnel; in contrast, the Personal Health Record (PHR) is an individual electronic record that the patient generally controls. Idiosyncrasies of notation in the health record were quantified by Wasserman (2011) who identified 278 different forms of documenting pyrexia in the narrative records of 465 children.

Conversely, polysemies or over-loading, terms with more than one unrelated meaning, as in the notation examples in section 1.4.3.4 above, where *d+* carried the three meanings *diarrhoea*, *drinking* and *defaecating,* introduce ambiguity, pose a classification challenge, especially where they occur with clinical meaning. Liu et al. (2001) found that a third of acronyms are overloaded, with the same character sequence representing several entities, and that their sense is frequently ambiguous even in their intended context. Although classically monosomies are considered to be more common in specialised text (Rees-Miller and Aronoff 2003), where words are overloaded they tend to have a greater number of senses in the clinical domain, with a mean of 2.6 senses per word in a dataset of words from published medical journal abstracts (Weeber, Mork, and Aronson 2001) and 5.1 senses per word in a dataset of ambiguous words generated from clinical narrative (Savova et al. 2008).

Word sense disambiguation describes the process of deciphering from a set of candidates which meaning a word is intended to have in a given lexical context; this is crucial to extracting meaning from complex narrative (H. Liu, Lussier, and Friedman 2001). Weeber et al. (2001) manually sense-tagged 5,000 instances of 50 ambiguous words within a corpus of abstracts retrieved from MEDLINE (National Library of Medicine 2016) using the UMLS Metathesaurus (National Library of Medicine 2009) as sense inventory. Liu et al. (2001) used these sense-tagged words with the New York Presbyterian Hospital Clinical Data Repository as a source of real-world clinical narrative and a wholly unsupervised approach, automatically deriving a sense tagged corpus. The derived sense-tagged corpus was then used as a training set for ambiguous word classification achieving an accuracy greater than 90% for each individual ambiguous term.

Later work by the same team investigated supervised word sense disambiguation and compared efficacy in the medical and standard English domains, concluding that supervised methods were suitable only when adequate tagged instances of each sense for a word were available. As might be expected there was a clear relationship between the achievable disambiguation precision and the number of instances available for training, achieving for example a precision of 91% for the abbreviation ASP with 141 gold standard instances available and 99% for the abbreviation APC with 2310 gold standard instances, both abbreviations had five identified senses (H. Liu, Teller, and Friedman 2004).

1.4.3.4.4 Negation

The consultation narrative is written in natural, human, language with complex juxtaposition of positive and negative information. The meaning of words or phrases can be altered, often to mean the direct opposite, by adjacent words. Contrast: ‘Fred did not have diarrhoea’, ’no d+’, ‘d+ settled’, ‘d+ not reported’ to ‘d+’. The first four phrases, in veterinary short-hand notation, indicate diarrhoea is not currently a problem, whilst the last that it is.

Regular expression based negation detection algorithms can be highly effective. NegEx for example achieved a specificity of 0.94 and sensitivity of 0.78 when developed as a negation identifier for discharge summaries (Chapman et al. 2001). When integrated into a pathology informatics network NegEx achieved a

precision of 0.84 and sensitivity (recall) of 0.8 against a gold standard corpus of negation annotations (Mitchell et al. 2004).

Further improvements in negation have been achieved by adding the Look-Ahead-Left-Recursive (LALR) parser to a system using indexed concepts from the Unified Medical Language System Metathesaurus (National Library of Medicine 2009) and regular expressions. Negfinder achieved specificity of 98% and sensitivity 95% for detecting negated terms in discharge summaries and operation notes (Mutalik, Deshpande, and Nadkarni 2001).

Cheng et al. (2017) demonstrated the importance of training context identification tools on an appropriate corpus. Although still disappointing in their efficacy when trained on the target corpus, Cheng et al. demonstrated an improvement in negation detection, in a random sample drawn from veterinary narratives collated by Vet Compass, to a precision of 89.5%, and recall 85.9% when training included oversampling of their veterinary corpus, compared to 75.2% and 63.1% respectively when trained on a human biomedical corpus. Detection of the scope of negation cues achieved a precision of 82.1% and recall 73.9%. The same work achieved a precision of 81.4% and recall of 54.3% in detecting speculation within their veterinary corpus.

1.4.3.4.5 Temporality

An understanding of the temporal relationship between events within the clinical narrative is vital to appropriate information extraction and interpretation. However identification of the order in which events occurred from the clinical narrative can pose considerable challenge (Keravnou 1996; Augusto 2005; Combi and Shahar 1997). Multiple concepts and temporal indicators within a sentence further complicates automated interpretation of the relative timing of events. Where events occur at an identifiable absolute time, temporal reasoning uses these anchors to determine relative timing of non-anchored events. Allen described thirteen mutually exclusive binary temporal relations and developed an algorithm capable of representing any interaction of these relations (J. F. Allen 1983). Allen's interval relationships have been extensively applied to medical natural language processing (Kahn, Tu, and Fagan 1991; Shahar 1997).

Within the clinical narrative relative time may not be stated explicitly and phraseology is frequently vague (Zhou et al. 2006). Within discharge summaries the majority (64%) of temporal assertions were found to be implicit, requiring domain knowledge and inference to be drawn, based on for example location within the report. (Hripcsak et al. 2005). Hyun et al. (2006) developed a system for identifying five assets of a temporal relationship; reference point, direction, number, time unit, and pattern. A similar system used by Zhou and colleagues categorised temporal expressions in the narrative of discharge summaries, they identified six main categories of temporal expressions. Their temporal-constraint-structure models temporal relations of an event by constraining its start and end points and qualitative and quantative relations to them. Using this system 97% of identified temporal expressions were effectively modelled (Zhou et al. 2006).

A system designed to extract information about relative timing of investigation from outpatient correspondence achieved precision (PPV) of 74% and recall (sensitivity) 56% (Gaizauskas et al. 2006). More recent work developed a system to infer temporal relations between events and time expressions achieved a sensitivity of 69% and precision of 70% using SVM and rules for coreferent pairs within the cTAKES-Temporal System applied to the annotated corpus previously used for the 2012 i2b2 challenge, and a sensitivity of 23% and precision of 53% using SVM within the corpus used for the 2015 Clinical TempEval challenge (Lin et al. 2016).

1.4.3.4.6 Incidental identifiers within narrative data

Electronic health records are patient-centred digital real-time chronicles of a patient's health, the healthcare they have received and communication between those involved. It is almost an inevitability therefore that the record is likely to contain information directly identifying the patient, by name, address, date of birth and indirectly by information that in combination with other data sources would allow identification for example, travel history, source of adoption, breed and age.

To facilitate the more ready availability of clinical data for research purposes de-identification of electronic health records is paramount. Where human and organisational identifiers are present within constrained fields this is readily automated. However, identifiers are found within the narrative fields, these may

belong to the person or animal to whom the records relate, their carers, health care providers, or insurer. In the interests of responsible and ethical data handling this sensitive data requires redaction in an intelligent manner that retains their contextual meaning within the narrative (Huang et al. 2010; Huang et al. 2009). With datasets commonly containing hundreds of thousands of records manual removal of identifying information becomes untenable (Dorr et al. 2006).

Even where client consent for the sharing of information has been gained, a duty to abide by data protection legislation persists. Much of this legislation is based on the guidance provided by the Organisation for Economic Co-operation and Development in 1980 (OECD 2010) including the UK Data Protection Act (UK Parliament 1998), European Union General Data Protection Regulations (European Union 2016) and US Health Insurance Portability and Accountability Act, HIPPAA (U.S. Department of Health and Human Services 1996).

Techniques used to redact identifiers within clinical free-text can largely be divided into two groups; rule-based and machine-learning, and hybrids of the two, using the same range of classification techniques as described previously to identify information likely to constitute an identifier, prior to redaction. De-identification systems are often designed to address the de-identification needs of a specific document type within a domain, within the clinical domain many of those described are validated in pathology reports (Berman 2003; Beckwith et al. 2006; Gupta, Saul, and Gilbertson 2004; Gardner and Xiong 2008; Thomas et al. 2002) or discharge summaries (Szarvas, Farkas, and Busa-Fekete 2007b; Neamatullah et al. 2008; Uzuner et al. 2008; Wellner et al. 2007). These are documents written explicitly in communication between clinical teams and would be expected to contain well-structured grammatical information, in contrast to contemporaneously documented clinical narrative.

With removal of identifiers comes the risk that an over-zealous tool would result in degradation of the quality and utility of the clinical narrative data. Although when evaluated in pre-existent systems the impact on the clinical information within the text field was thought to be small (Meystre, Ferrandez, et al. 2014); this is a potential source of reluctance to use automated de-identification when preparing healthcare records for research use. An estimated 95% of the highly structured and standardised, Health Level Seven (HL7), messages scrubbed

using the Medical De-identification System (MeDS) retained readability and were interpretable (Friedlin and McDonald 2008). Where de-identification has been thorough even the treating clinician may not recognise the identity of the patient from the de-identified narrative history (Meystre, Shen, et al. 2014).

A number of de-identification systems include research-specific data preservation techniques to minimise the risk of producing de-identified documents of poor research data quality. Such systems have commonly focused on the obfuscation of quasi-identifiers, characteristics able to be used to identify an individual from their uncommon combination so that the mean of a population remains the same whilst individual values are adjusted to reduce the likelihood of indirect identification via cumulative information (Gal et al. 2014; Sweeney 1996; Machanavajjhala et al. 2006; N. Li, Li, and Venkatasubramanian 2007; Lee et al. 2017). Tu et al. (2010) described the development of a system incorporating means of preserving the peculiarities of primary care medical records, including the use of eponymous syndromes and abbreviated forms (Gal et al. 2014; K. Tu et al. 2010). However, a system specifically designed for, or validated in, the removal of identifiers from veterinary narratives, with the preservation of clinically important features, was not identified.

1.5 Text-mining electronic health records for syndromic surveillance

Text-mining techniques can be used to rapidly and reproducibly access and begin to interpret the wealth of information recorded within the clinical narrative. Natural language processing techniques, using programmatic methods to enable computers to retrieve information from communications intended for humans, are utilised extensively to retrieve information from the biomedical literature and human health care records, but to date these techniques have not been used to their full potential in the veterinary medical field (Furrer et al. 2015; Dórea and Vial 2016; Anholt et al. 2015).

Surveillance systems that rely on diagnoses and laboratory investigation are hindered by their dependence on clinician suspicion of a diagnosis (Greene et al. 2012). Increasingly, the wealth of information within narrative clinical records is being utilised in surveillance systems (Chapman et al. 2005; Travers et al. 2013). An automated system able to classify consultation records for the presence of clinical signs, would generate signals to facilitate timely detection of

spatio-temporal trends in clinical signs (Conway, Dowling, and Chapman 2013). The quality of systems reliant on natural language processing is wholly dependent on the quality and breadth of free-text within the system (D. A. Campbell and Johnson 2001).

Whilst it undoubtedly contains a wealth of valuable information, the limitations of the narrative record must be considered when it is utilised as a data source. Following direct observation of consultations and examination of their associated electronic health record, Jones-Diette et al. (2017) found that only 64% of problems observed to be discussed during consultations were recorded in the electronic record and similarly 58% of actions observed to have been taken were documented. The documentation of actions was significantly affected ($p<0.001$) by the nature of action taken, with therapeutic or prophylactic treatment being more likely to be documented than watchful observation.

Where syndromic data can be automatically extracted the potential for real-time surveillance is more readily realised, because the need for time-consuming manual coding and reporting is avoided. Within companion animal health care a wealth of information is gathered during preventive health care visits (Shaw et al. 2008; N. J. Robinson et al. 2015), automated extraction from all routinely documented clinical narrative would ensure capture of this information with no additional demands on the practitioner. To achieve this in a responsible manner, respecting data protection regulations and the privacy of clinicians and owners, requires the prior development of a domain-specific research integrity preserving de-identification tool.

Chapter Two Materials & Methods

2.1 Data

2.1.1 The Small Animal Veterinary Surveillance Network

The Small Animal Veterinary Surveillance Network (SAVSNET), a joint venture between the University of Liverpool and the British Small Animal Veterinary Association (BSAVA), began collating veterinary consultation records in 2008. Since 2016 SAVSNET has been funded by the Biotechnology and Biological Sciences Research Council. In March 2018 SAVSNET had collated 3.4 million consultation records, regarding the health care of 1.2 million animals. This growing dataset holds a wealth of information regarding the UK's small animal population, its health and veterinary management.

There were two disjoint arms to SAVSNET's data capture, veterinary consultation related data and laboratory data, the work described here solely utilised the veterinary consultation data. Veterinary clinics were recruited directly by SAVSNET and on their agreement the clinic's practice management system (PMS) was adapted to transmit specified data in near real-time to the SAVSNET database. At the time of writing SAVSNET was able to integrate clinics where one of two PMS were being used, Robovet (Henry Schein Veterinary Solutions, Edinburgh, UK) and Teleos (Teleos Systems Limited, St Neots, UK). For reasons of data quality, only the narrative records of those consultations transmitted via the Robovet PMS have been utilised.

The SAVSNET project's ethical approval allowed for consent by opt-out. To achieve this, all participating veterinary clinics were required to display a large A3 poster describing SAVSNET in a prominent position in the waiting room, a reminder in the reception area and to have a frequently asked questions booklet available should an owner request more information about the project. Provided a client had not indicated that they did not wish to participate; at the end of a pre-booked consultation the clinician was presented with a graphical browser window and asked to assign the main reason for the consultation into one of ten heterogeneous categories. In a proportion of consultations, the main reason for presentation indicated by the clinician triggered a request for completion of a more detailed questionnaire.

The free-text narrative consultation record, clinician-assigned categorisation and, where one had been completed, questionnaire responses were transmitted

to the SAVSNET database accompanied by unique animal and consultation identifiers and additional information drawn from the PMS including the animal's signalment (age, breed, sex & neuter status), microchip number, whether they were insured, their vaccination history and the owner's registered postcode (Figure 2.1.a).

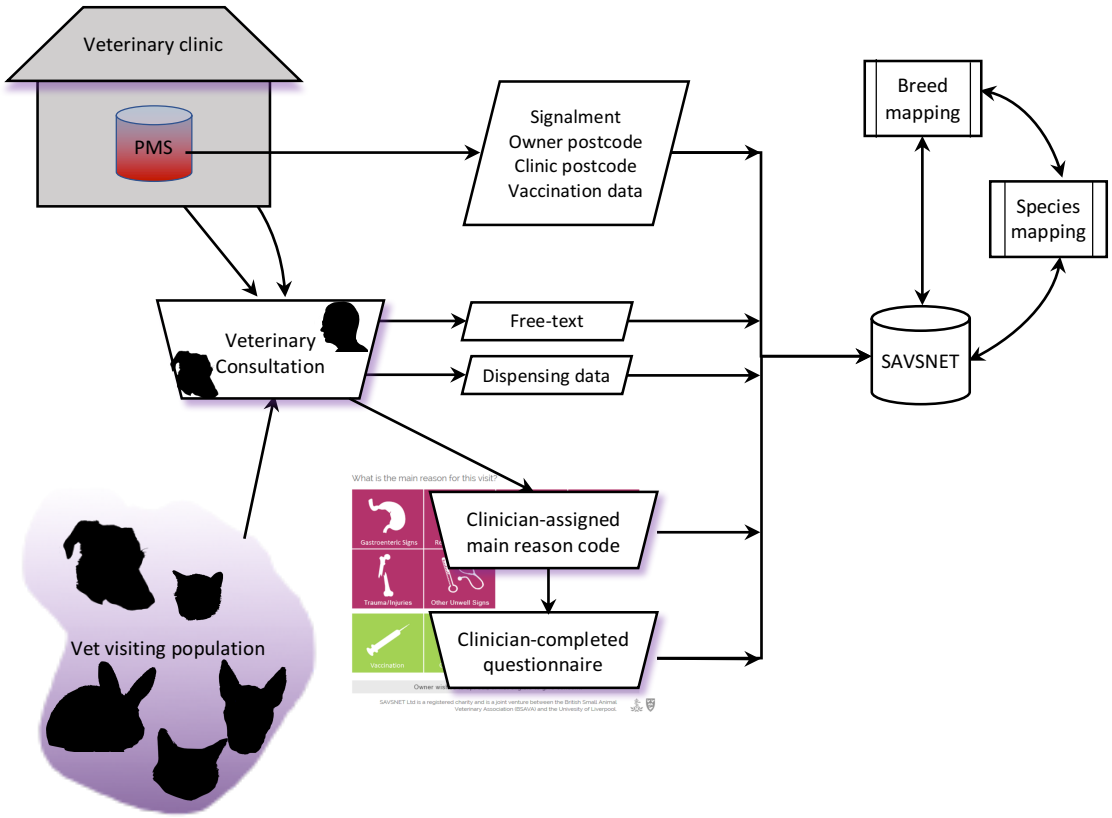

Figure 2.1.a: Infographic of SAVSNET data collation

The species and breed of each animal was stored in the PMS as two free-text fields, and as a result, there was considerable variation in the notation of both breed and species. During SAVSNET's data collation process, species were mapped to their English names (for example dog instead of canine, canis, canidae, Canis lupus familiaris). A manual mapping process assigned breeds to a recognised breed name on the first occasion that they were encountered and subsequent occurrences of the same breed name were mapped by the software. Breed maps were linked to species to improve mapping accuracy. Owing to the manner of mapping undertaken the process was not complete, at the time that this work was undertaken 6% of animals had not been assigned a mapped species or breed.

2.1.1.1 The clinical narrative field

The clinical narrative field held within the SAVSNET database consisted of a combination of the free-text clinical record, written by the attending clinician, formulaic drug dispensing labels generated by the practice management software and including instructions to the client; and in a small number of cases a structured block of text forming a clinical check list, with constrained responses and brief free-text remarks.

The text of the drug dispensing labels was not considered part of the free-text clinical narrative. A basic pattern recognition function was used to parse the dispensing labels and true narrative into separate fields. This function also cleaned white space, placing a full stop at the end of a string terminated by a line break and then converting all white space sequences to a single space. This replicated the sentence structure within the original narrative whilst removing formatting likely to hamper lexical and text analytics (Figures 2.1.b & 2.1.c). The Clancularius de-identification software described in Chapter five was used prior to human reading of narrative data.

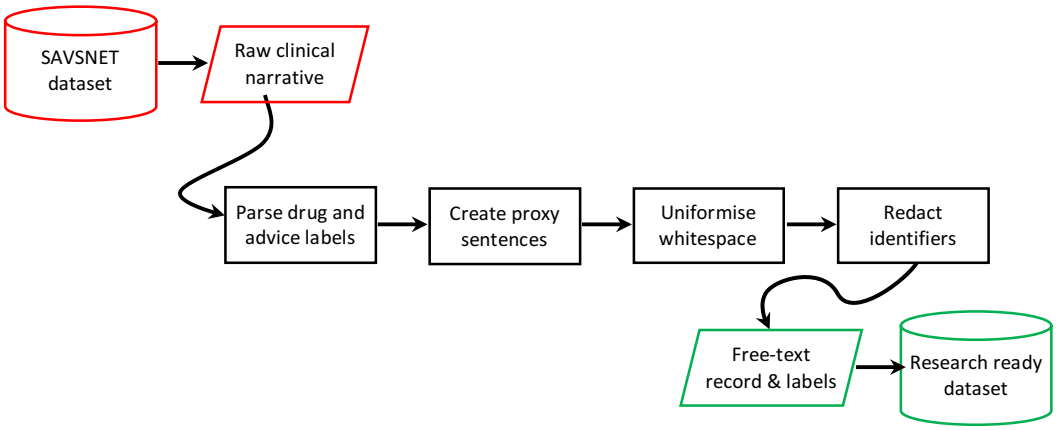

Figure 2.1-b: Pre-processing steps to produce a redacted narrative field for human reading with white space normalisation and proxy sentence creation to assist human reading and software processing.

As the data had passed through HTML, SQL and Python based systems, prior to this evaluation, and in some instances to avoid potential security issues with maliciously formulated SQL queries being embedded in form responses (SQL injection attacks), there was a degree of mis-encoding within the free-text field, with certain characters stored as HTML entities. The potentially clinically relevant HTML entities *'*, *'* *<*, *<*, *>*, *>* were changed to their respective UTF-8 characters using `pandas.Series.str.replace()`. The

codes & and & and any ampersand surrounded by whitespace were changed to the word *and*.

Raw narrative

JB. booster ducat+felv

ghc

color mm pink, ears/eyes/lnn normal, teeth ok, chest clear, heart regular and no murmer, hr136, abdomen palption nad, overweight, bcs 5

or d+ and occ v last month, now bar dude, mrs smith no concerns now

other cat, snowy has cough

a030a01/5lzx01

---------- Label ----------

Dispensed: 6 x Endectrid Cat Large >4kg

Instructions: Apply 1 pipette on the back of the neck every month Vet: Joe Bloggs

Processed narrative

<<name>>. booster ducat+felv. ghc. color mm pink, ears/eyes/lnn normal, teeth ok, chest clear, heart regular and no murmer, hr136, abdomen palption nad, overweight, bcs 5. or d+ and occ v last month, now bar dude, <<name>> no concerns now. other cat, <<name>> has cough. a030a01/5lzx01.

Dispensing label

- Label - Dispensed: 6 x Endectrid Cat Large >4kg. Instructions: Apply 1 pipette on the back of the neck every month Vet: <<name>>.

Figure 2.1.c: Example of the effect of processing the raw narrative to generate a research ready narrative field.

2.2 Software

2.2.1 The Python programming language

All exploratory, development and experimental work described in this thesis was undertaken in the Python programming language. Python is an open source, interpreted, object-oriented, programming language with dynamic semantics, it is intended to have an easy to learn syntax emphasising human readability, which along with its integral modularity aims to reduce maintenance and development costs. Because Python is an interpreted language, executing directly from the code, it has a fast edit-test-debug cycle, this does however come with processing costs as translation occurs each time the program is executed, generating a greater overhead for often executed programs than with a compiled language (Python Software Foundation 2016).

Python was first implemented in 1989 and released publically in 1991, the language was conceived by Guido van Rossum who remains principal author (Van Rossum 2009). The core philosophy of the language is summarised in the Zen of Python, which forms Python Enhancement Proposal 20 (Peters 2004), encapsulating the language's pragmatism and emphasis on clarity, making it the ideal language for a programming naive research scientist to harness.

The Zen of Python

Beautiful is better than ugly.
Explicit is better than implicit.
Simple is better than complex.
Complex is better than complicated.
Flat is better than nested.
Sparse is better than dense.
Readability counts.
Special cases aren't special enough to break the rules.
Although practicality beats purity.
Errors should never pass silently.
Unless explicitly silenced.
In the face of ambiguity, refuse the temptation to guess.
There should be one-- and preferably only one --obvious way to do it.
Although that way may not be obvious at first unless you're Dutch.
Now is better than never.
Although never is often better than *right* now.
If the implementation is hard to explain, it's a bad idea.
If the implementation is easy to explain, it may be a good idea.
Namespaces are one honking great idea -- let's do more of those!

Figure 2.2.a: The Zen of Python, PEP 20, by Tim Peters

2.2.2 Primary Python modules utilised

The SAVSNET database is a relational database managed by Microsoft SQL Server 2016 (Microsoft Corporation, Redmond, US). The database was queried via the pymssql (pymssql developers 2016) library dependent on the FreeTDS libraries (Bruns, Lowden, and Ziglio 2016), a simple database interface, using queries written in Transact Structured Query Language (T-SQL) and query results loaded directly into a Pandas dataframe (see below) using the `pd.io.sql.read_sql()` method.

Pandas (PyData Development Team 2017), the Python data analysis library, is a NumFOCUS-sponsored open source Python software library providing high-performance data handling tools. Pandas data structures comprise two value-mutable formats, one dimensional arrays of homogeneous data type, pandas.Series, and two dimensional mutable dataframes giving a tabular structure, pandas.DataFrame. Pandas is traditionally imported into the abbreviated namespace `pd` via the statement `import pandas as pd`, thus Pandas components commonly (including in this thesis) appear in code with the prefix `pd.` denoting that the class being instantiated or function called belongs to the Pandas library.

The data handling capability of Pandas was built on the n-dimensional arrays and broadcasting functions of NumPy (Numpy developers 2017), Python's fundamental package for scientific computing. Pandas functionality was augmented by directly accessing a number of NumPy functions. In the same manner as Pandas, traditionally NumPy is imported into the abbreviated namespace `np` and so reference to components of NumPy appear in code preceded by `np.`. Numpy was primarily used for its `np.where` function, which the author found more intuitive that that of Pandas.

The regular expression operations library, re, a core Python library, was utilised for text searching, matching and extraction capabilities, in addition to the string functions of Pandas which incorporate many of Python's built-in string methods and apply them across arrays. Where statistical tests and logistical regression was undertaken the SciPy (SciPy developers 2018) and StatsModels (Perktold et al. 2017) libraries were utilised, or a bespoke function written to meet the needs of the application. Data was visualised and plots generated using the two-dimensional plotting library Matplotlib (Hunter et al. 2018).

Initial work was undertaken in Python version 2.6 and associated libraries, as updated versions became available the additional functionality was utilised. Final versions of code were written in Python version 3.6.1 and utilised Pandas version 0.20.3, NumPy version 1.13.3, Matplotlib version 2.0.0, StatsModels version 0.8 and SciPy version 0.19.0.

2.2.3 Components of Python script

Variable A name given to a string or numeric value so that it can be reused, referred to, acted on and passed between pieces of code. Values are assigned to a variable with a single = sign:

```
in:    x = 1
in:    y = 'tomatoes'
```

Function A small command. A function names a piece of code in the same manner as variables give a name to a value. When defining a function, the syntax used is:

```
def functionName(arguments):
      doSomething
```

Class A class is a ring-fenced grouping of functions and data. Python is an object-oriented language, classes within Python confer structure and facilitate reusability of processes. When defining a class the following syntax is used:

```
class rhubarb(object):

    def __init__(self):
        self.yourName = input('What is your name? ')

    def printName(self):
        print('Your name is', self.yourName)
```

To instantiate a class, it is assigned to a name, for example to assign the name `x` to an instantiation of rhubarb:

```
in:    x = rhubarb()
```

`x` is now a ring-fenced object containing all of the functions and variables of the class rhubarb. On instantiation the class rhubarb's `__init__` function is initiated to provide values for the newly constructed object, in this case `__init__` asks for user input to define its `yourName` variable.

```
out:   What is your name?

in:    Fred
```

If the `printName()` method of the object we've called `x` is then called, using the command `x.printName()` the output message prints the value entered:

```
in:    x.printName()

out:   your name is Fred
```

The variable `yourName` can also be given a value directly once the object `x` has been instantiated:

```
in:    x.yourName = 'Barney'

in:    x.printName()

out:   your name is Barney
```

The same class can be instantiated multiple times within a programme without interference between the instantiations, an instance of an instantiated class is known as an object. Where `self` appears within a code snippet `self` is a variable of the instance of the class that is being used, `self.functionCalledX()` refers to the `functionCalledX()`, and likewise `self.variableCalledY` refers to the value of `variableCalledY`, within the specific instantiation of a given class. In this manner, the variables of multiple instantiations of a class can have different values. Within a class functions are referred to as methods.

2.3 Regular expressions

A regular expression is a codified representation of character ranges and patterns used to create a search pattern that will match the desired strings. The plasticity of regular expressions can be harnessed to match particular types of page layout, strings of text or specific words, as a result they lend themselves well to the tasks of text pre-processing, de-identification and classification via text-mining. Having encoded a regular expression, it can be used to identify whether the pattern is present, where in a string it is present, to extract the matching pattern from a string or to substitute it with another string.

Many programming languages feature regular expression functionality, either within their core or as adjunct libraries, with common syntax across languages but minor variation in their manner of application. All exemplar regular expressions cited within this thesis were applied within Python scripts and so

meet the constraints of the Python regular expression syntax (Python Software Foundation 2018).

Regular expressions are formed by combinations of literals, operators, constructs and quantifiers. Alphanumeric characters are represented by themselves within regular expressions, thus `cat` is a regular expression that will match the string *cat*. Character classes are represented in two ways: square brackets are used to represent a match to literals, or ranges of literals, within them, thus `[a-z]` matches any lower-case letter, there are also special character class representations: `\w` matches any alphanumeric character and the underscore, `\W` (with a capital W) matches any non-alphanumeric character. Similarly, `\s` represents any white space character whilst `\S` represents any non-whitespace character.

Adding a caret (`^`) as the first character within square brackets changes their function to match any character except those within the brackets, thus `[^aeiou]` matches any lower-case consonant. Where a regular expression includes a caret, other than as the first character within square brackets, this matches the beginning of a string, and similarly the dollar symbol (`$`) represents the end of a string.

Square brackets or the backslash can be used to create a regular expression to match a character that otherwise acts as an operator or quantifier, thus `[+]` and `\+` match the plus symbol.

Constructs are formed within parentheses, a capturing group consists of a regular expression within parentheses, thus `(\w+)` captures an alphanumeric pattern when it is found within a string. Parentheses can be used to group regular expressions in a similar manner to their application in mathematics, the non-capturing group is represented by a question mark and colon immediately inside the parentheses, thus `(?:\w+)` groups but does not capture an alphanumeric pattern. When regular expressions are applied within Pandas methods the non-capturing group avoids generating unwanted software warnings regarding capturing groups.

Quantifiers encode the number of repeats of a pattern element that should match. A question mark (`?`) encodes 0 or 1 matches to the element immediately preceding it, the asterisk (`*`) 0 or more and the plus sign (`+`) 1 or more; thus

`lo?se` will match *lose* and *lse*, `lo*se` will match *loose*, *lose* and *lse* and `lo+se` matches *lose* and *loose*, the latter two also matching spellings with an infinite number of *o*s. Quantifiers can also be encoded using the notation `{m,n}`, where m represents the minimum number of repeats and n the maximum, omiting m as in `{,n}` encodes up to n repeats, including no match and similarly omiting n, `{m,}` from m to infinity matching repeats.

By default, where a quantifier is used the maximum length of matching string available will be matched, this is termed greedy matching. Thus, for example, `the.*\.` will match from the word *the* to the last full stop in a string, whilst `the.*?\.` will match from the word *the* to the next full stop. This can be important in determining the permitted distance between matching elements within regular expressions within classifiers.

Wildcard functionality, common within many search engines, facilitates the creation of regular expressions where all characters are not defined. The period represents any character except a new line, thus `d.g` will match *dog*, dig and *d9g*. Character classes can be used as limited wildcards, permitting matching to the non-specified character only if it belongs to the class, for example `d[a-z]g` will match *dog* and *dig*, but not *d9g*.

Look around assertions add valuable functionality to regular expressions, they prohibit or mandate matching to a pattern dependent on the presence of a second pattern. Look around assertions are atomic, or have zero length, which means that they perform their function but are not themselves captured within a regular expression match. Using a positive look ahead, identified using the construct `(?=yourPattern)`, mandates that a pattern only match if it is immediately followed by the look ahead pattern, for example with the expression `pattern1(?=pattern2)` pattern1 will only match if immediately followed by pattern2. A negative look ahead, identified using the construct `(?!yourPattern),` has the opposite function, prohibiting matching if the look ahead pattern matches: `pattern1(?!pattern2).` Look behind assertions behave similarly: `(?<=pattern2)pattern1` and `(?<!pattern2)pattern1` permit and prohibit matching to pattern 1 if pattern2 does and does not immediately precede it respectively. In Python look head assertions can be variable length but look behind assertions must have a fixed character width.

The interpretation of a regular expression can be amended by specifying a number of options, known as flags. To render a regular expression case insensitive the `flags = re.I` or `re.IGNORECASE` option is used. Where regular expressions are written using white space to improve human readability, the `flags = re.X` or `re.VERBOSE` option is used, all non-escaped whitespace is then ignored in interpreting the regular expression. Multiple flags can be applied using the pipe (|) as in `flags = re.I|re.X`, this can also be used within the regular expression itself to match either the expression before the pipe or after the pipe, for example `cat|dog` matches the word *cat* or the word *dog*.

Chapter Three The need for information extraction from the free-text clinical record

3.1 Reliability of clinical coding systems

The summarisation of health care episodes into standardised clinical nomenclature is described as clinical coding. Historically, in human healthcare, patient care has been documented by attending clinicians, with care episodes initially coded by clerical staff for administrative purposes and post-discharge abstraction of information undertaken by specialists in clinical coding, with pertinent features encoded against standardised, internationally-recognised hierarchical dictionaries of terminology (Nouraei et al. 2016). Coded data has many uses, from health care funding allocation at a local and national level, through epidemiological study and health care quality audit to real time surveillance (MacIntyre et al. 1997; P. Cheng et al. 2009; Nouraei et al. 2016).

This chapter describes and evaluates the pre-existent, clinician-assigned, coding applied to the dataset of first opinion veterinary records whose narrative records formed the target, and development material, of de-identification and free-text classification methodologies described in later chapters.

3.1.1 Coding in human healthcare

With the introduction and increasing use of electronic health records, responsibility for clinical coding in human secondary care is evolving, with shift to clinician coding at the point of care delivery using SNOMED CT (National Library of Medicine 2017b) codes (Spencer 2016). Point of care clinical coding has been common in UK general practice since 1989, with the use of Read codes (Health Social Care Information Centre 2011). Weekly syndromic surveillance reports are generated from this primary care based coding (Public Health England 2015) and since 2004 the Read codes have been used to calculate a component of general practice funding allocation, via extraction of information related to Quality and Outcomes Framework indicators (NHS Employers 2016).

Considerable work has evaluated the efficacy of coding and the impact of discrepancy in coding in human healthcare, where commonly used classifications include the Read (Health Social Care Information Centre 2011) and SNOMED (National Library of Medicine 2017b) systems. Discrepancies in coding are common; an audit of discharge coding summarising emergency medical admissions amended an aspect of coding in 55% of the episodes, with primary diagnosis changed in 17% of cases (Nouraei et al. 2016). Discrepancy of the same order of

magnitude was found in an audit of Australian discharge data with a discrepancy rate of 53% overall and 22% in principle diagnosis (MacIntyre et al. 1997).

Assareh et al. (2016) found that chronic conditions were coded inconsistently, with a discrepancy incidence rate of 51% for hypertension and 26.7% for human immunodeficiency virus (HIV), i.e. over a quarter of patients noted to have HIV, a lifelong infection likely to impact management of any co-existent condition, and a half of those with hypertension, were not documented to have the respective conditions during a subsequent admission.

There are limitations to every clinical coding system but each also has its merits (J. R. Campbell et al. 1997). Many factors may influence the reliability of clinical coding. Much variation is attributable to hospital associated factors, including characteristics such as rurality but also individual undefined hospital characteristics (Assareh et al. 2016; Rangachari 2007; Santos et al. 2008; Lujic et al. 2014). The coverage, or inclusivity, of coding systems influences their ability to capture clinical information, where systems have poor coverage even assignment by specialised personnel will result in loss of a large amount of clinical information (Chute et al. 1996).

Relating case mix to funding, as for example with the UK's National Health Service primary and secondary care funding structures, introduces systematic bias of comorbidity coding (Steinbusch et al. 2007) with inadvertent financial incentive to code for the presence of diagnoses ('upcoding'). Consequently coding accuracy has been found to be influenced by payment systems (Assaf et al. 1993) and conversely appropriate payment to the institution may be impaired by inaccuracies in coding (Peeraully, Henderson, and Davies 2016).

An early study from Veteran Association hospitals in the United States found that a major source of discrepancy between documented free-text record and coding was the clinician failing to document events and the coding of resolved clinical issues as currently active (Lloyd 1985). Reliability also relates to the nature of the coding system used, even between progressive iterations of the same system (O'Malley et al. 2005).

The Clinical Practice Research Datalink (CPRD), formerly known as the General Practice Research Database (GPRD), collates human primary care information regarding approximately 6% of the UK population (National Institute for Health

Research 2018). Participating general practices provide information regarding each episode of illness, new symptom, and significant morbidity events coded with Read codes. Coding is in some circumstances recorded directly via the clinician and in others, such as transcription of secondary care diagnoses, by administrative staff. Two systematic reviews of diagnostic coding validity within this dataset found a median positive predictive value of 89% (range 24–100%).

3.1.2 Veterinary coding systems

The VeNom, veterinary nomenclature, coding system was developed in first opinion and referral settings through a drive to encourage the use of robust standardised terminology accessible to the clinician (Brodbelt 2012). The VeNom classifications consist of diagnoses, presenting complaints and administrative tasks, using terms standardised across UK veterinary institutions, to facilitate clinical audit and academic discussion. This coding system is integrated into the Veterinary Companion Animal Surveillance System (VetCompass 2017) and has been used to undertake a number of epidemiological studies, for example studying tail injuries (Diesel et al. 2010), risk factors for mast cell tumours (Shoop et al. 2015) and glucocorticoid use in veterinary primary care (O'Neill et al. 2012).

The standardised nomenclature of SNOMED CT is mirrored in its veterinary extension, VetSCT, maintained by the Veterinary Terminology Services Laboratory (VTSL) at the Virginia Maryland Regional College of Veterinary Medicine. VetSCT provides a standardized terminology for communication within the veterinary and public health communities, whilst maintaining inter-operability with SNOMED CT as applied in human healthcare (National Library of Medicine 2017a).

Although veterinary oriented clinical coding systems, such as VeNom and VetSCT, are incorporated into the practice management systems of many clinics their use is inconsistent with the risk of introducing bias to the coded data that they generate.

3.1.3 Clinician coding of SAVSNET consultations

Consultations within the SAVSNET dataset are collated via participating veterinary clinics (See Chapter two). At the end of a pre-booked consultation the clinician is presented with a graphical browser window (Figure 3.1.a) and asked to assign the main reason for the consultation into one of ten categories, or to indicate that the owner has opted-out of participation. These categories are a mixture of entity types; physiological systems (gastroenteric, respiratory), clinical signs (pruritus), disease

processes (kidney disease) and mechanisms (trauma) with catch-all 'other' categories (Table 3.1.a).

The SAVSNET interface is designed to require a single mouse click from the clinician, minimising the time-burden, and requires every consultation to be assigned to a category, unless the owner has chosen to opt-out, in an attempt to minimise the risk of introducing bias during the data collection process.

Table 3.1.a: Categories into which the attending clinician is asked to assign each consultation prior to data transfer to SAVSNET

Vet assigned category	Definition provided by SAVSNET
Ill-animal categories	
Gastroenteric	Signs including but not limited to: diarrhoea, vomiting, weight loss, poor appetite.
Kidney disease	Signs including but not limited to: polydipsia, polyuria, vomiting where kidney disease is a differential.
Pruritus	Signs including but not limited to: itching, scratching, pruritic otitis, chewing, licking, rubbing.
Respiratory	Signs associated with conditions affecting the upper and / or lower respiratory tract.
Tumour	Any suspected or confirmed benign or malignant neoplastic condition
Trauma	Animal suffering a trauma and / or a physical injury.
Other unwell	Signs that do not fit in other unwell animal categories including behaviour problems.
Healthy -animal categories	
Post-operative	If the animal has presented for post-operative care.
Vaccination	If the animal was booked in for a vaccination and was vaccinated.
Other healthy	Healthy animal presented for other reasons that do not fit in the vaccination or in the post-op check categories.

In addition to the categorical coding applied to all SAVSNET-collated consultations, in a small proportion of consultations, estimated as 5% at practice recruitment, the main reason for presentation indicated by the clinician triggers a request for completion of a more detailed questionnaire. Questionnaires are linked to a specific vet-assigned category of consultation, and comprise up to seven questions with on screen selection of appropriate answers via pointing device interactions (Table 3.1.b & Figure 3.1.b). Questionnaires are only triggered if the clinician selects gastroenteric, kidney disease, pruritus, respiratory or tumour as the main reason for visit.

Figure 3.1.a: Screenshot of the 'SAVSNET window' the graphical interface through which the attending clinician assigns a main reason for visit

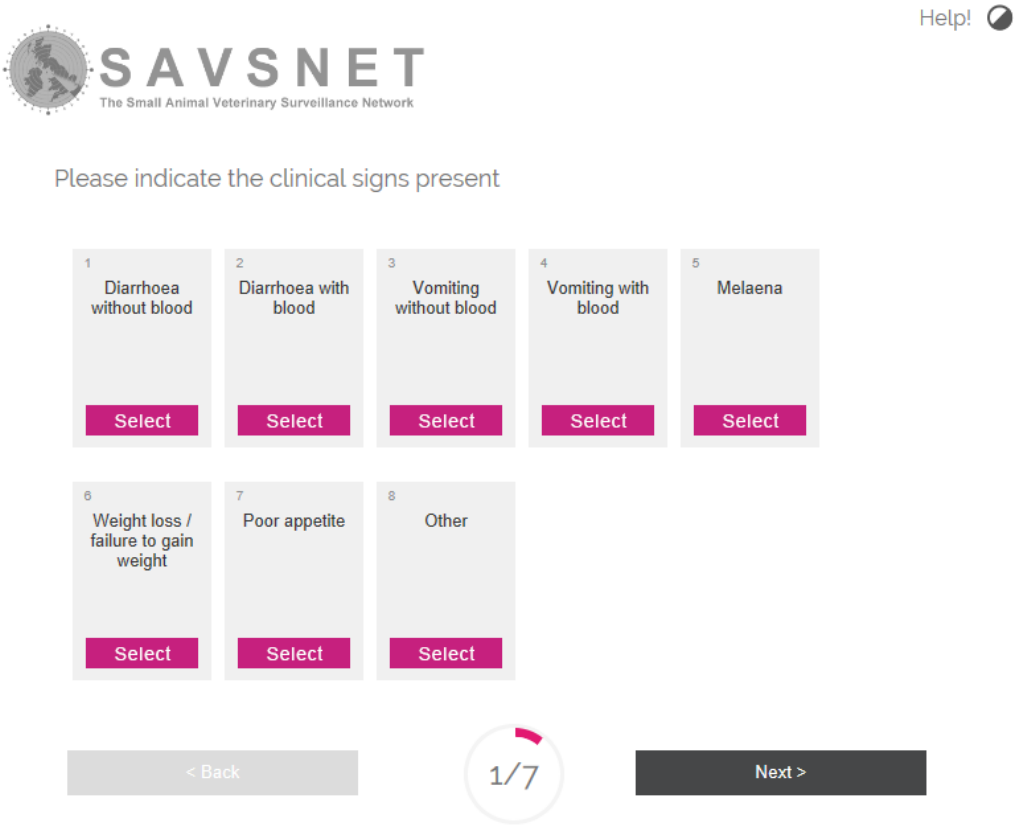

Figure 3.1.b: Screenshot illustrating the graphical interface through which clinicians respond to questionnaires. The question illustrated is the first of seven questions in the gastroenteric questionnaire.

Table 3.1.b: Example of the nature of questionnaire triggered when the respiratory option is selected as the main reason for presentation

Question	Options
Has this animal stayed at a kennels or cattery within the last 10 days?	Yes No Don't know
How long approximately has the pet had this episode of illness?	Less than 1 week Between 1 week - less than 1 month 1 month and over Don't know
The likely cause of this episode of illness is	Respiratory Cardiac Other Don't know
What diagnostic tests will be performed today for this episode of illness?	None Haematology Biochemistry Virology Bacteriology Parasitology Radiography Ultrasound Other
Has the animal returned from outside of the UK in the last 10 days?	Yes No Don't know
Please indicate the clinical signs present	Coughing Sneezing Nasal discharge Conjunctivitis and/or ocular discharge Drooling Dyspnoea Mouth ulcers Generalised depression / lethargy Pyrexia Other
How does this consultation relate to this episode of illness?	First presentation Revisit / check-up Don't know

Thus, the SAVSNET dataset has two, linked, methods of coding, with all consultations assigned a main reason for visit category and a small proportion of consultations also associated with a more detailed clinician-completed questionnaire. As coding and questionnaire completion is by the attendant clinician a degree of the error introduced during many clinical coding mechanisms, as used in human health care, could be expected to be mitigated as a tier of interpretation has been removed from the process.

The aim of this chapter was to quantify the efficacy of SAVSNET's clinician-assigned coding of main reason for visit and the questionnaire responses, using information available within the free-text record, and thereby explore the advantages that an automated coding system deriving information from the free-text record conferred.

3.2 Methods

3.2.1 Evaluation of clinician-assigned categorical classification

The reliability of the categorisation assigned by attending clinicians was gauged by manually reading by the author, who was medically trained and had extensive familiarity with the notation used within the veterinary narrative. A randomly selected sample of 1000 free-text records of consultations regarding cats or dogs, that had occurred during a 30-day period, was drawn from the SAVSNET dataset. A 30-day window was used to control for any effect of time since implementation on the efficacy of the classification system, and to capture a sample representative of the consultation mix captured in near real-time.

A Python script was used to present each consultation narrative in turn within a terminal window, to ensure author-assigned coding was blind to the clinician-assigned classification and all other database fields. All clinical signs that were described within the consultation narrative were documented and the apparent main reason for the consultation, based on information within the free-text record, was assigned using the same definitions that had been provided to clinicians (Table 3.1.a). Where a consultation was not associated with a free-text record this was noted and the Z method used to establish whether there was a difference in the proportion of ill and healthy animal consultations where this occurred.

The categorical classification assigned by the attending clinician and the author were compared and the sensitivity, specificity, and predictive value of the clinician-

assigned classifications were calculated, using the author's assigned classification as a proxy gold standard, whilst acknowledging its limitations as such. Where the narrative record was blank this was considered missing data and was not included in the calculations.

Confidence intervals were used to describe the uncertainty in the point estimate of proportions. These were calculated using the Normal Approximation to calculate the standard error of the sample proportion.

3.2.2 Trend in apparent proportional morbidity

The proportional morbidity (the number of animals presented for a given reason as a proportion of all reasons for visiting, during a given period of time), using the clinician-assigned main reason for visit classifications, was plotted over the three-year period from January 2015. For this purpose, any change in species proportion over time was controlled for by constraining the data to consultations regarding dogs.

The proportional morbidity for each year from 2015 was calculated, tabulated and plotted. The kidney disease category was introduced on the 6th March 2015, for this category calculations were based on data from the 7th March 2015 onwards, prior to that date it was anticipated that consultations would have been assigned to the 'other unwell' category. A one-way analysis of variance (ANOVA) test was performed on the data for each category to establish whether the proportional morbidity for that category differed across the three years studied at a significance level of $\alpha = 0.05$.

3.2.3 Quantification of data obscured by exclusive categorisation

One of the key disadvantages to a main reason for visit classification system is its exclusivity and consequent shrouding of a large amount of pertinent clinical information not encapsulated in the reason for visit, in a similar manner to a system with poor coverage (Chute et al. 1996). The manually-coded dataset of 1000 consultations was used to examine the distribution of clinical signs across the clinician-assigned categories. The proportion of manually identified respiratory and gastrointestinal clinical signs by clinician-assigned category was quantified.

3.2.4 Evaluation of SAVSNET questionnaire response

3.2.4.1 Respiratory questionnaire

A random sample of 1000 cat or dog consultations was selected from those where the attending clinician had assigned a respiratory main reason for consultation, completed the respiratory questionnaire and indicated that the animal was coughing. An additional 200 consultations were selected at random from respiratory questionnaire responses where the clinician had indicated that the animal was not coughing. These consultations were combined and shuffled, before being manually read, blind to the clinician's response, and allocated one of three classifications based on evidence in the free-text record: a) explicitly states that animal was not coughing; b) does not state that animal was coughing; c) animal was documented to be coughing. This latter classification included where there was a positive 'tracheal pinch' or tracheal sensitivity was documented and where the dog was documented to have 'kc' which was often used as notation for 'kennel cough' a syndrome of tracheobronchitis. Where category b applied, the number of consultations where no free-text record had been documented was also noted.

3.2.4.2 Gastroenteric questionnaire

In the same manner as for the respiratory questionnaire, a random sample of 1000 cat or dog consultations was selected at random from those consultations where the attending clinician had assigned a gastroenteric main reason for consultation, completed the gastroenteric questionnaire and indicated that the animal had diarrhoea. An additional 200 consultations were selected at random from gastroenteric questionnaire responses where the clinician had indicated that the animal did not have diarrhoea. The consultations were read and allocated three classifications analogous to those for the sample of cough consultations. Notation of HGE (haemorrhagic gastroenteritis), GE (gastroenteritis) and colitis was considered confirmation of diarrhoea being documented in these circumstances.

3.3 Results

3.3.1 Clinician-assigned categorical classification

Within the sample of 1000 consultations from January 2018, 62 (6.2%) consultations were not associated with a free-text record. This lack of documentation was disproportionately more common in consultations classified by the clinician as being for healthy animals ($p<0.01$), i.e. those consultations categorised by the clinician as other-healthy, vaccination and post-operative visits.

Results

The most common of the five specific ill-animal reasons for attendance according to clinician coding was pruritus, accounting for 4.3(95%CI: 3.04, 5.56)% of consultations, however manual assignment to the same categories found that the most common was a gastroenteric main reason for visit, with this being apparent in 7.25(95% CI: 5.59, 8.91)% of consultations (Figure 3.3.a).

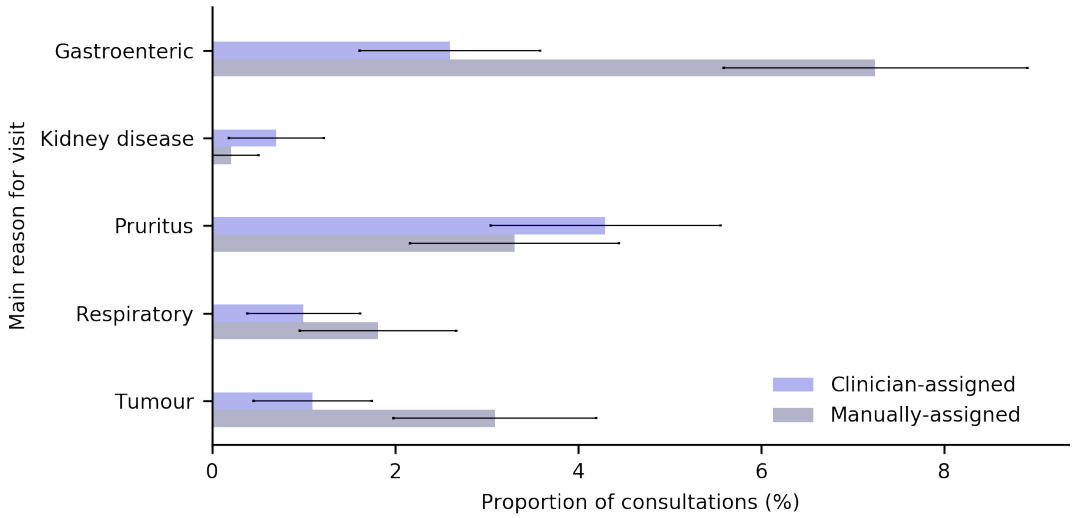

Figure 3.3.a: Apparent proportional morbidity, comparison of the clinician assigned main reason for visit and the apparent main reason for visit based on manual reading of the free-text record. Error bars represent the 95% confidence interval of the proportion.

The sensitivity of the clinician-assigned main reason for visit classification was low, although the specificity exceeded 96% for the five specific 'ill-animal' and the trauma categories, this obscured some limitations of the system given the disproportion of the true negative cases to true positives for each category (Table 3.3.a).

Table 3.3.a: Comparison of the vet assigned and apparent main reason for consultation in a random sample of 1000 consultations collated in the SAVSNET dataset in January 2018. *miss* indicates that the free-text record was blank, there were a total of 62 blank free-text records in the sample. *man* indicates the number of consultations where this appeared the main reason for visit on manual coding, and *vet* the number assigned to that category by the attending clinician.

Main reason		Apparent						Estimated		
for visit	vet	miss	man	tp	fp	tn	fn	sensitivity (%)	specificity (%)	PPV (%)
Gastroenteric	26	1	68	23	2	868	45	33.82(22.58, 45.06)	99.77(99.45, 100)	92(81.37, 100)
Kidney disease	7	2	2	1	4	932	1	50(0,100)	99.57(99.15, 99.99)	20(0, 55.06)
Pruritus	43	3	31	18	22	885	13	58.06(40.69, 75.43)	97.57(96.57, 98.57)	45(29.58, 60.42))
Respiratory	10	1	17	6	3	918	11	35.29(12.57, 58.01))	99.67(99.3,100)	66.67(35.87, 97.47)
Trauma	48	0	42	20	28	868	22	47.62(32.52, 62.72)	96.88(95.74, 98.02)	41.67(27.72, 55.62)
Tumour	11	0	29	8	3	906	21	27.59(11.32, 43.86)	99.67(99.3, 100)	72.73(46.41, 99.05)
Mean (un-weighted)								**42.06**	**98.86**	**56.35**

Results

On expanding evaluation to the full dataset, to examine the trend in proportional morbidity, there was no difference associated with the year of consultation in the proportion of consultations assigned by the clinician to the vaccination category (p = 0.13), for all other categories there was a difference associated with year of consultation (p<0.001 for each category). For example in 2015 the mean proportion of dog consultations assigned to the gastroenteric category by the attending clinician was 4.15 (95% CI: 4.08, 4.22)%, in 2016 the mean proportion was 3.4 (95% CI: 3.36, 3.44)% and in 2017 2.79 (95% CI: 2.75, 2.83)% (p<0.001). A similar pattern was seen across the five specific ill animal categories associated with a questionnaire (Figure 3.3.b) and the trauma and post-operative category. The opposite pattern was observed, with increasing clinician assignment between 2015 and 2017, in the other healthy and other unwell categories (Figure 3.3.c).

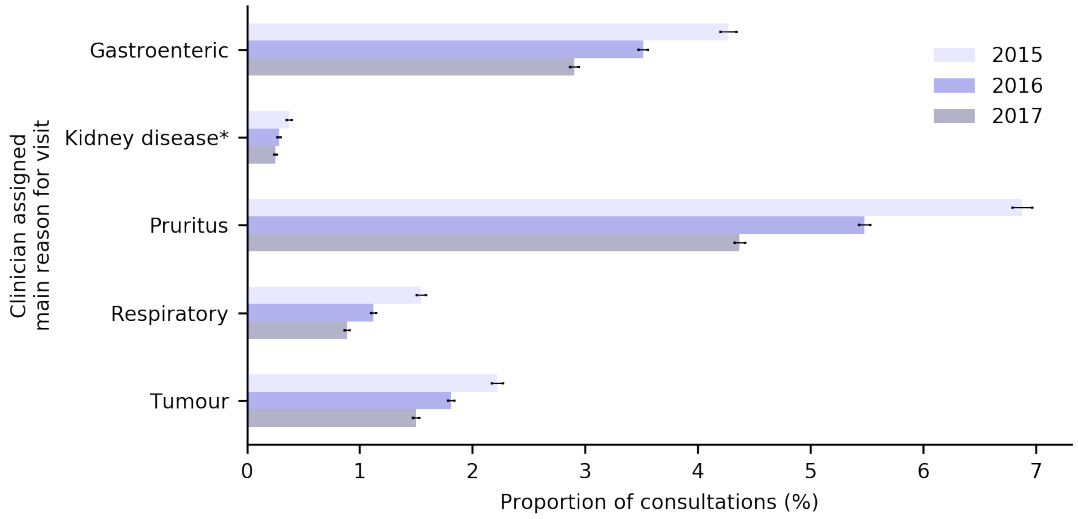

Figure 3.3.b: Bar plot of the proportion of consultations assigned by the attending clinician to each category associated with a questionnaire. * The kidney disease category and questionnaire was introduced on March 6th 2015, proportion for 2015 in this category uses only March 7th onwards as denominator.

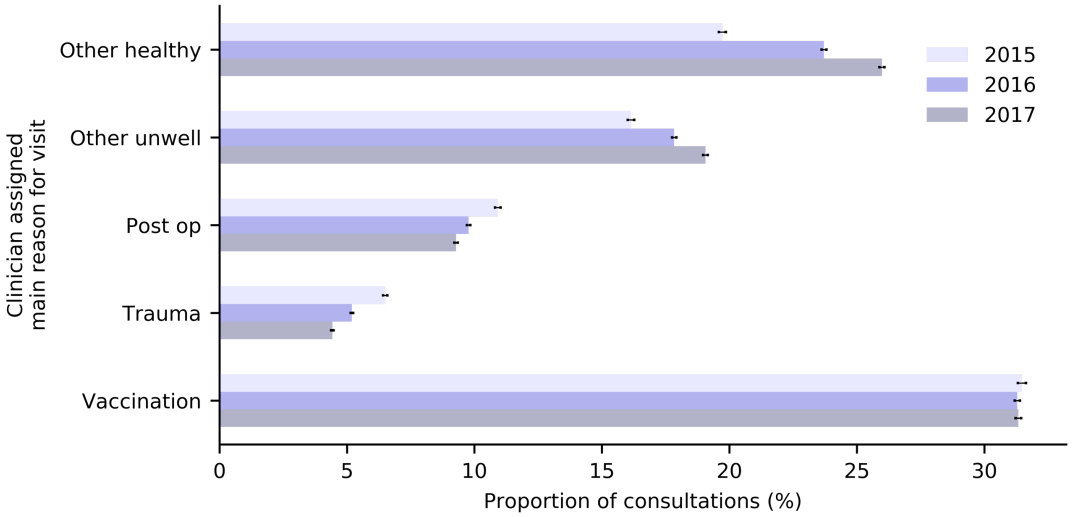

Figure 3.3.c: Bar plot of the proportion of consultations assigned by the attending clinician to each category that was not associated with a questionnaire.

The declining trend in assignment to the specific ill animal categories was also evident in time trend plots of the proportional morbidity for these categories (Figure 3.3.d).

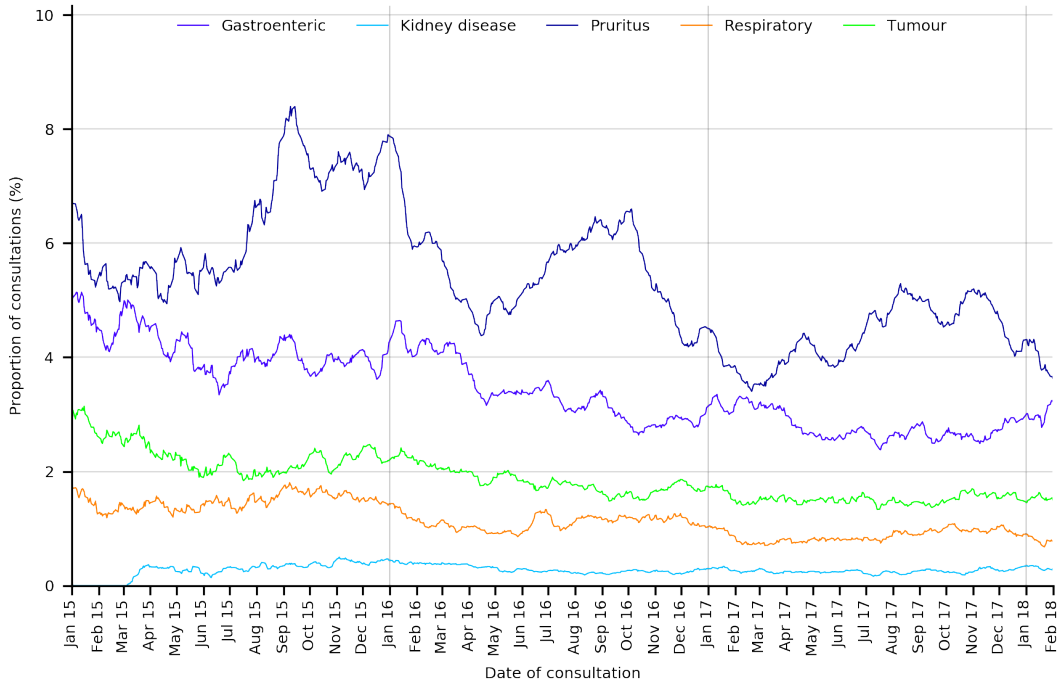

Figure 3.3.d: Temporal trend in clinician-assigned main reason for visit

3.3.2 Distribution of clinical signs by clinician-assigned category

Documentation of 1,115 clinical signs was identified within the sample of 1,000 consultations that were coded for the presence of clinical signs and apparent main reason for visit. Marginally more than half of these signs were found within the free-text record of consultations categorised by the clinician as being for ill animals, 52.91 (95% CI: 51.42, 54.4)% , the remainder, 47.09 (95% CI: 45.6, 48.58)% were identified within consultations ostensibly for healthy animals.

The majority, 74.25 (95%CI: 67.62, 80.88)%, of gastroenteric clinical signs identified from the free-text were recorded in the records of consultations classified by the attending clinician as having been for a main reason other than gastroenteric (Figure 3.3.e). Similarly only 17.46 (95% CI: 8.09, 26.83)% of respiratory signs identified on manual reading had been documented in the records of consultations classified by the clinician as having a respiratory main reason for the visit (Figure 3.3.f).

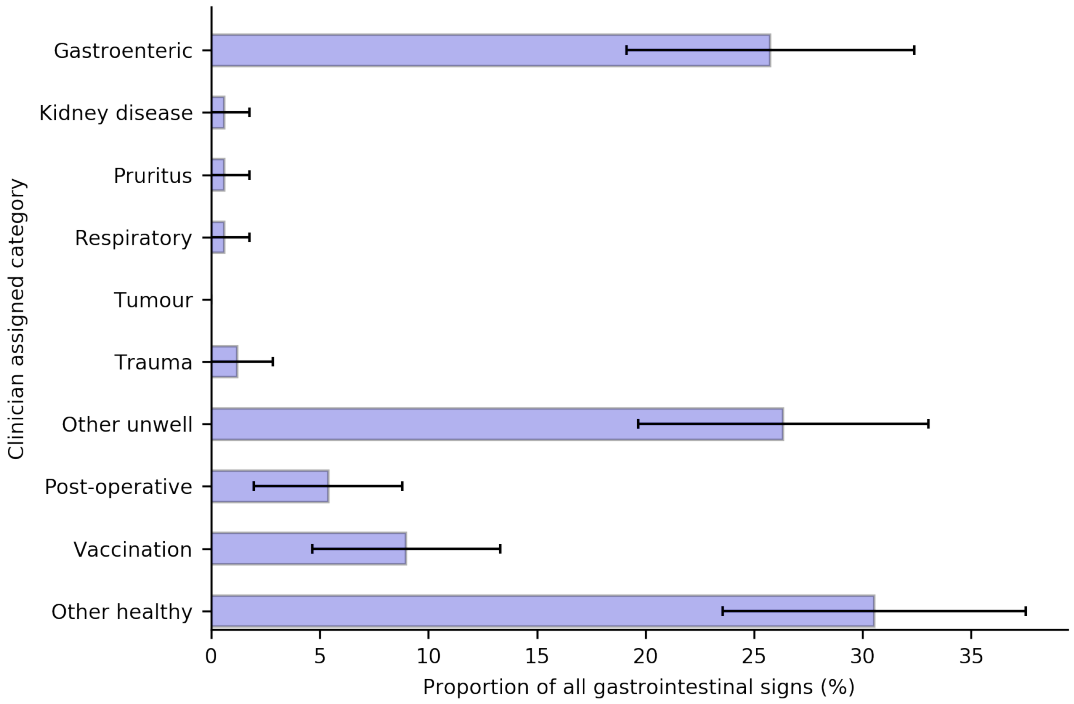

Figure 3.3.e: Distribution of gastrointestinal signs, identified by manual reading of the free-text record of 1000 consultations, in relation to the clinician-assigned main reason for visit category. Error bars represent 95% confidence interval of the proportion.

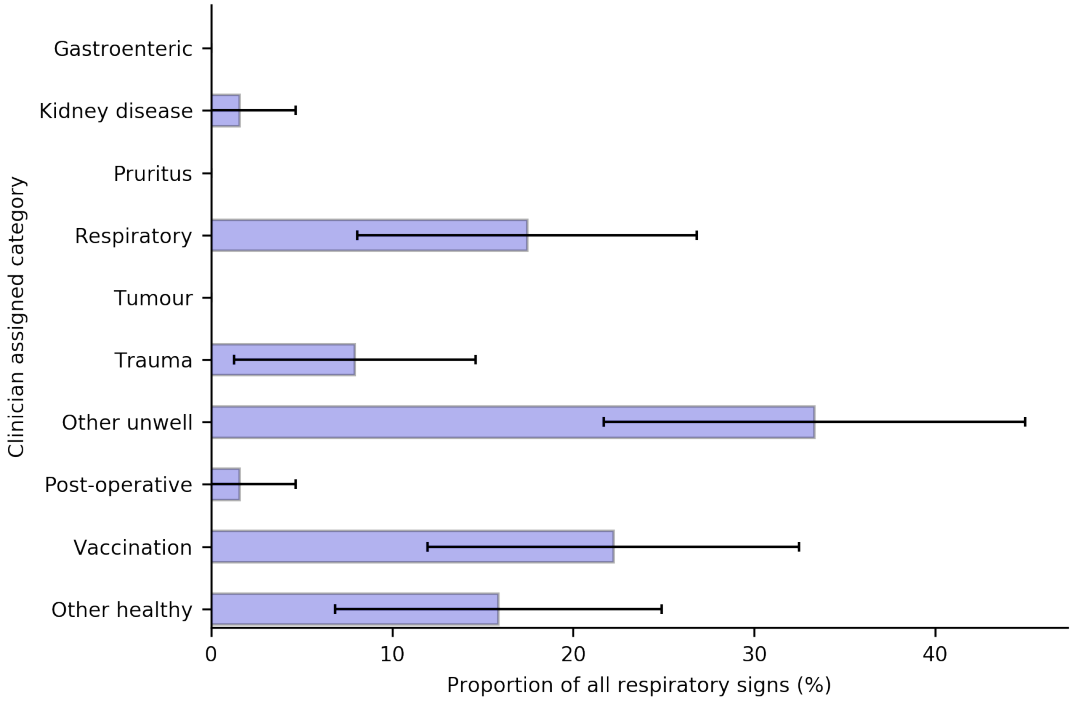

Figure 3.3.f: Distribution of respiratory signs, identified by manual reading of the free-text record of 1000 consultations, in relation to clinician-assigned main reason for visit category. Error bars represent 95% confidence interval of the proportion.

3.3.3 Predictive value of SAVSNET questionnaire response

Overall 2.6% of consultations within the SAVSNET dataset had an associated questionnaire. For four of the specific 'ill-animal' categories, approximately 1 in 4 consultations were associated with a questionnaire; the kidney disease questionnaire being deployed in only 4.5% of consultations assigned to the kidney disease main reason for visit (Table 3.3.b).

Table 3.3.b: Proportion of classified consultations

Main reason for visit	Consultations assigned to category	Linked questionnaire	Questionnaires completed	Consultations with associated questionnaire
Gastroenteric	96992	Gastroenteric	22852	23.6%
Kidney disease	8783	CKD (Feline)	394	4.5%
Kidney disease	5477	CKD (Canine)	247	4.5%
Pruritus	143747	Pruritus	34217	23.8%
Respiratory	38954	Respiratory	9075	23.3%
Tumour	49224	Tumour	11507	23.4%

Results

Evaluation of the questionnaire responses in comparison to information available within the free-text record found that in 81.2 (78.78, 83.62)% of questionnaires for coughing and 86.5 (84.38, 88.62)% for diarrhoea there was corroborative evidence that a sign recorded as present in the questionnaire was present in the presented animal. In 3.6 (2.45, 4.75)% of responses for coughing and 4.7 (3.39, 6.01)% for diarrhoea the free-text record provided evidence directly contradicting the questionnaire response, i.e. stated a sign was not present where the questionnaire stated it was. Excluding those consultations where a free-text record had not been recorded, the positive predictive value (PPV, precision) of a positive questionnaire response was 84.23(95% CI: 81.93, 86.53)% and 89.45(87.52, 91.39)% for coughing and diarrhoea respectively (Tables 3.3.c & 3.3.d).

Table 3.3.c: Comparison of respiratory questionnaire responses regarding presence of coughing with information within the narrative record of the associated consultation. *36 of these records had no free-text documented.

Manual classification	n=	Proportion (95% CI)
Not coughing	36	3.6 (2.45, 4.75)%
Not documented to be coughing	152*	15.2(12.97, 17.43)%
Documented animal was coughing	812	81.2(78.78, 83.62)%

Table 3.3.d: Comparison of gastroenteric questionnaire responses regarding presence of diarrhoea to information within the narrative record of the associated consultation. *33 of these records had no free-text documented.

Manual classification	n=	Proportion (95% CI)
No diarrhoea	47	4.7 (3.39, 6.01)%
Not documented to have diarrhoea	88*	8.8(7.04, 10.56)%
Documented animal had diarrhoea	865	86.5(84.38, 88.62)%

Difficulty arose in deciphering the validity of questionnaire responses where there was no direct evidence, within the free-text record, for the presence or absence of a clinical sign stated to be present in the associated questionnaire. These likely represented a mixture of consultations where the sign was present, but not documented, and where it was not present but its absence was not documented.

The predictive value of a clinician responding to a questionnaire indicating that an animal did not have a given clinical sign (negative predictive value) was similar to the positive predictive value. For example the negative predictive value for a response of no diarrhoea was 88.23 (95%CI: 83.76, 92.77)% i.e. for every 100

questionnaire responses indicating that an animal did not have diarrhoea, 12 of the animals appeared to have diarrhoea based on information present within the associated consultation record.

In contrast to the decline in allocation to main reason of visit categories, when a questionnaire was triggered the proportion of responses indicating that a given sign was present in the presented animal was relatively stable (Figure 3.3.g).

3.4 Discussion

Forming a measure of the efficacy of clinician-assigned main reason for visit categorisation posed a challenge. In addition to the limitations introduced by the mixed nature of the categories, there was difficulty in post-hoc abstract assignment of the main reason for the consultation, and constraint of summarising the wealth of information exchange occurring during a consultation into a single category.

The free-text record of many consultations includes an indication of the main reason for visit (Section 4.4.6) in the form of an opening statement (for example 'Boosters', or 'Cough again'), however this was not universal. In many consultations, even without an opening statement, the main reason was beyond significant doubt by virtue of the history and examination documented in the free text record. Thus, although it was challenging to correlate the clinician-assigned classification with a manual classification by human reading of the historical free-text record, as the information present within the documented record differs from that available to the attending clinician (Jones-Diette et al. 2017), a reasonable measure was achievable. This is the manner of comparison used in other published evaluations (Assareh et al. 2016; Lloyd 1985), and is the best retrospective evidence available. A valuable alternative would be contemporaneously gathering evidence by observing the consultation or interviewing the patient and or clinician, as was undertaken by Robinson et al. (2014; 2015).

The finding that a considerable proportion of consultations where the main reason for visit appeared to belong to one of the ill-animal specific categories but had not been categorised as such, was compounded by the finding that the majority of clinical signs belonging to the respiratory and gastrointestinal system were identified in consultations not assigned by the clinician to their parent category. The implication of this is that, if only the main reason for visit is considered, the majority of clinical signs would be neglected. It may be that this would filter the more severe

manifestations of clinical signs and so may have a place, however this would require further evaluation.

Within the acknowledged limitations of the evaluation undertaken, the efficacy of the clinician-assigned classification is concerning (Table 3.3.a). Impaired sensitivity for the key categories likely to be used in surveillance suggested that the majority of consultations that would be expected to be found in a given category were not classified as belonging to it, and there was poor reliability of those consultations that were classified to a category.

The limited options available and exclusive nature of the initial clinician-assigned main reason for visit was frustrating from a categorisation and surveillance perspective, and appeared to impair the clinician's ability to assign to the correct category. However, this simplicity did offer advantages and was designed to do so. The intention was to minimise the time required of the clinician and so encourage their contribution whilst attempting to gather high quality information.

The questionnaires deployed in a quarter of clinician-assigned gastroenteric and respiratory consultations were able to capture surveillance-valuable information; the respiratory questionnaire cough question had a positive predictive value of 81.2 (78.78, 83.62)% and the gastroenteric questionnaire's diarrhoea question a positive predictive value of 86.5 (84.38, 88.62)%. However, this overlooked the 75% of consultations assigned to these categories by the attending clinician where a questionnaire was not deployed, the consultations erroneously not assigned to the category and the large proportion of clinical signs present in animals presenting for other reasons and either identified incidentally or forming part of a wider complex of clinical signs.

3.4.1 Decline in clinician assignment to ill animal categories

Within the SAVSNET clinic interface, once a veterinary practice has committed to contribute to SAVSNET, all pre-booked consultations require the initial clinician-assigned category to be completed, before the consultation record can be finalised for payment, and if a questionnaire is triggered there is no option to quit without responding to each question. This may account for a degree of the apparently erroneous responses to questionnaires and categorisation, with the quickest or least demanding option being chosen.

The decline in the proportion of consultations assigned to the specific ill animal categories (Figure 3.3.e) may reflect a learned response. Clinicians familiar with the SAVSNET system will be aware that if they assign a consultation to gastroenteric, pruritus, respiratory or tumour categories there is a 1 in 4 chance that they will then be asked to complete a questionnaire.

This may account for the decline in classifications to these categories and increase in assignment to the other unwell and other healthy categories (Figure 3.3.f). Decline is however also observed in the trauma and post-operative categories which are not associated with a questionnaire, this raises the possibility of an alternative explanation, perhaps related to option location on the screen and ease of selection (Figure 3.1.a).

3.5 Conclusion - The promise of text-mining

Within the SAVSNET dataset of over 3 million consultation records, the clinician assigned categorisation identified 96,992 (3.17%) consultations where the main reason for visit was gastroenteric and 38,954 (1.27%) where the main reason was respiratory. These consultations are associated with 22,852 gastroenteric and 9,075 respiratory questionnaires, representing 0.75% and 0.3% of all collated consultations respectively. Consequently, even were the precision of classification and questionnaire response perfect, the wealth of information within the consultation record is sequestered. Text-mining directly from the free-text record would liberate this latent clinical and research rich information, enabling all consultation records, where a free-text record has been documented, to be included in surveillance and epidemiological study.

In contrast to the fixed surveillance measures provided by the SAVSNET clinician-assigned coding and questionnaires; classification via the free-text permits identification of any combination of clinical signs. Binary clinical sign indicators can be combined using Boolean terms to generate a syndrome indicator across the full dataset, these can be applied retrospectively or prospectively with responsive adaption of the barrage of classifiers dependent on surveillance and research needs. In contrast where questionnaire or clinical-coding is applied at the time of consultation, later amendments to that coding is not possible, and if definitions or questionnaires are altered they can only be applied prospectively.

The ability to generate indicators of information within unstructured free-text clinical records would confer capacity to utilise that information in the same manner as any other coded field in epidemiological studies and surveillance. Likewise, techniques to extract the values of numeric parameters from the free-text record would confer the ability to evaluate trends in those parameters on a large scale.

Chapter Four The small animal veterinary clinical narrative

4.1 Introduction

Computational extraction of information from narrative data, text-mining, requires an understanding of the vocabulary, structure, semantics and syntax of the language within the data being evaluated (Friedman, Kra, and Rzhetsky 2002; Harris 1991). This chapter will explore the nature of language and notation used in documenting first-opinion, small-animal, veterinary consultations, and describe differences in sentence metrics associated with timing of consultation and the species of animal presented. As a whole, the chapter characterises the small-animal veterinary clinical sublanguage.

4.1.1 Language and sublanguage

A linguistic domain, consisting of the subject field and associated vocabulary, is characterised by the union of its topic of reference, the register or degree of formality and the intent of the communication (Biber 1988). Domains are used within a discourse community where individuals communicate with a specific group of words and phrases (lexis) to convey information and feedback in pursuance of shared purpose (Swales 1990). Within a domain, the language used varies in response to the situation in which it is being used and relationships involved in the communication (diatypic variation) (Gregory 1967).

Harris provided much of the foundation work in describing the architecture of language and specialised sublanguages (Harris 1991; Harris 1981), suggesting that a language consisted of word sequences used to convey information occurring within constraints of dependency, likelihood and paraphrastic reduction, to convey information (Harris 1991). Paraphrastic reduction describes the intuitive changes to the structure of a sentence that occur by eliminating information that is redundant in context whilst maintaining its information content (Friedman, Kra, and Rzhetsky 2002). Within a language, dependency relations dictate the permissible sequence of words, with hierarchical dependence of words within a sentence on other words, and defined interdependence between classes of words.

Friedman, Kra and Rzhetsky (2002) used Harris' work to describe clinical sublanguages; they reported that within a sublanguage the grammar used is more specialised with dependency relations dictated by semantic constraints, related to the field of the sublanguage, in addition to the wider syntactic word

class constraints of the parent language. Where several pieces of information are being conveyed, paraphrastic reduction results in the elimination of redundant words and rearrangement of the operators and arguments whilst preserving the information conveyed. Within clinical sublanguages, telegraphic phraseology may occur, with information being conveyed by the minimum word sequence, as a result of the omission of words occurring at high frequency, paraphrastic rearrangement and the semantic dependency relations permitting the use and interpretation of, in the extreme, noun only phrases.

A sublanguage develops where there is restricted domain of reference and restricted, goal-oriented, purpose and mode of communication within a community sharing specialised knowledge (Kittredge 1983). The electronic health records used as a method of documentation and communication within small-animal veterinary practice provide the embodiment of Kittredge's description of a canonical example of a sublanguage (Figure 4.1.a).

> *"The best, canonical examples of sublanguages are those for which there exists an identifiable community of users who share specialized knowledge and who communicate under restrictions of domain, purpose, and mode by using the sublanguage. These participants enforce the special patterns of usage and ensure the coherence and completeness of the sublanguage as a linguistic system."*

Figure 4.1.a: Kittredge's description of the conditions in which a sublanguage occurs (Kittredge 1983).

4.1.2 Information recorded within the clinical narrative

A clinical consultation is a considerably more complex interaction than the classical premise that it is the sum of history taking, examination of the patient, diagnosis and management (Ledley and Lusted 1959). The consultation is now recognised to be an interactive and exploratory patient-centred process (Neighbour 2004), typically including the aggregation and exchange of an array of information encompassing health and socio-demographic factors, enveloped within the expectations of both clinician and client (Thorsen et al. 2001; Fischer and Ereaut 2012; Coe, Adams, and Bonnett 2008; Everitt et al. 2013).

In documenting a clinical record, it is expected that a complete record of examination, treatment and investigation is recorded alongside information conveyed between parties. In the United Kingdom these expectations are overseen by regulatory bodies such as the Royal College of Veterinary

Surgeons (RCVS) (Royal College of Veterinary Surgeons 2014) and General Medical Council (General Medical Council 2013). Figure 4.1.b demonstrates the detailed information that the RCVS Code of Professional Conduct for Veterinary Surgeons states should be included in the client and clinical record (Royal College of Veterinary Surgeons 2014).

Details of examination
Treatment administered
Procedures undertaken
Medication prescribed and/or supplied
Results of any diagnostic or laboratory tests
Provisional or confirmed diagnoses
Advice given to the client
Outline plans for future treatment or investigations
Details of proposed follow-up care or advice
Notes of telephone conversations,
Fee estimates or quotations
Consents given or withheld,
Contact details
Recommendations or discussion about referral

Figure 4.1.b:Information that should be recorded within the health record in accordance with the RCVS Code of Professional Conduct for Veterinary Surgeons

Veterinary consultations broadly fall into three types: presentation of an animal with a new complaint; ongoing management of a pre-existent condition; and routine and preventive veterinary health care (Everitt et al. 2013; N. J. Robinson et al. 2015). The number of issues explored during a consultation varies widely (N. J. Robinson et al. 2015) and it should not be assumed that a consultation primarily for one purpose does not include exchange, and thus potentially documentation, of information regarding other clinical issues (G. Jackson 2005). First opinion veterinary consultations are often allotted ten minutes when scheduling, but they range in duration from less than a minute to 37 minutes, with a median approximating to 10 minutes (N. J. Robinson et al. 2014; Everitt et al. 2013). These factors are likely to influence the volume of information documented in the narrative record of a consultation.

4.1.3 Structure of information for ease of abstraction

The complex interactions and iterative nature of clinical consultations (Everitt et al. 2013) mandate that documentation occurs in a manner that facilitates access

to information by the future reader. The length of sentences, complexity of words used and organisational cues influence reading speed and comprehension of language (Spyridakis 2000). Familiarity with the structure of text, conferred by the arrangement and relationships between pieces of information, creates an expectation of information availability and enhances the reader's ability to abstract information (Armbruster, Anderson, and Ostertag 1987). The factors likely to impact automated abstraction of information are not dissimilar to those affecting human abstraction, with pattern recognition and knowledge of structure and boundaries of packets of information being key factors (Friedman, Kra, and Rzhetsky 2002; Harris 1991).

This chapter explores the sublanguage of the small-animal veterinary clinical narrative as documented within electronic health records (EHR), describing features of sentence architecture and the structure and semantics of the language used in these clinical records. The consultation narratives collated by SAVSNET were used as the source of an exploratory corpus (University of Liverpool 2017).

4.2 Materials & methods

4.2.1 Preparation of a representative exploratory corpus

Consultation narratives within the SAVSNET dataset were retrieved alongside signalment and clinic information for each consultation. A small number of clinics introduced a clinical checklist template which constrained the text used in the affected consultations and contributed the majority of the text recorded during those consultations. Consultations containing these checklists were excluded from the dataset for the purposes of this analysis. Consultations where no species had been documented were excluded as this precluded later analyses. For this purpose, the species originally recorded by the veterinary practice, and not the subsequently mapped species created by SAVSNET, was utilised to avoid propagating any bias in mapping related to use of language.

For each clinic, the period of contribution (days between earliest and most recent consultation date), mean number of consultations contributed per day and total number of consultations contributed were calculated. An SQL query retrieved the data via the pymssql library version 2.1.3 (pymssql developers 2016) and the Pandas library version 0.20.3 (PyData Development Team 2017)

was used for data handling with all code written in Python version 3.6.1 (Python Software Foundation 2016).

Technical or work-flow issues had led to some clinics contributing to the SAVSNET dataset on only a small number of days, as a result if all clinics were to be equally represented the sample drawn from each clinic would need to be very small to enable inclusion of those that had contributed few consultations. A threshold of consultation sample size that permitted the inclusion of 95% of clinics was used as a pragmatic division, enabling the inclusion of small clinics whilst leaving sufficient data for lexical analysis. A randomly selected equal sample was drawn from those clinics that had contributed this threshold number of consultations to form the exploratory corpus. In addition to ensuring equal inclusion of smaller clinics, this negated the effect of over-representation of larger practices and those that had contributed for the longest time.

4.2.2 Species mapping

Species groups were created by mapping the species recorded by the veterinary practice to predefined groups for lexical analysis purposes. The vast majority of small animal consultations involved a cat, dog or rabbit, these were treated separately. The remainder of consultations were mapped to a diverse 'Uncommon' group as these represented animals seen uncommonly by the small animal clinician, and there were not expected to be sufficient examples for meaningful evaluation at species nor class level.

4.2.3 Designation of consultations as having occurred in- or out-of-hours

A consultation was considered out of hours if any of the following applied:

- It occurred on a Sunday
- It occurred after 12 noon on a Saturday
- It occurred between 8pm and 8am the following day
- It occurred on a public holiday (known as bank holidays in the UK)

All remaining consultations were considered to have occurred in-hours.

4.2.4 Methods of measuring sentence metrics

Corpus metrics were quantified and explored in relation to the species being treated and whether or not a consultation occurred within normal working hours.

4.2.4.1 Word count

Words were defined in a manner intended to capture the discrete clinically meaningful elements of the narrative text. These included sequences containing letters, numbers and the characters <, >, ^, +, ~. Hyphen (-) used to indicate the opposite of +, in for example *d-*, as in *no diarrhoea*, was only included within a word where it was the final character of the word, with the next character a non-word character, or at the end of the string. Full stop (.) was only permitted where it was used as a decimal point within a float. The number of words within each narrative was established using a pandas string function `pandas.Series.str.count(regex, flags = re.I)`, with the regular expression `(?:(?:[\d]+|(?<=\s))[\.]\d+|[a-z0-9<>\^+~\']+(?:\-(?=\W|$))?)`. Where analyses compared words to an external dictionary, and during calculation of lexical diversity, words were extracted as sequences of letters using the regular expression `[a-z]+`.

4.2.4.2 Word length & complexity

The number of characters within each word were measured by applying the Pandas string method `pandas.Series.str.count('\S')` to the series of extracted words within each narrative, this counted the number of non-whitespace characters within a string composed of only non-whitespace characters.

A bespoke script was created to generate an empirical syllable count approximation intended to return a meaningful number of syllables for words found within the clinical narrative (Figure 4.2.a). The Carnegie Mellon (CMU) Pronouncing Dictionary (Weide 1998), an open-source machine readable dictionary of 123,455 words and their pronunciation in North American English, was used as a source of validated syllable count for words of 4 or more letters, ignoring a terminal ‘s’. This was the closest available reference for British English pronunciation. The dictionary was imported as the `cmudict` module of Python’s Natural Language Toolkit ((NLTK Project 2015) `nltk.corpus.cmudict`). The CMU dictionary included some abbreviations in common use and traditional spellings for words such as ‘diarrhoea’. If a word was not found in the dictionary, a second attempt was made having replaced ‘ou’ with ‘o’ and ‘is’ with ‘iz’ to account for the dictionary using American English.

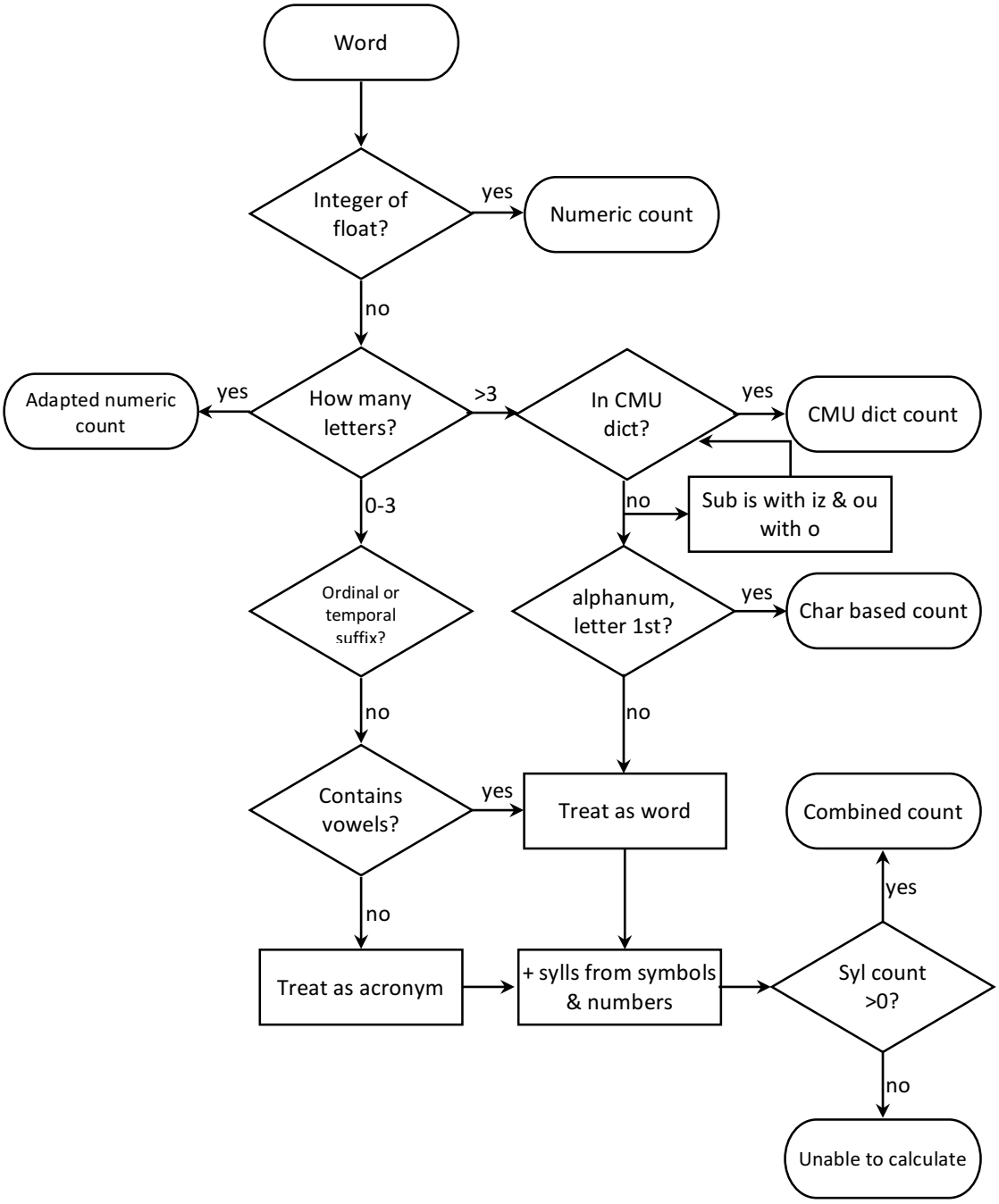

Figure 4.2.a: Flow chart of the bespoke syllable calculation process

Words containing fewer than 4 letters and those not found within the CMU dictionary, including incorrectly spelled words, numbers, veterinary neologisms and abbreviations, were dealt with using a series of bespoke Python methods encoding rules within regular expressions. Based on the conventions outlined by Flesch (Flesch 1948) that syllables should be counted with respect to the way a word would be pronounced if spoken: if the word was an integer or float the syllable count was based on integers below 100 being read as complete words

and those above 100 being read as a list of integers. Floats were also treated as a list of integers with the decimal point adding an additional syllable.

If the word contained more than three letters, ignoring a final ‘s’, and was an alphanumeric code beginning with a letter, it was expected to be read out loud as individual characters, and syllable count was calculated as such. If the word was not an alphanumeric code the syllable count was calculated as a word (Figure 4.2.b)

```
def sylsLongWord(self, word):
    sylCount = 0
    word = re.sub('''(?<!'''+cs+'''
        l)(?:(?<![bdfgtd])(?<!^[a-z]{2})(?<!\W[a-z]{2})ed|
        (?<!\W[a-z]{2})(?<!g)es)$|e$''','', word, flags = re.I|re.VERBOSE)
    if re.search(vs,word):
        sylCount += len(re.findall(vs+'{1,3}',word))
        sylCount += len(re.findall('(?<![tc])io|eum|iet|ia|^ree|yo',word))
    else:
        sylCount = len(re.findall(cs,word))
    sylCount += len(re.findall('(?<!'+vs+')'+vs+'y?ing$',word))
    if sylCount >1:
        sylCount -= len(re.findall(
            'ely$|(?<!^[a-z])(?<!\W[a-z])\wre(?!'+vs+')',word))
    sylCount += len(re.findall(vs+'sn|dn', word))
    return sylCount
```

Figure 4.2.b: Method used to calculate syllables within a word with 4 or more letters, provided it was not an alphanumeric code beginning with a letter. The variables vs and cs represent vowels including y and consonants respectively.

Three letter abbreviations were often overloaded with more than one meaning as an abbreviation, their letter sequences often also formed a real word. The CMU dictionary convention was to expand abbreviations to their constituent words and return the syllable count of those words. This required mitigation to avoid the problem of the number of syllables in an expanded abbreviation of a non-veterinary word or phrase being returned in lieu of the veterinary use, consequently words of three or fewer letters were not passed to the CMU dictionary. Instead the syllable count was derived as an approximation that matched short words, and abbreviations which were often so common as to be considered words within the clinical lexicon, and in addition temporal and ordinal short hand (Figure 4.2.c).

```
def sylsAbbrvShortWord(self, word):
    sylCount = 0
    if re.search('\d+(?:hr|dy|wk|mo|yr|[dwmy]|nd)$', word):
        sylCount = 1
    elif re.search(vs,word):
        sylCount += len(re.findall(vs+'{1,3}',word))
        sylCount += len(re.findall('(?<![tc])io|eum|iet|ia|^ree|yo',word))
        if sylCount >1:
            if re.search('e$',word):
                sylCount -= 1
    elif not re.search('\d(?:st|rd|th)$',word):
        sylCount =  len(re.findall('[a-vx-z]',
                    re.sub('(?<!c)s$','',word)))+len(re.findall('w',word))*3
    return sylCount
```

Figure 4.2.c: Method used to calculate syllables with a word of 3 or fewer letters. This method was designed to account for two and three letter abbreviations.

The final step was to add the appropriate number of syllables contributed by numbers and symbols. If a word reached the end of this process and still had a zero syllable count it was returned as a missing value and the word was omitted from calculations.

4.2.5 Sentence length

In view of the nature of the text being examined, a bespoke sentence tokenisation function was used to split narratives into their component sentences (Figure 4.2.d). This was necessary because methods such as NLTK library's sent_tokenize() method (NLTK Project 2015) relied to a degree on normal grammar and capitalisation, which was not likely to be reliable within the veterinary clinical narrative (NLTK Project 2015).

```
def sentTokenizeJN(textString):
    textString = re.sub('"','',textString)
    sentenceS = re.findall('(.+?(?:[\.!?](?=\s|$)|$))', textString)
    sentenceS = [re.sub('^[\.]?\s*','',
                    sentence.strip()) for sentence in sentenceS]
    return sentenceS
```

Figure 4.2.d: Bespoke function to parse clinical narrative into component sentences independent of capitalisation.

4.2.6 Numeric content

The proportion of consultations containing a float or integer and the volume of floats and integers, where they were present, was assessed using regular expressions. As numbers in isolation carry little clinical meaning, it was likely that their presence would impact the word and syllable count to a greater extent

than their own contribution, and perhaps also influence the structure of the narrative. This was assessed and the metrics previously evaluated were reviewed with further stratification into those consultations containing at least one float or integer, and those without.

4.2.7 Capitalisation

Regular expressions were used to establish whether each sentence within a narrative began with a capital letter or a number, which couldn't be capitalised. Descriptive statistics were calculated for all sentences within a narrative, excluding those beginning with a number, and for all sentences except for the opening sentence which had been subjectively observed to differ from later sentences with the inclusion of capitalised clinician initials and opening statements. Where the first sentence was excluded this also excluded those narratives that consisted of only one sentence.

4.2.8 Lexical diversity

Lexical diversity was calculated using the method developed by Mueller and as described by Maas (Maas 1972) and implemented using a bespoke function (Figure 4.2.e). This method was chosen as it accounted for the effect of the length of the text on diversity. A value, termed Maas a, was generated for each narrative record reflecting the lexical richness of the language used within it, the greater the value the more often words were repeated within the text (Figure 4.2.e part c).

a) $a^2 = \frac{\log n - \log v}{(\log n)^2}$

b)
```
def maasIndex(n,v):
    return np.sqrt(np.log(n/v)/np.square(np.log(n)))

n is the length of text, number of words or tokens
v is the vocabulary size of the text, number of unique words     or
tokens
```

c)
```
Example values for a text of 100 words:
        100 unique words (n=100, v=100), a = 0
        90 unique words (n=100, v=90), a = 0.07
        50 unique words (n=100, v=50), a = 0.18
        1 unique word (n=100, v=1), a = 0.47
```

Figure 4.2.e: Maas formula used to calculate lexical diversity (a) and the equivalent Python function (b). With example values as a guide to interpretation (c).

Given the nature of the language used it was not felt appropriate to stem words prior to this analysis. The words derived from tokenisation using the [a-z]+ regular expression were used to control for variations in symbolisation of the same word and remove the effect of numeric content.

4.2.9 Statistical analysis

Descriptive statistics of the above measures were calculated in a univariable manner with stratification by the species of animal to which a consultation narrative referred and whether the consultation had occurred in- or out-of-hours. As the distributions were right skewed, logarithmic transformation was used in an attempt to transform to a normal distribution and improve homoscedasticity. Minor deviations from normality were tolerated due to the central limit theorem. An analysis of variance (ANOVA) test was performed to establish equality between the species groups with regard the parameter being investigated. Where there was a significant difference, at the α = 0.05 level, pairwise Students two sample t test was used to test the hypothesis that there was no difference in the means of the species groups. A Bonferroni correction was used to account for the multiple pairwise testing, $\alpha_{\text{critical}} = 1 - (1 - \alpha_{\text{adjusted}})$ giving an α_{adjusted} of 0.0047. Confidence intervals were used to describe the uncertainty in the point estimate of proportions and means, these were calculated using the Normal Approximation to calculate the standard error.

Python's SciPy.stats library was used to perform the ANOVA and t-tests, stratification and within group calculations were performed using the data handling functionality of Pandas.

4.2.10 Methodology for assessing use of language

4.2.10.1 Comparison of word frequency to standard English

The most common 1000 and 10,000 words within the exploratory corpus was compared to the most common words within the British National Corpus (BNC) the proportion of words shared between the common words of the two corpora was calculated. The BNC is a collection of samples of written and spoken British English language from the late 20th century. The corpus comprises 100,000,000 words of narrative text from a wide range of sources including newspaper articles, academic literature and essays (BNC Consortium 2007).

4.2.10.2 Evaluation of the use of language

The semantic nature of the most common 1000 non-numeric words was examined and tabulated in addition to their frequency of abbreviation and overloading. Word and ngram frequencies within the exploratory corpus were examined and described using string processing and data handling methods in Pandas.

A random sample of 100 narratives drawn from the exploratory corpus, was manually read by the author and annotated for the presence and nature of abbreviations and key thematic features based on clinical experience and familiarity with the domain. The manner of notation used in documenting common themes was explored by examining the concordance of phrases. A bespoke method, `regexConcordance()` was developed. This was largely analogous to Python's Natural Language Tool Kit `nltk.text.concordance()` method (NLTK Project 2015), with additional functionality to permit exploration of the concordance of a regular expression, rather than a stemmed word, within a series of individual narratives. The bare bones of the code used can be found in Figure 4.2.f. and an example of output using a basic regular expression intended to identify instances of diarrhoea in Figure 4.2.g.

```
import pandas as pd
import re, textwrap
pd.options.mode.chained_assignment = None
wrap = textwrap.TextWrapper(width = 60)

class regexConcordance(object):
    def __init__(self, df, narrative = 'narrative'):
        self.df = df
        self.width = 40
        self.num = 10
        self.narrative = narrative
        self.prior = '(?<![a-z0-9])'
        self.aft ='(?![a-z0-9])'

    def findItLoop(self, returnDf = False):
        regexString = ''
        while regexString != 'quit':
            while regexString.strip() == '':
                regexString = input('Enter regular expression: ')
                if regexString.strip().lower() == 'quit':
                    break
                try:
                    regex = re.compile(
                       self.prior+'(?:'+regexString+')'+self.aft, flags = re.I)
                    self.df['regexMatch'] = self.df[
                                            self.narrative].str.contains(regex,
                                            na = False).astype(int)
                    regexMatches = self.df[self.df['regexMatch']==1]
                    print('{:,}'.format(regexMatches.shape[0]),
                        round(regexMatches.shape[0]/self.df.shape[0]*100, 2),
                        '''% narratives contain a matching string
                        ''')
                    if self.num<regexMatches.shape[0]:
                        regexMatches = regexMatches.sample(self.num)
                    for narr in regexMatches[self.narrative]:
                        for matchingString in regex.finditer(narr):
                            if matchingString.start()-self.width<0:
                                start = 0
                            else:
                                start = matchingString.start()-self.width
                            if start == 0:
                                addOn = 0-(matchingString.start()-self.width)
                            else:
                                addOn = 0
                            if matchingString.end()+self.width>len(narr):
                                end = -1
                            else:
                                end = matchingString.end()+self.width
                            print(' '*addOn+narr[start:end])
                        print('\n')
                except Exception as e:
                    print('That didn\'t work', e)
                regexString = ''
        if returnDf:
            return self.df
```

Figure 4.2.f: regexConcordance.py. Bare bones of the method designed to allow exploration of concordance of a regular expression within a series of strings held within a pandas dataframe.

```
Enter regular expression: d[+]|(?<![a-uw-z])dia(?![mtg])|loos|slop
15,802 9.68 % narratives contain a matching string

in herself, sleeping alot. no vomitting/diarrhoea. no change in her condition. cann

eave food,if taken out are on a lead,no diarr noted,,admit for sc fluids 2x 20mls w

t gastrointestinal diet. No vomiting or diarrhoea was noticed recently by his Owner

coat good.  in good condition. needs to loose some weight aim for 5.5kg.

 parrafin. Advised caution as can cause D+, no palpable obstruction felt but is oe

M M salmon pink and CRT < 3 secs. No V+/D+ today. DUDE. Urination abnormal. Urinat
ood X1 day for 5 days and to stop if V+/D+. Spinal palpation unremarkable.

       "OR no problems, EDUD fine no V+ D+ coughign sneezing. phsyical exam NAD, g

                  "RE EXAM EARS/TEMP. no d+, managing ears with sid cleaning and dr

 today ,still eating, u/F as normal, no d+ seen. On exam today is v bright and ale

lmon pink and CRT < 3 secs. DUDE. No V+/D+. HR 80. Rr 16. Normal aus. NAD abdomina
```

Figure 4.2.g: Example of the use and output of the regexConcordance() method. The string matching the input regular expression is centred to facilitate ready visualisation of the concordant phraseology. Text coloured green matched the regular expression, blue and purple highlights the adjacent context.

Where words or phrases of interest were identified, the proportion of narratives in which they occurred within the 163,240 consultation narratives of the exploratory corpus was established using Pandas string methods. Where there was more than one meaning for a term, overloading, the extent of this and the relative semantic frequency, the relative frequency with which a term carries a given meaning, within this corpus was established by reading a sample of 100 concordant narrative strings and where necessary the narratives as a whole. The use of abbreviations and symbols within the clinical narrative was examined using the same method. The nature of language usage and the regular expressions used in identifying it were tabulated alongside regular expression matching frequency.

4.2.11 Estimation of vocabulary size

In estimating the size of the vocabulary used within the veterinary clinical narrative only sequences of letters were used. This avoided erroneously amplifying the apparent vocabulary size by the inclusion of numbers, and concatenations including numbers and symbols, with a limited range of clinical meanings but a multitude of variations.

The narratives were transformed to lower case and words extracted into a Pandas series using the regular expression `[a-z]+` , the `pandas.Series.value_counts()` method was then used to generate a word frequency dataframe. Pandas sample and slice criteria were used to draw a sample of words occurring at periodic frequencies within the exploratory corpus. Where there were sufficient in a given frequency range a sample of 500 words was selected. Where there were less than 500 words meeting the criteria, all words were selected.

In this manner samples of those words occurring once, twice, ten times, and at incrementally greater frequencies were selected from the dataframe of unique words present within the corpus. Word samples were manually read and words that contained spelling or typographic errors identified. Where words were veterinary neologisms, contractions or abbreviations, these were considered correctly spelled. Where there was doubt as to correct spelling, online dictionaries were checked. Where there was uncertainty as to whether the word was a neologism, abbreviation or misspelling the context in which it had been used was examined by referring back to the exploratory corpus and extracting the narratives in which the word had been used via the `regexConcordance()` method.

The frequency of spelling errors at each reference point and the apparent number of words occurring between the reference point and the preceding point were used to calculate the true vocabulary size having accounted for spelling errors. The Wilson method (Wilson 1927) was used to estimate confidence intervals as it better dealt with proportions at or approaching zero.

4.3 Results 1: Exploratory corpus & sentence metrics

4.3.1 Description of the exploratory corpus

At the time of writing the SAVSNET dataset contained 2,330,373 consultation narratives, 71,854 narratives were excluded because the animal's species had not been documented, a further 4,563 narratives contained a software generated checklist and were excluded. The exploratory corpus was generated from the remaining 2,253,956 consultation narratives regarding 901,129 animals of forty species.

The consultation narratives were contributed by 391 veterinary clinics. Although the most recently joining practice had been contributing for over 7 months, the period of contribution (days between oldest and most recent collated record) ranged from 1 day to 3 years, with a median 4,124 narratives collated per clinic, range 10 to 32,998 narratives. There were a small number of clinics that had contributed either for a very short period of time or very few consultations, or both.

The 0.05 quantile of contribution by a clinic to the SAVSNET dataset was 440 narratives/clinic, i.e. 95% of the clinics had contributed at least this number. The narratives of 371 clinics were included in the exploratory corpus. Among these clinics the median period of contribution was 576 days, range 45 to 1,149 days. Mean contribution of a clinic per week ranged from 5.5 to 328.2 narratives, median contribution 57.2 narratives per week per clinic.

Sampling 440 narratives from each clinic that had contributed at least this number (the 0.05 quantile threshold of contribution) created a corpus of 163,240 narratives, containing consultation narratives regarding 141,428 animals. The majority of these consultations occurred in-hours, with 2.73% (n=4,463) categorised as out-of-hours consultations.

Cats (28.94%), dogs (67.84%) and rabbits (1.75%) accounted for the vast majority of consultations. The remaining 1.47% comprised 2,393 consultations and consisted of consultations regarding an array of small mammals (68.62%), birds (16.59%), reptiles (8.78%), large and farm animals (1.71%), equine (0.71%), and small numbers of wild animals, fish, insects and unspecified other exotic species (3.55%).

4.3.2 Sentence metrics

A typical (median) narrative contained five sentences of seven four-letter words, a total of 41 words and 66 syllables (Table 4.3.a). As consultation records regarding dogs formed 68% of the population, the distribution and any associations within the dog consultation sub-corpus heavily influenced that of the overall population.

Table 4.3.a: Descriptive metrics of narratives in the overall corpus.

	Mean (95% CI)	Range	Median	IQ range
Narrative				
Sentences	6.15 (6.13, 6.18)	1- 82	5	3- 8
Words	53.16 (52.94, 53.39)	1- 835	41	20- 73
Syllables	86.27 (85.91, 86.64)	1- 1,653	66	32- 118
Lexical diversity	0.073 (0.0727, 0.0732)	0- 1.2	0.084	0.048-0.103
Sentence				
Words	9.46 (9.42, 9.49)	0.5- 269	7	5- 10
Syllables	15.28 (15.22, 15.34)	0- 449	11	7.5- 16
Word				
Syllables	1.655 (1.653, 1.657)	1-18	1	1
Characters	4.581 (4.578, 4.584)	1- 15	4	4

4.3.3 Word count

Word count ranged from one to 835 words, mean word count was 53.16 (95% CI: 52.94, 53.39). The median word count, for both definitions of a word, was 77% of the mean, indicating a considerable positive skew to the distribution of word counts. The median clinically meaningful word count was 41 (IQ range: 20, 73) (Table 4.3.b). There was a significant effect of species group on word count of consultations at the $\alpha = 0.05$ level, $F(_{3,\ 163235}) = 14.11$, $p<0.001$.

Table 4.3.b: Word counts within the exploratory corpus using three definitions of a word.

		Word count within each narrative			
Word definition	**Regular expression**	**Range**	**Mean (95% CI)**	**Median**	**IQ range**
Clinically-meaningful characters	(?:[\d]+|(?<=\s))[\.]\d+|[a-z0-9<>\^+~\']+(?:\-(?=\W|$))?	1- 835	53.16 (52.94, 53.39)	41	20- 73
Letters only	[a-z]+	0- 811	51.89 (51.67, 52.11)	40	19- 71

The word count for consultations regarding rabbits was lower than that of each of the other species group, with a mean of 49.21 (95% CI: 47.57, 50.84) words per narrative compared to 53.46 (95% CI: 53.19, 53.74), 52.75 (95% CI: 52.33, 53.16) and 52.26 (95% CI: 50.46, 54.06) words for dogs, cats and uncommon species respectively (p <0.001) (Table 4.3.c).

With the exception of the uncommon species group (out-of-hours group n=93) there was a significant difference in the word count between in and out-of-hours consultation narratives at the alpha = 0.05 level, with a mean of 13 more words used when documenting out-of-hours consultations (p <0.001), out-of-hours mean word count was 66.27 (95% CI: 64.67, 67.87) compared to 52.8 (95% CI: 52.57, 53.02) in-hours (Figure 4.3.a).

Table 4.3.c: Descriptive metrics stratified by species

	Dog Mean (95% CI)	Median	IQ range	Cat Mean (95% CI)	Median	IQ range	Rabbit Mean (95% CI)	Median	IQ range	Uncommon Mean (95% CI)	Median	IQ range
Narrative												
Sentences	6.12 (6.1, 6.15)	5	3- 8	6.27 (6.22, 6.31)	5	3- 8	5.73 (5.56, 5.89)	5	2- 8	5.66 (5.47, 5.85)	4	2- 8
Words	53.46 (53.19, 53.74)	41	20- 73	52.75 (52.33, 53.16)	41	20- 72	49.21 (47.57, 50.84)	37	16- 69	52.26 (50.46, 54.06)	41	19- 71
Syllables	86.76 (86.31, 87.21)	66	33- 118	85.76 (85.08, 86.44)	66	32- 117	78.59 (76, 81.17)	59	27- 110	83.1304 (80.26, 86)	64	30- 114
Sentence												
Words	9.55 (9.51, 9.59)	7	5- 10	9.21 (9.14, 9.27)	6.5	4.5- 10	9.06 (8.81, 9.3)	6.5	4.5- 10	10.51 (10.17, 10.84)	8	5- 11
Syllables	15.43 (15.36, 15.5)	11	8- 16.5	14.91 (14.81, 15.02)	11	7- 16	14.49 (14.11, 14.87)	11	7- 16	16.61 (16.09, 17.13)	12	8- 18
Word												
Syllables	1.656 (1.654, 1.658)	1	1	1.657 (1.654, 1.66)	1	1	1.637 (1.623, 1.65)	1	1	1.602 (1.591, 1.613)	1	1
Characters	4.57 (4.566, 4.574)	4	4	4.598 (4.592, 4.604)	4	4	4.66 (4.64, 4.687)	4	4- 4.5	4.661 (4.637, 4.685)	4	4- 4.5

Table 4.3.d: Descriptive metrics stratified by timing of consultation

	In-hours Mean (95% CI)	Median	IQ range	Out-of-hours Mean (95% CI)	Median	IQ range
Narrative						
Sentences	6.12 (6.09, 6.14)	5	3- 8	7.43 (7.26, 7.61)	6	3- 10
Words	52.8 (52.57, 53.02)	40	20- 72	66.27 (64.67, 67.87)	53	28- 90
Syllables	85.66 (85.29, 86.03)	65	32- 117	108 (105.31, 110.7)	86	45- 145
Sentence						
Words	9.44 (9.41, 9.48)	7	5- 10	9.98 (9.78, 10.18)	7	5- 10.5
Syllables	15.26 (15.2, 15.31)	11	7.5- 16	16.14 (15.82, 16.45)	12	8- 17
Word						
Syllables	1.655 (1.653, 1.657)	1	1	1.654 (1.644, 1.664)	1	1
Characters	4.581 (4.578, 4.584)	4	4	4.589 (4.572, 4.606)	4	4

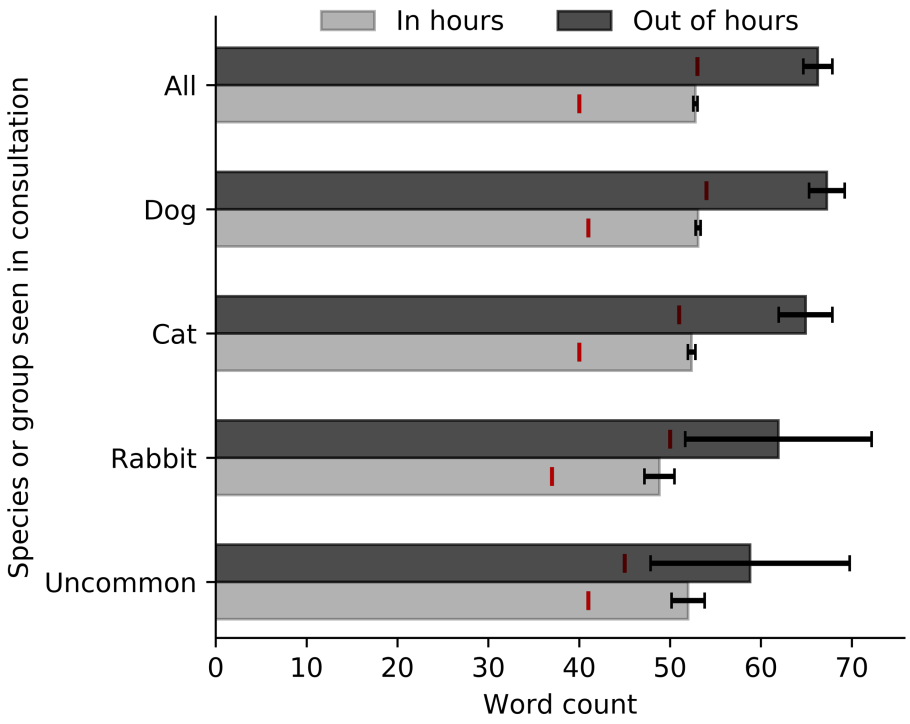

Figure 4.3.a: Bar plot of the mean word count within consultation narratives with stratification by the timing of consultation and species examined. The red bar represents the median and the error bars the 95% confidence limits of the mean.

The mean narrative word count of a veterinary clinic created a right skewed distribution (Figure 4.3.b). The upper quartile of the mean clinic word count was 66.74 words, median 49.88 words.

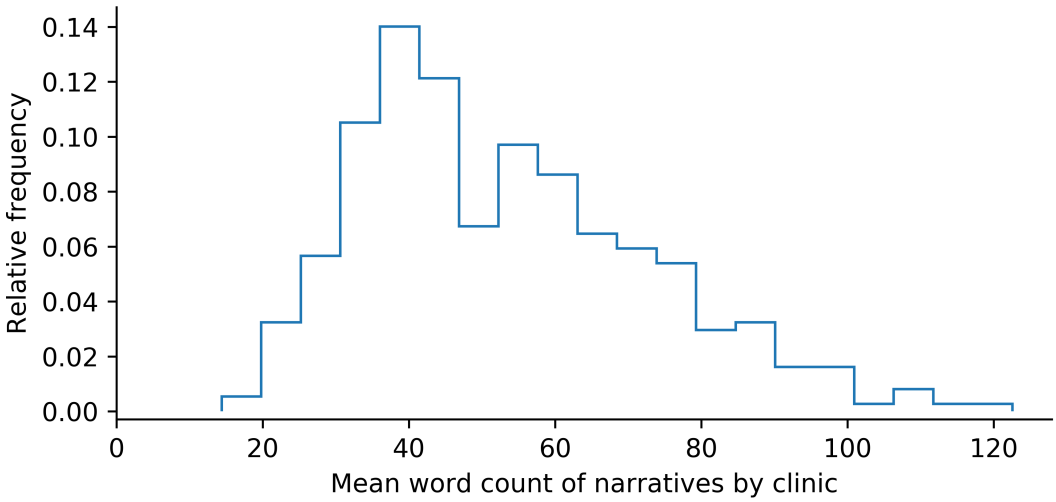

Figure 4.3.b Relative frequency histogram of the mean word count of clinical narratives collated from a veterinary clinic.

4.3.4 Word length and complexity

The mean length of a word in the veterinary narrative was 4.581 (95% CI: 4.578, 4.584) characters and 1.655 (95% CI: 1.653, 1.657) syllables. There was a significant effect of species group on word length at the $\alpha = 0.05$ level, $F_{(3,\ 163235)} = 55.55$, $p<0.001$.

The words used in recording consultations regarding dogs were slightly shorter than those used for other species at 4.57 (95% CI: 4.566, 4.574) characters compared to the next shortest which was consultations regarding cats at 4.598 (95% CI: 4.592, 4.604) characters ($p<0.001$).

On average a word consisted of 1.655 (95% CI: 1.653, 1.657) syllables. There was a significant effect of species group on the mean number of syllables in each word at the $\alpha = 0.05$ level, $F_{(3,\ 163235)} = 29.0$, $p<0.001$. The consultations of uncommon species were documented using words containing fewer syllables than those for the other species groups ($p < 0.001$). There was no difference between the length or complexity of words used in documenting consultations that had occurred in- and out-of-hours.

As might be anticipated a similar relationship was seen in the total syllable count with consultations regarding rabbits having a lower syllable count than consultations regarding dogs or cats (p <0.001). Rabbit consultations contained a median 59 syllables whilst those of cats and dogs contained 66 and uncommon species 64 (Figure 4.3.c).

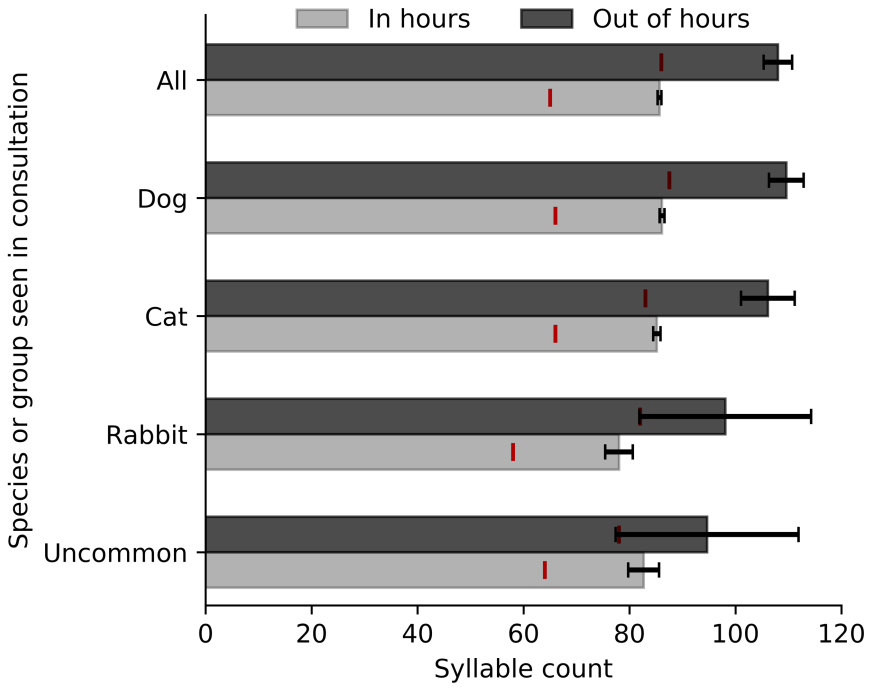

Figure 4.3.c: Bar plot of the mean syllable count within consultation narratives with stratification by the timing of consultation and species examined. The red bar represents the median and the error bars the 95% confidence limits of the mean.

4.3.5 Sentences

The mean sentence length was 9.46 (95% CI: 9.42, 9.49) words and 15.28 (95% CI: 15.22, 15.34) syllables, with 6.15 (95% CI: 6.13, 6.18) sentences in each narrative. There was a significant effect of species group on word length at the α = 0.05 level, $F(_{3,\ 163235}) = 66.83$, p<0.001. Uncommon species and dogs had consultations with the longest sentences (p <0.001) at a mean length of 10.25 (95% CI: 9.96, 10.54) and 9.59 (95% CI: 9.55, 9.64) words respectively compared to 9.15(95% CI: 9.09, 9.21) words for consultations regarding cats and 9.26 (95% CI: 9.02, 9.49) words for consultations regarding rabbits.

Consultations occurring out-of-hours comprised longer sentences than those occurring in hours, this was a significant difference in consultations regarding

dogs and cats ($p < 0.001$). The overall difference was small with a mean sentence length of 9.46 (95% CI: 9.43, 9.5) words for in hours consultations compared to 9.85 (95% CI 9.62, 10.07) words in out-of-hours consultations.

The distribution of mean sentence length from individual clinics was right skewed with a median of 8.91 words per sentence, upper quartile of 10.49 words and maximum of 31.04 words (Figure 4.3.d)

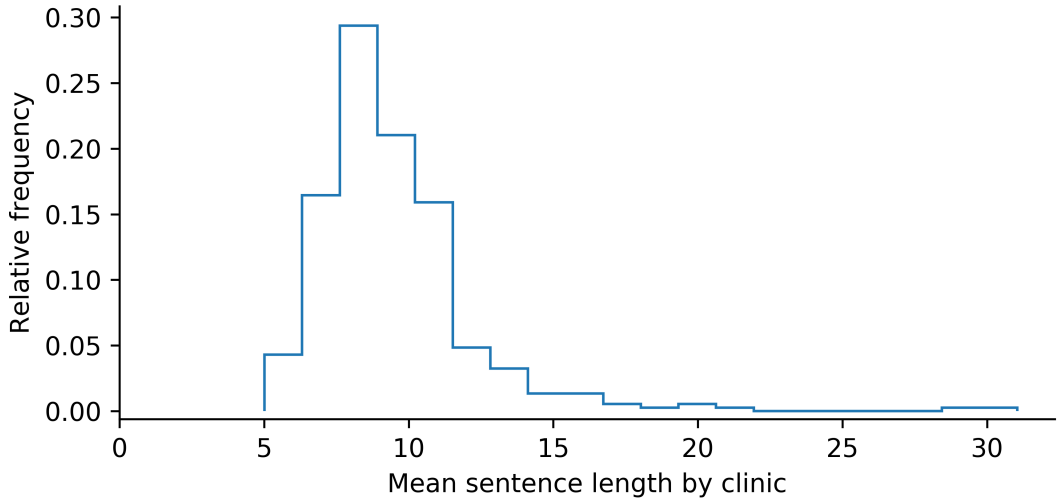

Figure 4.3.d Relative frequency histogram demonstrating the distribution of mean sentence length by veterinary clinic.

4.3.6 Numeric content

Overall 69.94 (69.72, 70.16)% of consultations contained at least one float or integer. There was a significant effect of species group on the likelihood of a narrative record including numeric content at the $\alpha = 0.05$ level, $F(_{3,\ 163235}) = 135.82$, $p<0.001$. There was a clear dichotomy in the likelihood of consultation narratives containing a number with regard the species group that the consultation related (Table 4.3.e). Narratives regarding the veterinary management of cats and dogs had an odds ratio of 1.75 (95% CI: 1.65,1.85) for including at least one integer or float compared to the other two species groups.

Table 4.3.e: Numeric content of narrative field

	Narratives containing a number	Numbers present		
	Proportion of species corpus (95% CI)	Mean (95% CI)	Median	IQ range
Dog	70.51 (70.24, 70.78)%	2.38 (2.36, 2.39)	1	0- 3
Cat	69.99 (69.58, 70.4)%	2.34 (2.31, 2.37)	2	0- 4
Rabbit	59.22 (57.42, 61.03)%	1.47 (1.39, 1.54)	1	0- 2
Uncommon	55.62 (53.63, 57.61)%	1.35 (1.27, 1.42)	1	0- 2

Where at least one number was present, there were considerably more numbers within the narratives of dogs and cats compared to those of rabbits and uncommon species, mean 3.36 (95% CI: 3.34, 3.38) floats and integers compared to 2.45 (95% CI: 2.38, 2.52) respectively, median 2 for both.

Re-examining the number of syllables within the narrative whilst further stratifying for whether consultations contained a number showed that the differences between in- and out-of-hours consultations persist within the strata, despite the consequent reduction in size of each stratum (Table 4.3.f)

Table 4.3.f: Syllable length of narratives stratified by the presence of numeric content and timing of the consultation.

Species group	Proportion of corpus (n)	Mean (95% CI)	Median	IQ range
Consultations containing at least one number				
In-hours	97.16% (110,925)	102.95 (102.48, 103.42)	83	47-138
Out-of-hours	2.84% (3,247)	126.92 (123.61, 130.23)	105	61- 168
Consultations with no numeric content				
In-hours	97.52% (47,851)	45.6 (45.21, 45.99)	33	15- 62
Out-of-hours	2.48% (1,216)	57.5 (54.59, 60.4)	44	19- 79

4.3.7 Capitalisation

Ignoring the 2.2% of sentences that began with a number, 57.6 (95% CI: 57.49, 57.68)% of sentences began with a capital letter. When the first sentence, and thus also those narratives containing only one sentence, was ignored this rose to 58.81 (95% CI: 58.56, 59.07)% of sentences.

There was a significant effect of species group on the likelihood of sentences beginning with a capital letter at the $\alpha = 0.05$ level, $F(_{3,\ 163235}) = 12.34$, $p<0.001$. There were notably fewer sentences beginning with a capital letter in the narratives of consultations regarding the uncommon species group, median 50% compared to a median of 66.67% for dogs, cats and rabbits (p <0.001)

There was no difference in capitalisation of sentences in relation to whether the consultation occurred out-of-hours.

4.3.8 Lexical diversity within clinical narratives

Mean intra-narrative lexical diversity over the whole exploratory corpus was 0.073 (0.0727, 0.0732) median 0.084 (IQ range: 0.049, 0.103). As demonstrated in Figure 4.2.e less word repetition and therefore a greater range of words is associated with a lower value for Maas' a statistic.

There was a significant effect of species group on the lexical richness of narrative records at the $\alpha = 0.05$ level, $F_{(3,\ 163235)} = 165.42$, $p<0.001$. Interpretation of this measure is challenging as the distribution of Maas diversity (Figure 4.3.e) was bimodal by virtue of a significant minority of consultation narratives having a diversity of zero, i.e. all words within the narrative occur only once, this was less common in consultations regarding dogs than other species (p <0.001) and may have in part been responsible for the apparently less rich language used in consultations regarding dogs compared to those regarding any other species group (p <0.001) (Table 4.3.g).

Table 4.3.g: Maas within narrative lexical diversity (a), narratives consisting of a single word excluded.

Species	Narrative count	Mean (95% CI)	Median	Proportion of narratives where a = 0.00 (% & 95% CI)
Overall	162,377	0.073 (0.0727, 0.0732)	0.084	23.23 (23.02, 23.43)
Dog	24,880	0.0748 (0.0745, 0.0751)	0.0867	22.47 (22.22, 22.71)
Cat	11,689	0.0691 (0.0686, 0.0695)	0.0817	24.74 (24.35, 25.13)
Rabbit	756	0.0697 (0.0679, 0.0715)	0.0827	26.46 (24.84, 28.08)
Uncommon	591	0.0699 (0.0681, 0.0717)	0.0843	24.7 (22.9, 26.42)

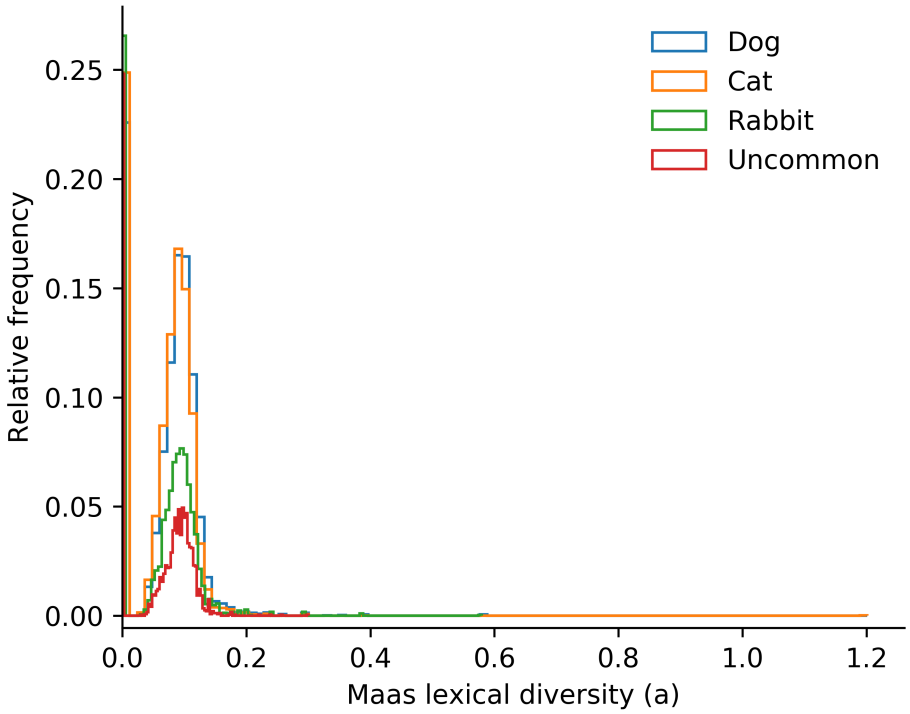

Figure 4.3.e: Maas lexical diversity within narratives, stratified by the species being seen. A quarter of narratives have a Maas a of zero creating a bimodal distribution.

Post hoc analysis of those narratives comprising over 20 words (the lower quartile of the narrative word count) demonstrated that the apparent difference between cats and other species persisted except for the uncommon species group, and in this reanalysis consultations regarding cats had notably richer narratives (ie. a lower value of Maas a) than the other species groups ($p < 0.01$) with the mean for cats 0.0834 (95% CI: 0.083, 0.0837) compared to 0.0875 (95% CI: 0.0873, 0.0877) for dog consultations, 0.0857 (95% CI: 0.0844, 0.087) for rabbits and 0.0871 (95% CI: 0.0859, 0.0884) uncommon species. Given the distribution that this adjustment to the data generated (Figure 4.3.f) this appeared a more appropriate interpretation although should be viewed with consideration to the risks of post hoc reinterpretation.

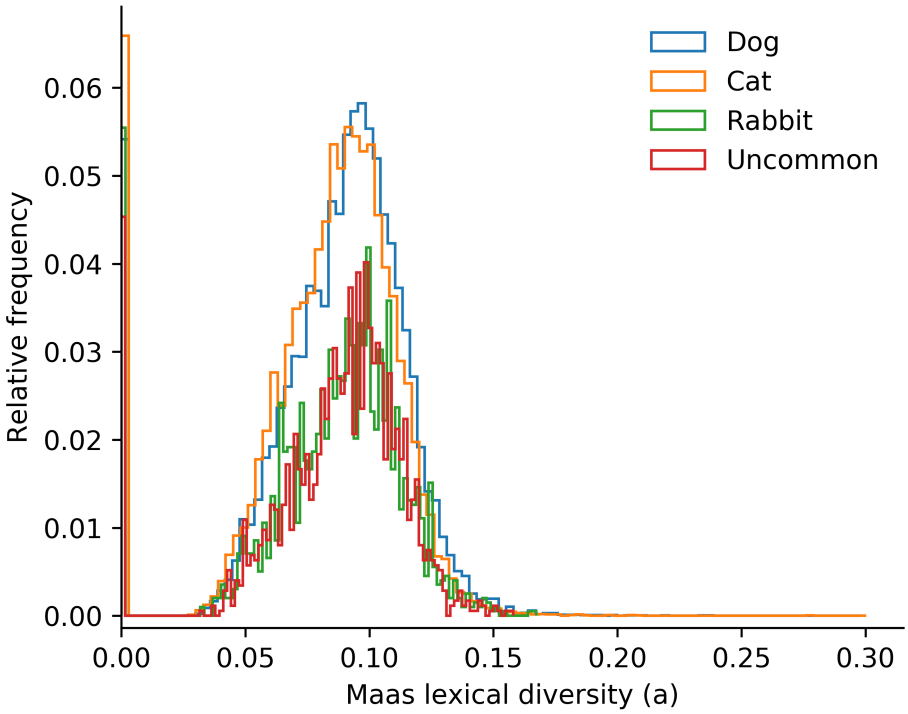

Figure 4.3.f Maas lexical diversity within narratives, stratified by species being seen. Narratives comprising greater than 20 words, this reduced the peak at zero to 1 in 20 narratives.

There was more word repetition, and thus less diversity, in consultations occurring out-of-hours: within narrative mean lexical diversity 0.0788(0.0776, 0.08), compared to those occurring in-hours where the mean was 0.0727(0.0725, 0.073), (p <0.001).

4.4 Results 2: Use of language within the veterinary clinical narrative

For ease of reading phrases and abbreviations in the following sections have been written in italics and not quotation marks, percentages given without a confidence interval refer to the proportion of narratives within the exploratory corpus that contain the preceding phrase, the confidence interval is included in the respective tables. The regular expression used to explore the corpus are provided in tables following each subsection. Where a confidence interval or other form of likelihood has been cited within the text this indicates the estimate was derived from examination of concordance of the phrase in a sample of its uses within the exploratory corpus.

4.4.1 Comparison to standard English word frequencies

On comparing the most common 1,000 words of the exploratory corpus and the British National Corpus 3,310 (33.1%) were shared, a similar proportion (34.91%) of the most common 10,000 words were shared between the two corpora.

4.4.2 Semantic type of common words

Having discounted words that were solely floats or integers, 182 of the most common 1000 words in the exploratory corpus were clinical signs or findings from history or examination, 118 words provided anatomical information,105 words conveyed objective or relational temporal information and 20 the trajectory of illness.

4.4.3 N-gram frequency

The most common bi- tri- and quad grams referred to follow-up arrangements (Table 4.4.a). Discounting these and examining the next most frequent demonstrated subtle differences between the species groups (Table 4.4.b & 4.4.c).

Table 4.4.a: Most common ten ngrams across whole exploratory corpus. Numeric content has been replaced by *n*.

Bigrams	Trigrams	Quadgrams
in *n*	next appointment in	next appointment in *n*
n n	appointment in *n*	appointment in *n* weeks
n weeks	nothing abnormal detected	appointment in *n* week
next appointment	in *n* weeks	appointment in *n* days
appointment in	bcs *n n*	bright alert and responsive
n days	in *n* week	nothing abnormal detected on
nothing abnormal	in *n* days	defecating urinating drinking eating
abnormal detected	alert and responsive	mm pink and moist
at home	n n n	lab request references generated
n week	pink and moist	exam nothing abnormal detected

Table 4.4.b: Most common bigrams by species group, excluding those relating to follow up arrangements.

Dog	Cat	Rabbit	Uncommon
nothing abnormal	nothing abnormal	at home	to be
abnormal detected	abnormal detected	nothing abnormal	has been
at home	at home	abnormal detected	for n
has been	no concerns	no concerns	few days
no concerns	hr n	on exam	not eating
in for	abdo palp	has been	abnormal detected
to be	n kg	myxo rhd	nothing abnormal
abdo palp	has been	in for	on exam
nad on	in for	abdo palp	if no
hr n	bcs n	a little	need to

Table 4.4.c: Most common quadgrams by species group, excluding those related to follow-up arrangements.

Dog	Cat	Rabbit	Uncommon
bright alert and responsive	bright alert and responsive	bright alert and responsive	bright alert and responsive
nothing abnormal detected on	nothing abnormal detected on	defecating urinating drinking eating	been in owners possesion
defecating urinating drinking eating	defecating urinating drinking eating	nothing abnormal detected on	body condition score n
mm pink and moist	mm pink and moist	and responsive defecating urinating	condition score n n
lab request references generated	lab request references generated	palp nothing abnormal detected	defecating urinating drinking eating
exam nothing abnormal detected	exam nothing abnormal detected	alert and responsive defecating	for a few days
v d c s	no v d c	responsive defecating urinating drinking	eating and drinking well
o has no concerns	v d c s	o has no concerns	not eating as much
no v d c	o has no concerns	exam nothing abnormal detected	palp nothing abnormal detected
mucous membranes pink and	palp nothing abnormal detected	eating and drinking well	nothing abnormal detected on

4.4.4 Abbreviations

Of the 1000 most commonly used words, 140 were used as abbreviations within the veterinary clinical narrative and 52 (37.1%) of these were overloaded, that is they had at least two meanings within this corpus. The abbreviations included 7 words that contained or were solely composed of symbols, three of which were plus signs (+, ++ and +++).

In the manually read random sample of 100 narratives drawn from the exploratory corpus 84 (95% CI: 76.81, 91.19)% contained at least one abbreviation, with a median of 2 (IQ range: 1, 7) different abbreviations per narrative. There were 538 abbreviations in this sample, 238 unique. The number of abbreviations within an individual narrative ranged from zero to 34, with a maximum of 28 different abbreviations within a single narrative.

The most frequent abbreviation in the corpus was *o*, abbreviating *owner*, and accounting for 0.92% of all words used within the corpus. *o* was present in 31% of the narratives in the exploratory corpus, and used by all clinics. The next most commonly occurring abbreviation also related to the owner, *or* accounted for 0.55% of all words used, this was a combination of its use as the conjunction *or* and as an abbreviation for *owner reports*, commonly documented during history taking.

There were three abbreviations for antibiotics (*ab, abs, abx*) in the most common 1000 words, and 4 variations of the phrase, *eating drinking urinating defaecating* (*eddu, edud, eduf, dude*) and all of the component letters of the same. These abbreviations and their counterparts using + and / were present in 13.1% of all narratives in the corpus.

4.4.5 Use of symbols

The plus (+) sign, was present in 1 in 6 narratives within the exploratory corpus and was commonly used to indicate presence or absence of a clinical sign, to describe the magnitude of a sign and to abbreviate a physiological event. For example *d+ since yesterday* indicated that the animal has had diarrhoea since yesterday, whereas *d+u+d+e+* was used to indicate that intake and elimination of food and fluids was normal (Table 4.4.d). For the most part (81% of uses) a

single plus sign was used, however increasing numbers were used to describe increasing severity or frequency of a clinical sign.

Table 4.4.d Examples of the overloaded use of the plus symbol

Example of usage	Meaning
d+ since yesterday	Has had diarrhoea since yesterday
no d+	Does not have diarrhoea
d+u+d+e+	Drinking urinating defaecating and eating
d+++	Either lots of, severe or frequent diarrhoea

The minus sign was used as an antonym to the plus sign, *d-* to mean *no diarrhoea* for example. This was challenging to quantify as the symbol – used as both a minus sign and a hyphen was present in 44% of narratives. The regular expression `(?<!\w)[vdcs][on]?[-]`, which would find the common short hand forms of no *vomiting, no diarrhoea, no cough and no sneeze*, matched in 0.45% of narratives.

The less than sign (<) was present in 4.82% of narratives, 71.1% of the uses of the symbol were as part of the statement that capillary refill time was less than 2 or 3 seconds (variants of *CRT<2s*). The greater than sign was present in 0.73% of narratives but not used in a conventional manner, in a sample of 100 narratives containing the symbol, 12% used it to mean greater than, either in relation to a number or to indicate that one limb was affected to a greater degree than another. Other uses of the symbol included its use to indicate worsening clinical signs, as punctuation where a colon may more commonly be used in standard English, and to indicate causative pathways of disease, thought processes and management plans with *x therefore did y* represented as variants of *x* > *y*.

4.4.6 Common themes identified

Commonly noted themes were as shown in Table 4.4.e below, these reflect the clinician gathering and acting upon information through their interaction with the animal and owner. The following sections explore how these themes are documented and can be identified within the clinical narrative.

Table 4.4.e Clinical themes commonly identified within small-animal veterinary narratives

Theme
Reason for visit
Owner concerns
Lack of owner concerns
Clinical history
General
Presentation specific
Owner described signs
Social history
Examination findings
Statement of general wellbeing or illness
General examination
Presentation specific examination
Differential diagnosis
Management
Preventive medicine
Safety net

4.4.6.1 Reason for visit

The majority, 71 (95% CI: 62.1,79.9)%, of consultation narratives began with a statement summarising the reason for the animal being brought to the veterinary surgeon on that occasion. These reasons for visit could be broadly grouped into specific owner concerns relating to a new or ongoing clinical issues: review of an ongoing clinical issue; or preventative health care including vaccination, weight management, parasite prophylaxis, and post-operative review.

The level of detail contained in the reason for visit statement ranged from a full and detailed sentence; *lame left fore acute onset after walking on rough ground*, to one or two abbreviated words such as *still d+*, meaning still has diarrhoea, and *eag+++*, meaning expressed anal glands which were extremely full.

4.4.6.2 Owner concerns

The thoughts and concerns of the owner were documented within the reason for visit statement and clinical history, often specifically noted alongside other pertinent information, 4.5% of narratives contained a variant of the phrase *owner concern (2.84%), worry (0.61%) or thinks (1.97%)*. Conversely the absence of any concerns was documented in 9.1% of consultations (Table 4.4.f).

Table 4.4.f Regular expressions used to gauge the volume of contents regarding owner concerns.

Phrase	Regular expression	Regular expression matched within narrative n=	Proportion (%)
owner concerns	(?<!no\s)(?<!\w)o\w*\W*(?:conce|worr|think|tho\w*t)	7,390	4.53(4.43, 4.63)
no concerns	no\W*(?:o\w*)?\W*(?:conce|worr)	14,778	9.05(8.91, 9.19)

4.4.6.3 History

Where it was documented, clinical history was largely divided into the general history of the animal's health, their bodily functions and that of the specific reason for the current visit. Documentation of the history varied from broad statements of all being well, through lists of negative features to discursive description and documentation of the owner's account of events. Where the presence of an illness or clinical signs was documented, its frequency, severity, exacerbating and relieving factors, including previous response to treatment, were documented to varying extents.

4.4.6.3.1 Common owner described signs and functions

Key features of the general history were whether the animal was eating, drinking urinating and defaecating (2.5%) normally. Short hand was used to describe normal bodily functions *dude or d+u+d+e+* (7.2%) or its synonyms *eddu* (1.9%) and *eduf* (1.4%) were often found, and less commonly variants such as *eadrurdef* (0.03%). One of these variants occurred in 13% of consultation records (Table 4.4.g).

Table 4.4.g: Regular expressions used to gauge the volume of contents regarding an animal's normal bodily functions.

Phrase or word	Regular expression	Regular expression matched within narrative n=	Proportion (%)
eating drinking urinating defaecating	(?:(?:eat|drin|urin|def)\w*\W*(?:and|or)?\W*){4}	4100	2.51(2.44, 2.59)
dude	(?<!\w)dude(?!\w)	11708	7.17(7.05, 7.3)
dude complex	(?<!\w)d\W*u\W*d\W*e\W*(?!\w)	11818	7.24(7.11, 7.37)
eddu	(?<!\w)e\W*d\W*d\W*u\W*(?!\w)	3033	1.86(1.79, 1.92)
eduf	(?<!\w)e\W*d\W*u\W*f\W*(?!\w)	2205	1.35(1.29, 1.41)
eadrurdef	eadrurdef	49	0.03(0.02, 0.04)

The presence or absence of common clinical signs that would generally be identified in the history rather than examination was documented by some clinicians (Table 4.4.h). Where the animal did not present with a sign, the words *no*, *not*, *none* and minus symbol *(-)* were used to indicate absence of a sign. Signs were individually documented in both their full and abbreviated forms. For example, a variant of the phrase *no diarrhoea,* indicating that the animal did not have diarrhoea, was found in 3.3% of narratives, including; *d-* (1.9%), *no d+* (0.4%), *no diarrhoea* (0.4%) and *no d* (0.3%). This was further complicated where on occasion the statement of absence followed the sign, rendering identification of to which sign the *no* or *none* referred challenging where signs were listed without punctuation.

Table 4.4.h: Regular expressions used to gauge the volume of contents regarding owner reported clinical signs.

		Regular expression matched within narrative	
Phrase or word	**Regular expression**	**n=**	**Proportion (%)**
no d excluding dr and die	no\W+d(?!ie|r)	5340	3.27(3.18, 3.36)
d-	d[\-]	3121	1.91(1.85, 1.98)
no d	no\W*d(?!\w|[+])	412	0.25(0.23, 0.28)
no d+	no\W*d[+]	713	0.44(0.4, 0.47)
no diarrhoea	no\W*diar+h?[ohea]{2,}	608	0.37(0.34, 0.4)
vomit diarrhoea cough sneeze	(?:(?:vom|dia|cou|sne)\w*\W*(?:(?:and|or)\W*)?){4}	488	0.3(0.27, 0.33)
vdcs in any order	(?:[vdcs]\W*){4}	6335	3.88(3.79, 3.97)
vdcs	v\W*d\W*c\W*s	1988	1.22(1.16, 1.27)

Grouping of signs occurred, most notably *vomiting* and *diarrhoea*, reflecting signs that are common features of many diseases, signs commonly presented to the veterinary surgeon and signs that commonly co-existed. The order of a sequence of documented signs appeared not to be random, for example an abbreviation for *vomiting and diarrhoea* (e.g. *v+/d+)* occurred in 4.7% of consultation narratives, the short hand for *vomit* occurred first in 94.5% of these notations.

Similarly, variants of the phrase *vomiting, diarrhoea, coughing and/or sneezing*, which occurred in 0.3% of narratives, could be found in short hand form, for

example *vdcs,* with varying punctuation, in 3.9% of narratives. Ignoring the symbols there were 24 theoretical permutations of *vdcs* (possible permutations = n!/(n-r)! = 4!/1 = 4*3*2*1 = 24) however the order of the abbreviations was *vomiting diarrhoea coughing sneezing* in 31.4% of these (1.2% of narratives). On reading it was apparent that in some narrative records there had been automated expansion of abbreviations, this may account for a degree of the preference for specific sequence of signs.

4.4.6.3.2 Social history

The animal's social history, most notably how long they had lived with their current owner, was also documented, especially where an animal was attending for the first time or there was a concern regarding their behaviour. *Biop*, present in 1.3% of narratives, was used to denote the animal had *been in owner's possession* for the specified period of time. *Biop* was overloaded, also being used as an abbreviation of biopsy, when a sample of tissue is taken for examination, however on examining the concordant phraseology 99 in 100 uses were referring to how long the animal had lived with its current owner.

Pertinent information regarding the owner that impacted the animal's health or care were occasionally present within the clinical history. This included information of the owner being, or imminently going to be, absent and the animal cared for by another party, owner illness or disability, and financial constraints impacting the available management options.

4.4.6.4 Examination

Where a clinical examination was documented this could be divided into the animal's demeanour, a general examination of the major bodily systems and a specific examination of any systems or body areas of current or previous concern. Thus, if an animal presented with a limp there would be likely to be comment on the animal's general appearance, a musculoskeletal examination, and perhaps also examination of their neurological, gastro-intestinal and cardio-respiratory system. Documentation of the clinician's examination ranged from brief non-specific comment to extensive documentation of both normal and abnormal findings. This range of verbosity is illustrated in Figure 4.4.a, the two narrative records are for animals that presented with recurrent unilateral otitis externa, narrative a has been abridged with an additional 60 words describing management strategies redacted.

a) "Re-ex l ear. Flare-ups every few months, owner has been using cleaner and surolan left over from march, but feels no improvement this time around. Does appear to fully resolve in between flare-ups, rest of skin great at the moment. R ear - clear, no concerns. L ear - doesn't tolerate touching, has lichenification, stenotic ear canal, erythema, crusty beige discharge, very uncomfortable... For now, treat with cleaning/drops and steroids, steroids should help reduce irritation and allow owner to clean better. R/V 7 days."

b) "oe left ear still v bad - discuss tymp memb. advise e/d and re-examine if not improved."

Figure 4.4.a: Contrast of the degree of verbosity used to describe consultations for animals with similar presentations.

4.4.6.4.1 Statement of general well-being

A range of statements were used to denote that the clinician had seen an animal that appeared generally well. There was overlap with the statement of no owner concerns but these statements tended to imply the clinician rather than owner's opinion.

The phrases; *aok* (0.8%), *all* ok (4.2%), *all fine* (2.6%) or *all good* (0.3%) occurred in 7.8% of narratives. When more specifically describing the healthy animal in terms of their demeanour, statements such as *bright* (7.4%) and *alert* (3.7%) were used, the phrase *bright alert responsive* occurred in 3% and its abbreviated form *bar* in 12.4% of narratives. The demeanour acronym *bar* was found in combination with the bodily function acronym as *bar dude* in 2.2% of narratives, denoting *bright alert responsive defaecating urinating drinking and eating are all fine* (Table 4.4.i).

Table 4.4.i: Regular expressions used to gauge the volume of statements of general well-being.

		Regular expression matched within narrative	
Phrase or word	**Regular expression**	**n=**	**Proportion (%)**
all ok/fine/good	`(?<!\w)al*\W*ok|all\W*(?:fine|good)`	12717	7.79(7.66, 7.92)
aok	`(?<!\w)aok`	1298	0.8(0.75, 0.84)
all ok	`(?<!\w)all\s?ok`	6844	4.19(4.1, 4.29)
all good	`(?<!\w)all\s?good`	410	0.25(0.23, 0.28)
all fine	`(?<!\w)all\s?fine`	4235	2.59(2.52, 2.67)
bright	`bright(?!\Wred|green|yel|pink)`	12082	7.4(7.27, 7.53)
alert	`(?<!less\s)alert`	6113	3.74(3.65, 3.84)
bright alert responsive	`bright\W*alert\W*(?:and|&)?\W*responsive`	4840	2.96(2.88, 3.05)
bar	`(?<!\w)bar(?!\w)`	20244	12.4(12.24, 12.56)
bar dude	`(?<!\w)bar\W*dude`	3508	2.15(2.08, 2.22)

Conversely where an animal appeared generally unwell the words lethargic (1.8%), *quiet* (1.7%), off colour (0.4%), *not [him/her/them] self or selves* (0.4%), *flat* (0.4%), *subdued* (0.3%), *depressed* (0.2%), *dull* (0.2%) and *withdrawn* (0.04%) were used (Table 4.4.j) These words were overloaded, with different meanings dependent on context.

Table 4.4.j: Regular expressions used to gauge the volume of statements of illness.

Phrase or word	Regular expression	Regular expression matched within narrative n=	Proportion (%)
lethargic	le[th]{2}ar	2874	1.76(1.7, 1.82)
quiet	quiet	2719	1.67(1.6, 1.73)
not self	not\s\w*\s?sel[fv]	696	0.43(0.39, 0.46)
off colour	off\W*col[ou]+r	659	0.4(0.37, 0.43)
flat	flat	692	0.42(0.39, 0.46)
subdued	subdu(?!r)	475	0.29(0.26, 0.32)
dull	(?<!\w)dull(?!\w)	330	0.2(0.18, 0.22)
depressed	depres+ed	266	0.16(0.14, 0.18)
withdrawn	withdrawn	66	0.04(0.03, 0.05)
qar	(?<!\w)qar(?!\w)	1052	0.64(0.61, 0.68)

Within the veterinary narrative *quiet* was used when referring to a quiet sound, quiescent inflammatory process, advising an owner to keep an animal quiet, and as a misspelling of quite, 4 in 5 uses referred to a quiet demeanour, this use was also abbreviated to *qar*, denoting *quiet alert responsive*, which was present in 0.6% of narratives. *Flat* was used most commonly to describe a lesion and also in reference to where the animal lived, level ground, geometric position or stance, approximately 1 in 3 uses of *flat* within this sub language referred to an ill or collapsed animal's condition.

4.4.6.4.2 General examination

A word of the lemma *exam* (23%) often preceded or followed the findings of the clinical examination (Table 4.4.k). *On exam* (5.2%)*, on examination* (0.7%), and *general examination* (0.01%) were less commonly used than the short hand *exam* (13.6%) alone. Where an indication was present, a range of abbreviations were used to denote that what followed was a description of clinical examination, *pe* (5.8%), *ce* (8.2%), *clin* (1.5%), and *px* (0.2%). *px* was

overloaded, also denoting planned management or prescription but these uses accounted for less than 1 in 10 occurrences of *px*.

Table 4.4.k: Regular expressions used to explore indicators of an examination being documented.

		Regular expression matched within narrative	
Phrase or word	**Regular expression**	**n=**	**Proportion (%)**
Lemma exam	`exam`	37540	23.0(22.79, 23.2)
on exam	`(?<!\w)on\sexam\W`	8536	5.23(5.12, 5.34)
on examination	`(?<!\w)on\sexamination`	1128	0.69(0.65, 0.73)
general examination	`(?<!\w)general\sexamination`	19	0.01(0.01, 0.02)
exam not on	`(?<!\w)(?<!on\s)exam(?!\w)`	22224	13.61(13.45, 13.78)
ce	`(?<!\w)c\W*e(?!\w)\W*(?!\w)`	13320	8.16(8.03, 8.29)
pe	`(?<!\w)p\W*e(?!\w)\W*`	9478	5.81(5.69, 5.92)
clin	`(?<!\w)clin(?!\w)`	2521	1.54(1.48, 1.6)
px	`(?<!\w)px`	386	0.24(0.21, 0.26)

As part of the general clinical examination it was common to document the *capillary refill time*. This is the time taken for an external capillary bed to refill following blanching by pressure, reflecting peripheral perfusion and a crude, but valuable, assessment for haemodynamic compromise. Capillary refill time is commonly measured in an animal by gentle manual pressure on the gum. The results of this assessment could be found in 7.4% of narratives, 87.2% of them using the abbreviation *crt* (Table 4.4.l).

Table 4.4.l: Regular expressions used to explore indicators of a gross examination of hydration and haemodynamic status.

		Regular expression matched within narrative	
Phrase or word	**Regular expression**	**n=**	**Proportion (%)**
cap refill	`(?<!\w)crt|cap\w*\W*ref`	12050	7.38(7.25, 7.51)
crt	`(?<!\w)crt`	10509	6.44(6.32, 6.56)
mm	`(?<![a-z])mm(?!\w)`	13527	8.29(8.15, 8.42)
mm not num	`(?<!\d)(?<!\d\s)(?<![a-z])mm(?!\w)`	10780	6.6(6.48, 6.72)
num mmhg	`\d\s?mm\s?hg`	300	0.18(0.16, 0.2)
mms	`(?<!\w)mms`	2329	1.43(1.37, 1.48)
mmpm	`(?<!\w)mm\W*pm`	159	0.1(0.08, 0.11)
tacky	`tac[hk]y(?!\w)`	611	0.37(0.34, 0.4)
tent	`(?<\w)tent`	1438	0.88(0.83, 0.93)

Apparent hydration status of an animal was a key feature of the documented examination, this was usually measured with reference to the condition of the animal's oral mucosa, alongside exclusion of pallor. Within the sub language of the veterinary clinical narrative the abbreviation *mm* usually referred to the mucous membrane and not millimetres. The singular *mm* could be found in 8.3% of consultation narratives, 79.7% of these (6.5% of all narratives) contained the abbreviation in a form not preceded by a number, implying it was not being used as a measurement. In 10.9% of the narratives where a number did precede the abbreviation *mm* (0.18% of all narratives), the abbreviation was being used as the unit of blood pressure measurement *mmhg or mm hg*, abbreviating *millimetres of mercury*.

Variations on the *mm* abbreviation included its plural, present in 1.4% of consultation narratives and *mmpm*, abbreviating *mucous membranes pink and moist*, present in 0.1% of narratives. Where the mucous membranes were found to be dry they were described as tacky (0.34%). An alternative measure of hydration status was assessment for 'skin tenting', this is described where an animal is dehydrated and skin raised between the vet's fingers fails to rapidly return to lying flat. the stem *tent* was present in 0.88% of narratives.

The acronym of *nothing adverse, or abnormal, detected*, *nad,* was present in 17.6% of consultation narratives. This was used both generally, to imply the animal was well, and specifically in respect of a body system, examination or investigation. Where the latter were found normal the abbreviation of *within normal limits, wnl* (5.5%), *normal* (19.7%), *n* (1.4%), and *norm* (0.1%) were also used. Other acronyms indicating normality were identified, but used less commonly, for example no significant finding was abbreviated as *nsf* in 0.2% of consultation narratives and *no abnormality found* as *naf* in 0.04%. Where a clinical sign or concern had been documented but the remainder of the examination was normal the phrase *otherwise well* (1.1%) or variations of similar (2.5%) were used (Table 4.4.m).

Table 4.4.m: Regular expressions used to explore indicators of normality

Phrase or word	Regular expression	Regular expression matched within narrative n=	Proportion (%)
nad	(?<!\w)nad(?!\w)	28642	17.55(17.36, 17.73)
wnl	(?<!\w)wnls?(?!\w)	8958	5.49(5.38, 5.6)
normal	(?<!\w)norma	32150	19.69(19.5, 19.89)
n	(?<!\w)n(?!\w)	2298	1.41(1.35, 1.46)
norm	(?<!\w)norm(?!\w)	157	0.1(0.08, 0.11)
naf	(?<!\w)naf(?!\w)	66	0.04(0.03, 0.05)
nsf	(?<!\w)nsf(?!\w)	240	0.15(0.13, 0.17)
otherwise well	otherwise well	1709	1.05(1.0, 1.1)
otherwise well variation	(?:otherwise|else|rest)\W* (?:well|fine|good|nad|wnl|naf)	4113	2.52(2.44, 2.6)

4.4.6.5 Differential diagnosis

Where an animal was presented with clinical signs or owner concerns some clinicians documented a summary of their current list of likely underlying issues, their differential diagnosis. This was usually preceded by the phrase *differential diagnosis* (1%), *ddx* (0.9%) or *differential* (0.2%). Less formal, discursive, documentation of the clinician's impression was often preceded by *could be* (1.6%) (Table 4.4.n).

Table 4.4.n:Regular expressions used to explore indicators of a differential diagnosis

Phrase or word	Regular expression	Regular expression matched within narrative n=	Proportion (%)
differential diagnosis	dif+er+e?nti\w*\s?diag	1680	1.03(0.98, 1.08)
differential	dif+er+e?nti\w*(?=\W)(?!\s(?:diag|count|white))	384	0.24(0.21, 0.26)
ddx	d\W*d\W*x	1496	0.92(0.87, 0.96)
could be	could\sbe	2655	1.63(1.57, 1.69)

4.4.6.6 Management

It was uncommon for a specific aetiology or diagnosis to be documented. The majority of consultations described pragmatic empirical management of the animal's signs and owner's concerns. Management strategies were often preceded by the word *plan* (6.6%) or try (6.6%). Abbreviations of plan were inconsistent and overloaded, *p* (2.7%) for example was used to mean *plan*,

often as *p:* (0.3%) or p- (0.4%) however, it was also used as part of abbreviations for *pink and moist* when referring to hydration status (0.9%), as an abbreviation for *polish* in *s&p* (0.2%), *scrape and polish*, and within the acronym for *polyuria and polydipsia, pupd*, which was usually written *pupd* (0.7%) or pu/pd(0.6%) but occasionally with interposing symbols as p/u-p/d (0.1%) (Table 4.4.o).

Table 4.4.o: Regular expressions used to explore indicators that what followed was the management plan.

Phrase or word	Regular expression	Regular expression matched within narrative n=	Proportion (%)
plan	`(?<!\w)plan(?!\w)`	10791	6.61(6.49, 6.73)
try	`(?<!\w)try`	10702	6.56(6.44, 6.68)
p	`(?<!\w)p(?!\w)`	4350	2.66(2.59, 2.74)
p:	`(?<!\w)p\s?:(?!\w)`	517	0.32(0.29, 0.34)
p-	`(?<!\w)p\s?\-(?!\w)`	633	0.39(0.36, 0.42)
p\W*m	`(?<!\w)p\Wm(?!\w)`	1512	0.93(0.88, 0.97)
s\W*p	`(?<!\w)s\W*p(?!\w)`	371	0.23(0.2, 0.25)
pupd	`pupd`	1212	0.74(0.7, 0.78)
pu/pd	`pu\W+pd`	1040	0.64(0.6, 0.68)
p/u-p/d	`p/u\-p/d`	150	0.09(0.08, 0.11)
(?<!\w)rx	`(?<!\w)rx`	1866	1.14(1.09, 1.19)

Overloading was also evident with the abbreviation *rx* (1.1%). From the latin verb *recipere*, *to take*, *rx* is commonly used to denote a prescription in human medicine. However, in the veterinary sublanguage this use accounts for less than 1 in 10 occurrences of *rx*. The usual meaning is as an abbreviation of *review*, denoting either that a consultation was a review for the specified reason or the circumstances in which the animal should be reviewed.

4.4.6.7 Preventative veterinary medicine

The clinical narrative commonly included evidence of the clinician having advised the owner regarding preventative health care and husbandry for welfare benefit. This included advice regarding parasite prophylaxis, neutering and breeding, diet and weight management, dental care and general husbandry. Parasite control was a common feature, differentiation between notation that a parasite has been observed and where it has been checked for, discussed or prophylaxis provided would be essential if surveillance for parasites was

undertaken via text-mining. The stem *flea* was present in 7.89% of narratives, in a sample of 100 occurrences of the stem 91 (95% CI: 85.39, 96.61)% were notations of preventative health care, the remainder reflected where fleas or flea dirt had been observed. Similarly the stem worm was present in 7.2% of narratives, 99 (95% CI: 97.05, 100)% of these occurrences reflected prophylaxis or its discussion. Preventive measures were often noted together in long hand and as abbreviations, for example *weigh and worm*, *flea and worm* (Table 4.4.p).

Table 4.4.p: Regular expressions used to explore notation of preventive health care

Phrase or word	Regular expression	Regular expression matched within narrative n=	Proportion % (95% CI)
flea	`(?<![a-z])flea(?![a-z])`	10,379	6.36(6.24,6.48)
	`(?<!\w)flea\w*`	12,884	7.89(7.76, 8.02)
f&w variants	`(?<!\w)f\W*w`	500	0.31(0.28, 0.34)
worm stem	`(?<!\w)worm\w*`	11,758	7.2(7.07, 7.33)
booster	`(?<!\w)bo+st\w*`	15,524	9.51(9.37,9.65)
vac stem	`(?<!\w)vac\w*`	30,614	18.75(18.56, 18.94)
vacc or booster	`(?<!\w)vac\w*|(?<!\w)bo+st\w*`	40,672	24.92(24.71, 25.13)
kc	`(?<!\w)kc(?!\w)`	11,982	7.34 (10.82% of dogs)
felv	`(?<!\w)felv(?!\w)`	6578	4.03 (13.92% of cats)
hpc	`(?<!\w)hpc(?!\w)`	5,402	3.31(3.22,3.4)
ghc	`(?<!\w)ghc(?!\w)`	1212	0.74(0.7, 0.078)

Vaccinations were referred to in a range of ways, the word booster was present in 9.51% of narratives, and the stem *vac* in 18.75% with a combined occurrence in 24.92% of narratives. The illness against which a vaccination was targeted was often used in isolation to denote or discuss the vaccination or the animal's vaccination status, *kc* abbreviating kennel cough was present in 10.82% of dog consultation narratives and *felv* abbreviating *feline leukaemia virus* in 13.92% of cat consultation narratives. This was complicated by the similar use of these abbreviations when the illnesses themselves were being discussed.

4.4.6.8 Safety net

There was extensive evidence of clinicians having ensured that the animal's owner knew in which circumstances they should take further action, usually returning for clinical review, this is known as providing a safety net or safety netting (Table 4.4.q) The most common forms of this were phrases implying *if x then y* (2.3%) and the phrase *if no improvement* (1.6%) which was also used in abbreviated form *ini* (2%) and *inb* (0.3%). The conjunction *if* (27%) often denoted a safety netting phrase, for example; *if not* (4.1%), *recheck if* (0.6%), *reex if* (0.5%), *if deteriorates* (0.4%), *review if* (0.3%) and *come back if* (0.2%).

Table 4.4.q: Regular expressions used to explore indicators of a differential diagnosis

		Regular expression matched within narrative	
Phrase or word	**Regular expression**	**n=**	**Proportion (%)**
if x then y	(?<!\w)if\s(?:\w+[^\.]){1,4}then	3827	2.34(2.27, 2.42)
if no improvement	if\sno\simp	2674	1.64(1.58, 1.7)
ini	(?<!\w)ini(?!\w)	3267	2.0(1.93, 2.07)
inb	(?<!\w)inb(?!\w)	481	0.29(0.27, 0.32)
if	(?<!\w)if(?!\w)	44044	26.98(26.77, 27.2)
if not	(?<!\w)if not	6640	4.07(3.97, 4.16)
recheck if	(?<!\w)r\w+ck\sif	948	0.58(0.54, 0.62)
re-ex if	(?<!\w)re+\W*e?x\sif	808	0.49(0.46, 0.53)
if det	(?<!\w)if det	696	0.43(0.39, 0.46)
review if	(?<!\w)rev\w*\sif	456	0.28(0.25, 0.3)
come back if	come\sba?ck\sif	310	0.19(0.17, 0.21)

4.4.7 Estimation of vocabulary size

The exploratory corpus contained a total of 8,678,599 words, this included 116,808 unique words. Utilising only words that were a string of letters, the corpus contained 847,0311 words with 93,211 unique words, the apparent vocabulary size. Over half (56.39%) of the words occurred only once within the corpus. Samples were taken at random at frequencies of 1, 2 and 10 occurrences within the exploratory corpus. Above this there were insufficient to take a sample and all words within each frequency range were examined (Figure 4.4.b)

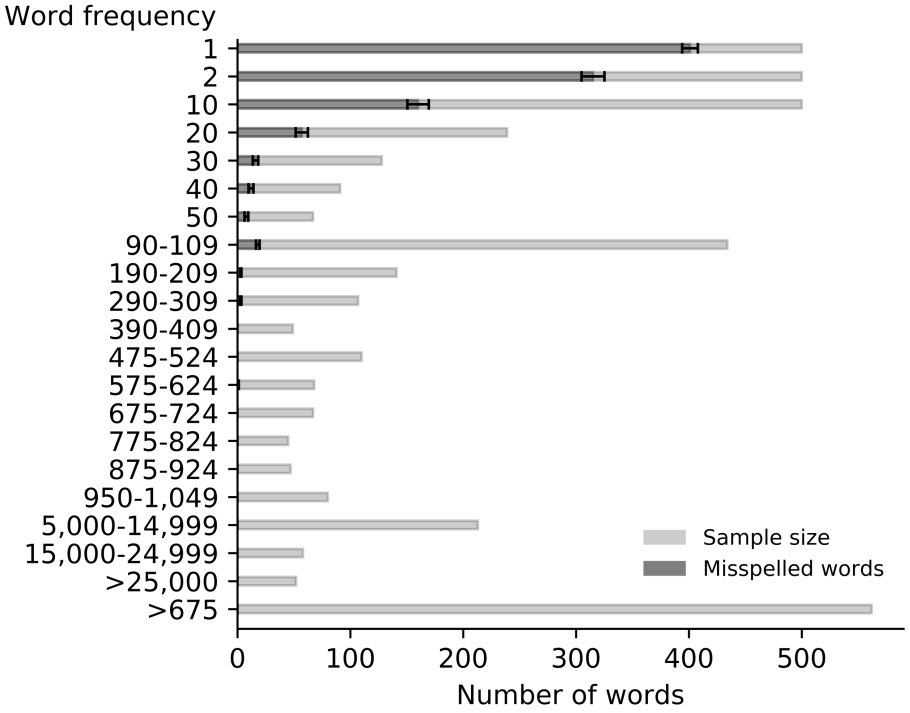

Figure 4.4.b: Size of sample and number of misspelled words at intervals of word frequency in the exploratory corpus. Confidence interval represents 95% confidence limit using Wilson's method.

The majority of words with an occurrence in the exploratory corpus of 2 or less were misspelled, 80.2 (95% CI: 76.48, 83.46)% and 63 (95% CI: 58.68, 67.12)% for those occurring once and twice respectively. There was a single spelling error within the 68 words with a frequency of 575-624, and no other errors above a word frequency of 309 (Table 4.4.s). The true vocabulary size, estimated from the periodically sampled misspelling rate, was 37,354 (95% CI 33,791, 40,852) words, 40% of the apparent vocabulary size. Weighting the word frequencies within the exploratory corpus by the spelling error rates identified suggested that 2.42% (95% CI: 2.02, 3.6)% of words within the corpus were misspelled.

Table 4.4.r: Estimation of spelling and typographic error rate by word frequency within the exploratory corpus. * Samples selected at random from words meeting frequency criteria. $ total sample size above this point.

Word frequency	Sample size	Spelling errors	Misspelled Proportion	95% CI
1	500*	401	80.2	76.48, 83.46
2	500*	315	63	58.68, 67.12
10	500*	160	32	28.06, 36.21
20	239$	57	23.85	18.89, 29.64
30	128	16	12.5	7.84, 19.34
40	91	12	13.19	7.71, 21.65
50	67	8	11.94	6.18, 21.83
90-109	434	18	4.15	2.64, 6.46
190-209	141	3	2.13	0.73, 6.07
290-309	107	3	2.8	0.96, 7.92
390-409	49	0	0	0, 7.27
475-524	110	0	0	0, 3.37
575-624	68	1	1.47	0.26, 7.87
>675	562$	0	0	0, 0.68

4.5 Discussion

This work provides an overview of the sentence metrics of the veterinary clinical narrative, an understanding of which is paramount for the design of effective text-mining tools targeted to extracting information from the veterinary clinical narrative for surveillance and cohort study purposes.

The broad range of narrative length, from the extremely brief with 17% of narratives contained only a single sentence, or less than ten words, to the verbose 1% containing over 500 words combined with the similarly variable sentence length, with up to 269 words identified within a sentence must be borne in mind during development of any text-mining algorithm. Such brevity and verbosity would need to be accounted for within pattern recognition mechanisms with steps to avoid excluding the extremes without impairing precision in the more typical narratives, or compromise that only the more typical records were likely to be detected but that this may provide higher quality data. The wide spread of mean word count and sentence length across the 371 clinics of the exploratory corpus highlighted the potential risk of introducing a geographically bias to text-mining tools by failing to take account of the varying degrees of verbosity used in documenting the clinical narrative.

The greater number of words used in documenting consultations occurring out-of-hours may reflect the different case mix seen at these times with a likelihood that animals were more seriously ill or injured and potentially unknown to the treating clinician. Such circumstances would likely be associated with a more extensive gathering and exchange of information and the need for more detailed clinical record, especially where it may also be used to communicate with the animal's regular in-hours clinician. In addition, clinics providing an out-of-hours service to clients registered elsewhere for daytime care may mandate a minimum required level of information to be documented, to better facilitate handover of care to the animal's usual clinician.

The reasons for the lower mean word count observed in consultations regarding rabbits can only be speculated. The consultation observation work of Robinson et al. (2015) may provide insight, with their observation that fewer problems were discussed for rabbits than cats ($P<0.001$) or dogs ($P<0.001$). Alternative explanations include differences in the nature of presentations seen in rabbits compared to the other species, for example rabbits may be brought in less often for routine preventive treatment or more commonly whilst in extremis. Availability of treatments, clinician and client expectations and experience may all have a role. That rabbit consultation length also differs from that of the uncommon species group is particularly perplexing.

The disparity in word length and complexity between species would, for human reading, affect the ease with which they were read, shorter and less complex words being more easily human read (Flesch 1948; McLaughlin 1969; Gunning 1952). For software reading of the text, and development of text-mining algorithms for application within this sublanguage, the challenges are not the readability but the lack of consistency across the species. This has implications for both development and validation of text-mining tools, with the need to either have specific tools adapted to the records of a given species, or to validate tools separately across the species acknowledging that performance may differ between the species.

Sentence length would similarly be expected to affect human readability (Spache 1953; Flesch 1948; Gunning 1952; Dale and Chall 1948), its implications for text-mining however is on the appropriate parsing of strings of words, in order to correctly relate their contained information. The longest

sentence in the exploratory corpus contained 269 words, 1 in 10 narratives contained a sentence over 30 words long. This posed the challenge of ensuring that extended clauses were captured, for example descriptions of diarrhoea where reference to bowel and consistency are tens of words apart, without inadvertently using pieces of information from multiple clauses within the same sentence as though they belonged to a single clause, as for example where a long sentence describes the absence of conjunctivitis but the presence of sneezing.

Conversely 63.5% of narratives contained at least one sentence composed of three or fewer words. Sentences of this nature posed different but equally onerous challenges to the development of effective automated information extraction tools. Such sparse use of language impairs sentence interpretation and highlights the telegraphic nature of much of the veterinary clinical sublanguage.

The greater numeric content within dog and cat consultations compared to those of the rabbit and uncommon group may reflect a greater availability of investigations, and a greater evidence base in managing these more commonly seen species. Alternative explanations could include that these animals are more readily handled and as a result it may be easier for the clinician to measure parameters such as the heart rate and temperature and to draw blood for laboratory investigations. The frequency with which physiological parameters are documented in the narrative record is examined further in Chapter seven.

The greater lexical diversity (lower Maas a statistic) in consultations regarding cats, compared to other species, may reflect differences associated with the animals, owners or clinicians, or interactions between each. A potential explanation may be that the nature or number of problems presented during a cat consultation differs from those for other species, perhaps in itself reflecting differences in the challenges of handling a cat compared to a dog for example. Work by Robinson et al. (2015) found no difference in the number of problems presented during cat and dog consultations, which would suggest it may be more likely to be the nature rather than number of problems. It may be that the nature of interaction that occurs between clinicians and cat owners results in a more diverse form of documentation, or that some clinicians are preferentially consulted regarding cats and those clinicians have a more rich writing style. If

consent and ethical approval permitted, the identities of clinicians could be used to examine whether the differences between species persisted within the consultations of individual clinicians.

4.5.1 Comparison to standard English

As would be expected, there is little doubt that the language used in documenting the veterinary clinical narrative in the UK differs from standard English. The extensive use of abbreviations and words with alternative meaning reflects the constrained domain and purpose of the clinical narrative, in keeping with the descriptions of Kittredge on the circumstances in which a sublanguage will develop (Kittredge 1983). These notations were also highlighted by recent work utilising a corpus drawn from a similar UK first opinion veterinary dataset (K. Cheng, Baldwin, and Verspoor 2017).

The relatively small degree of commonality between commonly used words of the veterinary clinical sublanguage and the British National Corpus is in keeping with work exploring the co-existence of words in a corpus of human medical records and purpose made medical vocabularies, including SNOMED (National Library of Medicine 2017b) and the UMLS Metathesaurus (National Library of Medicine 2009) and other gazetteers which found that 40% of words within the medical records could not be identified within these vocabularies (Hersh and Campbell 1997). This piece of work also identified similar issues of neologisms, contextually appropriate shortening and compression of words or phrases. This has considerable implications for the use of such gazetteers as a resource in designing text-mining algorithms, especially for the veterinary clinical narrative where the techniques are more nascent than in human medicine and gazetteers yet to be developed, it may be that their development would not be judicious and energies could be better targeted to real world language use.

4.5.2 Atypical grammar & spelling

The lack of capitalisation in 4 out of 10 sentences in conjunction with the previously noted very long and very short sentences, highlights the need for any text-mining system not to rely on the usual grammatical rules of the English language when attempting to extract information from the English veterinary clinical sublanguage. It must be borne in mind that the pre-processing of

narratives created pseudo-sentence structure where the clinician or practice management software had included line breaks.

The extensive use of non-English words, either by virtue of misspelling, abbreviation, colloquialism or technical language, makes the inclusion of an ability to match similar but non-identical phraseology essential to any system designed to extract information from the veterinary consultation narrative. An alternative, or adjunctive, approach would be to institute a spelling correction algorithm within the pre-processing pipeline, this would be likely to be problematic given the overloaded vocabulary. Preliminary exploration of the efficacy of stemming, lemmatisation and part of speech tagging within this corpus showed it to be sufficiently poorly efficacious as to abort further work in that direction. This was to be anticipated given the tools being used were not trained on the veterinary clinical sublanguage.

Spelling correction within the veterinary clinical sublanguage would require the formation of a domain specific dictionary of correctly spelled words. This would permit identification of non-words, by their absence from the dictionary, and identification of the dictionary word formed by the minimum number of substitution, insertion or deletion operations via the minimum Damerau–Levenshtein edit distance (Damerau 1964). This process may be augmented by information regarding word frequency within a corpus and phonetic similarity (Lai et al. 2015). This process would not address dictionary words that formed strings without syntactical or semantic sense, language models can be used to identify the likely intended word in these situations .

The approach taken here, with ability to detect relevant notation and context even where there were substantial spelling errors, was considered more appropriate given the syntactic and semantic peculiarities and diversity of the sublanguage. It is likely that a dictionary and language model for consultations regarding individual species may be required to achieve an acceptable level of accuracy in spelling correction. Even were these resources in place the efficacy of spelling correction achieved was felt unlikely to improve on the efficacy of the more forgiving approach of working around the error rate within the corpus.

4.5.3 The need for context sensitivity

The non-standard and polysemic use of abbreviations and symbols within the veterinary clinical sublanguage poses a challenge to their automated interpretation. As with much of the information within the narrative record, and indeed within most natural language, comprehension of the context in which words have been used is critical to extracting the correct information. Software recognition of the presence of any word, abbreviation or symbol is of little value if the context in which it has been used cannot also be interpreted.

This is well illustrated by the findings of the brief exploration of the documentation of preventive health care. Searching for the word flea or worm, or their plural would, primarily identify preventive measures and not actual infestation or its consequences. This particular example also highlights the need for validation of the efficacy of any classification system; as both prophylaxis and infestation may exhibit a seasonal pattern, with extra efforts being made to prevent infestation when the risk of such is greatest, there is potential for false reassurance that a non-context sensitive classifier was detecting a seasonal pattern, and thus was behaving as anticipated.

Identification of the themes present and their manner of delineation within the narrative of a veterinary consultation is valuable in identifying boundaries of pieces of information, providing landmarks adjacent to which particular pieces of information may be found. This is of value in designing text-mining algorithms. However, given the wide-ranging volume of information documented in the narrative, and the sparse documentation used to record a significant proportion of consultation, the landmarks cannot be relied upon as the only marker of an event.

Recognition of common notation indicative of the reason an animal's owner sought veterinary attention on a given occasion has the potential to assist in identifying and extracting that information using text-mining techniques. This is not without limitations however in that a single, exclusive, reason for visit poorly summarizes the wealth of information communicated with a veterinary consultation. It would however provide scope for validation of other systems of classification, including the clinician-assigned main reason for visit categorisation applied to the SAVSNET dataset, and also as a secondary

source of information for use as an adjunct to more finessed concurrent information extraction.

The, wholly appropriate, documentation of aspects of the consultation such as owner concerns, differential diagnosis and safety netting advice introduces an array of information that does not directly describe the clinical situation of the animal presented. Without concurrently understanding the context in which it was documented, any text-mining classification system would carry the potential for a high rate of false positive classifications.

It is clear on reading clinical narratives in this corpus that in some cases automated expansion of abbreviations has been incorporated into the clinic software. It is likely that some of the findings described here reflect this. However, whilst consequently the text does not always directly reflect what was typed by the clinician, it is the material from which any text-mining will extract information, thus the amended language, and occasional imprecision, resulting from automated abbreviation expansion must be incorporated into any such information extraction process.

4.5.4 Conclusion

This work describes the previously undocumented veterinary clinical sublanguage and highlights the paramount importance of domain knowledge and domain specific context sensitivity in the development of any text-mining algorithm targeted to the veterinary clinical narrative.

Chapter Five Redaction of incidental identifiers within free-text veterinary clinical records

5.1 Introduction

This chapter describes the legal and ethical need for data to be handled in a manner that protects identities and techniques that have been used to address this key challenge in the re-purposing of clinical free-text for research applications. Preliminary work to quantify the volume of identifiers present within the free-text records of the SAVSNET dataset are described alongside foundation experiments that aimed to optimise programmatic processes within the de-identification software. The development of Clancularius, de-identification software specifically tailored to the veterinary clinical sublanguage, is then described and its efficacy assessed within the SAVSNET and Bristol Cats Study datasets.

5.1.1 The need for de-identification

Electronic health records (EHR) are widely and increasingly utilised in health care settings (WHO Global Observatory for eHealth 2006); in the United Kingdom the majority of community-based health care is documented within an EHR (Payne et al. 2011; D. Robinson and Hooker 2006). In addition to its clinical functionality, the data contained within the EHR is a valuable source of information regarding health and wellbeing at a population level (Lund 2015).

Even where client consent for the sharing of information has been gained, a duty to abide by data protection legislation persists. Much of this legislation is based on the guidance provided by the Organisation for Economic Co-operation and Development in 1980 (OECD 2010) including the UK Data Protection Act (UK Parliament 1998), European Union General Data Protection Regulations (European Union 2016) and US Health Insurance Portability and Accountability Act, HIPPAA (U.S. Department of Health and Human Services 1996). In the UK the Data Protection Bill 2017 was introduced in September 2017 and when enacted will supercede the Data Protection Act 1998 and incorporate the requirements of EU General Data Protection Regulations (UK Parliament 2017a).

Among other requirements, European legislation dictates that where an organisation holds information regarding identifiable living people it must: only collect information for a specified purpose; and only store that which is needed

for as long as it is needed (UK Parliament 1998; European Union 2016). Although the clinical information within veterinary clinical records for the most part relates to an animal or herd, information regarding the owner and other third parties may also be documented, for example in explanation of financial constraints, owner ill health and travel plans. This is recognized by the Royal College of Veterinary Surgeons (RCVS) which advises against the inclusion of such information within the clinical record itself, except for a statement of the need for limitation of treatment as a result of owner circumstances (Royal College of Veterinary Surgeons 2016).

The need for conscientious data security in healthcare research was highlighted by Nelson (2015) who showed that 42% of data breaches in the United States during 2014 were breaches of healthcare related data. The majority of breaches resulted from oversight by the data user rather than unpermitted access by an external party.

Where research relies on the coded and constrained fields of the EHR, such as gender and clinical coding, protection of individual identities can be achieved by omission of identifier fields (e.g. postcode, name and address), and the use of only aggregate data, providing summary statistics.

When research involves examination of the free-text narrative content in clinical records; there is a risk that the record will contain sufficient information, either in isolation or on linking to other available data, to identify the attending clinician, owner, and potentially other third parties. This poses a challenge to the use of the clinical narrative for research purposes. International standards of de-identification are currently being developed by the International Organization for Standardization (ISO/IEC JTC 1/SC 27 IT Security techniques 2016).

Redaction of sufficient personal identifiers to render all parties unidentifiable is desirable. The clinical narrative is potentially the richest source of detailed information regarding disease and, as such it is critical to develop methodologies to access this resource in an appropriately de-identified manner. This would permit respectful and ethical use of narrative data for research purposes with minimal likelihood of any party being inadvertently identified by the researcher.

5.1.2 Techniques previously described

Techniques used to redact identifiers within clinical free-text can largely be divided into two groups; rule-based and machine-learning, each with their own relative merits. De-identification systems are often designed to address the de-identification needs of a specific document type within a domain, many of those described are validated in pathology reports (Berman 2003; Beckwith et al. 2006; Gupta, Saul, and Gilbertson 2004; Gardner and Xiong 2008; Thomas et al. 2002) or discharge summaries (Szarvas, Farkas, and Busa-Fekete 2007b; Neamatullah et al. 2008; Uzuner et al. 2008; Wellner et al. 2007). These documents are written explicitly in communication between clinical teams and would be expected to contain well-structured grammatical information, in contrast to the language and grammar used in contemporaneously documented clinical narrative, which is often produced under time pressure.

5.1.2.1 Rule-based de-identifiers

Rule-based systems require domain knowledge in the construction of their dictionaries, regular-expressions and other pattern-matching systems. They may be poorly generalizable to other datasets whilst conversely being readily adaptable with the addition of new patterns or rules (Deleger et al. 2013).

Simple alphanumeric patterns, such as postcodes, policy numbers and dates are commonly identified using regular expressions (Meystre et al. 2010). Names of people and places are identified via dictionaries constructed from census and similar collated information (Beckwith et al. 2006; Neamatullah et al. 2008; Friedlin and McDonald 2008). Where software is tailored to a specific institution, known staff and patient identifiers may be used (Neamatullah et al. 2008; Fielstein, Brown, and Speroff 2004; Friedlin and McDonald 2008). If the clinical records being de-identified contain identifier fields, for example within an XML structure, database table or document header, these can be exploited by the de-identification tool, with further occurrences of the known identifiers being redacted within the free-text fields (Friedlin and McDonald 2008; Beckwith et al. 2006; Gupta, Saul, and Gilbertson 2004).

Dictionaries are augmented by pattern matching, for example recognition of preceding titles or neighbouring context likely to indicate a name (Friedlin and

McDonald 2008; Beckwith et al. 2006; Thomas et al. 2002; Neamatullah et al. 2008). Common words and medical terms are used to disambiguate identifiers with other uses (Friedlin and McDonald 2008; Neamatullah et al. 2008). This becomes more problematic within veterinary clinical narratives where the patient is the animal and male and female owners are commonly differentiated by title alone, as for example 'unwell, mr reports vomiting, mrs has seen occasional blood'.

An alternative rule-based approach is to redact all words not found within clinical lexicon data sources, such as the Unified Medical Language System (UMLS) metathesaurus (National Library of Medicine 2009) and the Medical Language Extraction and Encoding System (MedLEE) lexicon (Friedman et al. 1994). Such techniques lend themselves to the structured information of pathology reports (Berman 2003) and communication between secondary and primary care (Morrison et al. 2009), but perhaps less so to the telegraphic and often unstructured first opinion clinical record where coherence would likely be severely impaired, an issue reported even in the former data types(Berman 2003). These systems rely on the prior existence of a wealth of lexical and semantic natural language processing infrastructure, resources not currently available for the veterinary clinical narrative.

5.1.2.2 Machine-learning de-identifiers

Machine-learning de-identification systems have the advantage of not needing their developer to recognize, nor create dictionaries of, the information to be redacted. Thus, domain knowledge is less critical. These systems however require large annotated training sets and this time-consuming process in itself requires extensive domain understanding.

Machine-learning de-identification algorithms rely on features of the text to identify the data to be redacted. Most system utilise lexical features, such as capitalisation (Gardner and Xiong 2008; Szarvas, Farkas, and Busa-Fekete 2007a; Taira, Bui, and Kangarloo 2002a; Uzuner et al. 2008; Wellner et al. 2007), punctuation (Taira, Bui, and Kangarloo 2002a; Uzuner et al. 2008) and numeric content (Gardner and Xiong 2008; Szarvas, Farkas, and Busa-Fekete

2007a; Uzuner et al. 2008; Wellner et al. 2007; Taira, Bui, and Kangarloo 2002b).

Part-of-speech identification is utilised as a means to identify syntactic features (Gardner and Xiong 2008; Taira, Bui, and Kangarloo 2002a; Uzuner et al. 2008), this poses a challenge in the non-grammatical heavily-abbreviated veterinary consultation narrative with its own sublanguage. As with rule-based systems semantic features are included in machine-learning algorithms, augmenting the functionality with dictionaries of identifiers (Szarvas, Farkas, and Busa-Fekete 2007b; Taira, Bui, and Kangarloo 2002a; Uzuner et al. 2008; Wellner et al. 2007), non-medical (Wellner et al. 2007) and clinical terms (Szarvas, Farkas, and Busa-Fekete 2007b; Taira, Bui, and Kangarloo 2002a). These are usually augmented by pattern-matching for items such as postcodes and telephone numbers. Features of the documents themselves are also incorporated into some algorithms, including section headers containing known identifier fields (Szarvas, Farkas, and Busa-Fekete 2007a; Uzuner et al. 2008).

5.1.3 Preservation of data utility

With redaction of identifiers comes the risk that an over-zealous tool would result in overscrubbing, degrading the quality and utility of the clinical narrative data. Although when evaluated in pre-existent systems the impact on the clinical information within the text field was thought to be small (Meystre, Ferrandez, et al. 2014); this is a potential source of reluctance to use automated de-identification when preparing healthcare records for research use. An estimated 95% of the highly structured and standardised, Health Level Seven (HL7), messages scrubbed using the Medical De-identification System (MeDS) retained readability and were interpretable (Friedlin and McDonald 2008). However, where de-identification has been thorough even the treating clinician may not recognise the identity of the patient from the de-identified narrative history (Meystre, Shen, et al. 2014).

Overscrubbing, resulting in impaired specificity and low precision (positive predictive value) of the de-identification algorithm, is not quantified for some systems, but can be assumed to be significant because of the method used, prioritising redaction of all that could be an identifier over identification of what is

likely to be an identifier (Morrison et al. 2009; Berman 2003). Beckwith et al. (Beckwith et al. 2006) evaluated their original rule-based version of HMS Scrubber and reported a precision of 42.4% when the system was applied to pathology reports whilst MeDs reported an average of 1.7 false positive identifier redactions (over scrubs) per message accounting for 8% of the apparent identifiers (Friedlin and McDonald 2008).

A number of de-identification systems include research-specific data preservation techniques to minimise the risk of producing de-identified documents of poor research data quality. Such systems have commonly focused on the obfuscation of quasi-identifiers, characteristics able to be used to identify an individual from their uncommon combination (Gal et al. 2014; Sweeney 1996; Machanavajjhala et al. 2006; N. Li, Li, and Venkatasubramanian 2007). Tu et al. (2010) described the development of a system incorporating means of preserving the peculiarities of primary care medical records, including the use of eponymous syndromes and abbreviated forms. However, a system specifically designed for, or validated in, the redaction of identifiers from veterinary narratives, with the preservation of clinically important features, has not been described.

5.1.4 Software specification

The intention of this work was to develop a system that would be able to minimise the likelihood of identifying individuals or veterinary practices from information available within the veterinary narrative, even when this was linked to prior knowledge or information available from other sources (Figure 5.1.a). The software, named Clancularius, needed to reliably identify name and address components included within unstructured and ungrammatical veterinary narrative text, without degrading the quality of clinically-relevant information available for research and surveillance purposes.

Primary aims

1. Redaction of owner, clinician and third party identifiers:
    a. Person and place names
    b. Postcodes
    c. Telephone numbers
    d. Email addresses
    e. Microchip numbers
2. Facility for adaptive research specific information preservation
3. Minimal degradation of clinically-relevant data quality
4. Ability to redact names absent from dictionaries in specified contexts

Secondary aims

1. Redaction of animal names
2. Redaction of clinician initials

Figure 5.1.a: Declared aims of the de-identification process.

The system was designed for use within UK small animal veterinary narratives and optimised for use within the narrative field of the Small Animal Veterinary Surveillance Network (SAVSNET) dataset. Software was to run on MacOS, Linux and Windows based operating systems and be able to process text files, dataframes stored as text files (csv format) and strings from a database, with output of the same formats. The aim was for a system that did not require significant computer proficiency to implement once developed.

5.1.5 Hypothesis

The veterinary clinical narrative provides a unique de-identification challenge as a result of its unmapped sublanguage with non-standard phraseology, telegraphic and abbreviated style. A rule-based classifier using logically-chosen dictionaries, sequential processing and data-masking can significantly reduce identifiers while maintaining research usability of records.

5.2 Preliminary observations with results directing methodology development

A series of experiments were performed to evaluate key features required to inform development of the automated de-identifier, these included:

- Evaluating and optimizing means of collecting ‘gold-standard’ de-identified records through semi-automated techniques
- Defining likely identifiers that would require redaction
- Verification of the principle of name pairs with meaning being unlikely to be paired within real names
- Comparing efficacy and processing speed of word matching methods

All code was implemented in Python version 3.6.1 (Python Software Foundation 2016) on a Macintosh iMac computer running MacOS 10.12.6 ('Sierra'). Python library versions are provided in Table 5.3.a.

5.2.1 Methods

5.2.1.1 Optimising manual de-identification

Comparison was made of the time taken to complete pattern recognition and decision-making steps of de-identification. The median word count of the non-parsed narrative field within the SAVSNET dataset, where a word was considered an alphanumeric sequence, was established using regular expression pattern counting within a Pandas dataframe. Three batches of 100 narratives of this length were selected at random from the SAVSNET dataset.

A Python script was written to present narratives in random order, requiring the user to record any identifier that was unique to that narrative. The software recorded the time from the narrative being presented to completion of de-identification. Where no identifiers were present this involved two keyboard key presses. Where an identifier was present this required a yes/no decision and entry of identifiers present in each of ten categories (Figure 5.2-a).

Preliminary observations with results directing methodology development

Each of the three batches of 100 narratives was manually de-identified by the author, assisted by the Python script, with a break of an hour between batches. These narratives were also fed to the full de-identifying package under construction as added optimization through identification of any discrepancies and errors. Discrepancies were examined and errors identified.

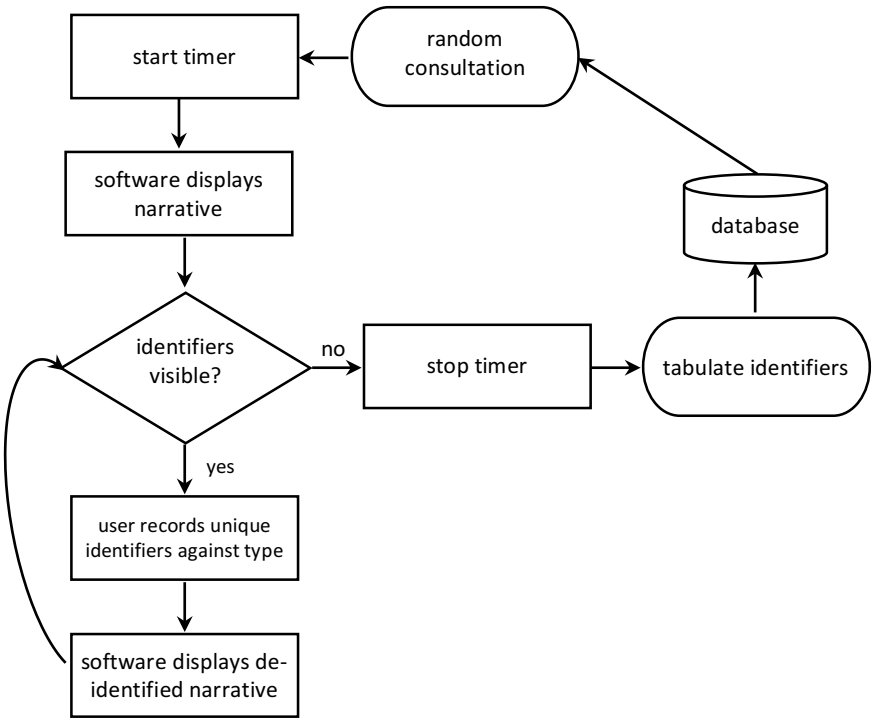

Figure 5.2-a: Python assisted semi-manual de-identification process

5.2.1.2 Defining the extent and nature of identifiers present within veterinary clinical narratives

A random sample of 1000 consultation narratives was taken from the SAVSNET dataset. This was calculated as an adequate sample size to determine the extent of identifiers within narrative records, based on an estimated 25% of records containing identifiers. Each narrative was annotated for the presence and type of identifiers using the semi-manual process outlined above. This process was intended to maximise the sensitivity of identifier recognition and document the nature of identifiers present within this sample, which was representative of the SAVSNET dataset as a whole.

5.2.1.3 Inter-operator validation of manual de-identification

Two samples of 100 consultation narratives were drawn from the sample of 1000 narratives used in section 5.2.1.2 above. These samples were independently semi-manually coded by project supervisors, using the same assistive software. Where there was disagreement in the manual identification of identifiers the narrative was reread blind to the original assessment. Discrepancies were noted and the inter-user agreement assessed.

5.2.1.4 Verification of the principle of name pairs with meaning being unlikely to be paired

A fundamental principle of the proposed de-identification system was that two names with a real-world meaning were unlikely to occur together as a name. The de-identifier name dictionary, developed as described in section 5.3.4, was used to annotate names extracted from the RCVS register of Veterinary Surgeons (Royal College of Veterinary Surgeons 2015) as to whether the first fore name and last name were present within its dictionaries and had been flagged as a non-safe word, that is a word that carried relevant meaning and as such should only be redacted by the de-identifier if it occurred within a name or address pattern. Those names that consisted of non-safe first and last names were quantified.

5.2.1.5 Comparing efficacy and processing speed of dictionary based single word matching methods

In the development of de-identification software able to work on a large scale; with dictionaries in excess of a hundred thousand entries and datasets containing millions of consultation narratives, a balance of speed and efficacy was a key determinant of methodologies able to be implemented in a real-world functional software tool. The aim of this experiment was to determine the method of reliable word matching with the shortest processing time in specified circumstances.

The processing-speed of two pattern-matching methods were tested. Many variables may affect processing speed and the reliability of the matching process, those able to be controlled, or counterbalanced by an alternative approach, were included in this experiment. Manner of searching was the

fundamental difference to be tested, two key Python methods of matching dictionary entries were evaluated;

a) Pandas Series.isin() method, which returns a boolean series showing whether each element in the series is exactly contained in the passed sequence of values
b) Regular Expression findall() method which looks for regular expression matches in a string.

A Python script was written to test each of four word finding techniques with a word series ordered in ascending and then descending word length order (Table 5.2.a). Four remaining variables; size of dictionary: word order; narrative length and manner of joining the dictionary; were incorporated into the experiment such that each manner of searching was examined with one of two methods of joining, short and long dictionaries and narratives and with a dictionary ordered by ascending and descending word length.

Test phrases were generated using a corpus of words drawn from the text of the Wikipedia page regarding Monty Python (wikipedia.org/wiki/Monty_Python). An identifier-free experimental corpus was generated by extracting alphanumeric sequences from the text of the page and removing any found within the word name dictionary. The corpus was reduced to 500 words at random using pandas DataFrame.sample(). The name ‘Frederick’ was chosen as the test identifier because it contained other identifiers within it, *‘Fred’, ‘Derick’* and *‘Rick’*.

Table 5.2.a: Word finding methods examined in examination of efficacy and processing speed

Experiment	Method call	Notes
findall() simple join	regex.findall(phrase)	regex formed using a simple pipe join of a list. '|'.join(dictionary as list of words)
findall() look around join	regex.findall(phrase)	regex formed using a negative look-around join to prohibit matching where a letter preceded the matching string and any letter except s followed the matching string. The same look arounds were also applied to the first and last items of the joined list using a bespoke function. '(?![a-rt-z])|(?<![a-z])'.join(dictionary as list of words)
phrase.isin(dict)	narrDf[narrDf.phrase.isin(dictSeries)] ['phrase'].tolist()	narrDf was a dataframe formed by splitting the passed phrase on white space, using `narrDf = pd.DataFrame({'phrase':re.findall('\S+', phrase)})`, and dictSeries a Pandas series containing all of the dictionary 'safe name' words. This method selects the words in the phrase that are also found in the dictionary.
dict.isin(phrase)	dictSeries[dictSeries.isin(narrDf.phrase.unique())].tolist()	Vice versa to `phrase.isin(dict)`. This method selects the words in the dictionary that are also in the phrase.

Preliminary observations with results directing methodology development

An iterative loop was created (Figure 5.2-b) whereby:

a) Ten words were sampled from the experimental corpus using the pandas `DataFrame.sample()` method
b) The ten words and a multiplier from zero to five were used, with the additional word ‘Frederick’, to generate a string in random order containing a single identifier. Where the multiplier was zero the string contained only the name 'Frederick'.
c) Each string was passed to a function to test each word identification method with the dictionary in both ascending and descending word length order. This test was repeated fifty times for each string and each test.
d) The words identified, mean time taken and the standard deviation of the mean were tabulated for each test and string.

The loop was repeated 100 times using two dictionaries:

a) All ‘safe words’ within the de-identifier dictionary, this included 172,113 words
b) A four-word dictionary containing the name ‘Frederick’ and its derivatives.

This generated information regarding the processing time and efficacy for each method with incremental lengthening of string whilst controlling for the number of identifiers present and words liable to interact with the dictionary.

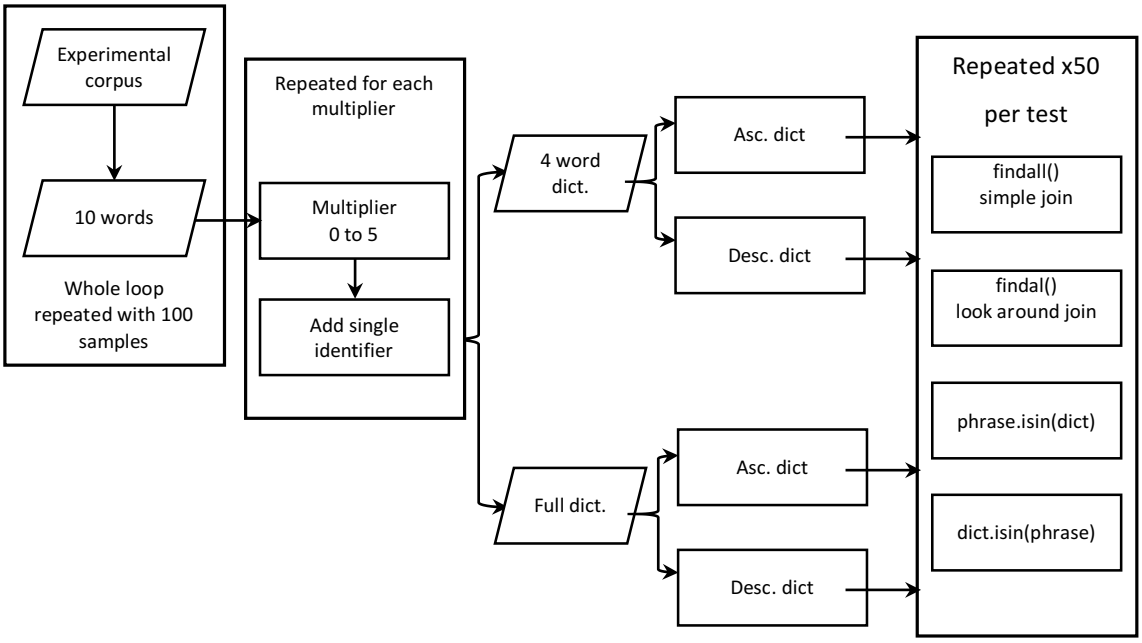

Figure 5.2-b: Iterative loop used to examine the effect of word finding technique, the word dictionary, stored as an ordered series and narrative length on processing speed and word finding efficacy

5.2.2 Results of preliminary work

5.2.2.1 Manual de-identification efficacy

The median time taken to de-identify each of the 300 median length narratives was 13.39 seconds, range 3.87s to 68.66s. Of seventeen missed identifiers (false negatives), all occurred after the 48th narrative in the batch (Figure 5.2.c), χ^2 = 16.66, p<0.001 for a difference in error rate in the first fifty narratives of the sample compared to the second fifty.

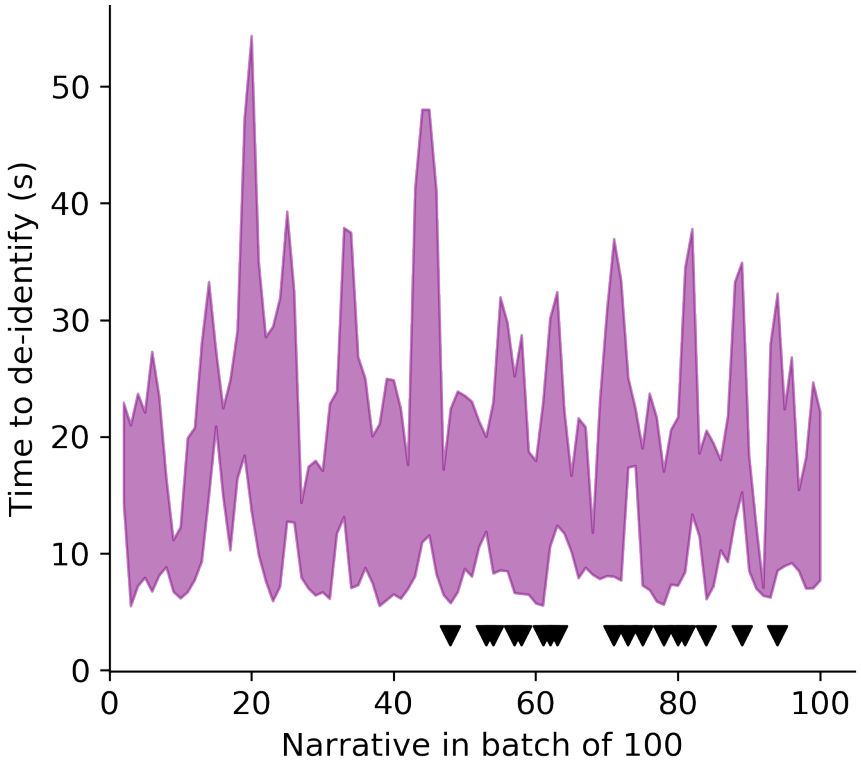

Figure 5.2.c: Durability of semi-manual de-identification. The shaded area represents the time taken to de-identify using a semi-manual approach over three samples of 100 narratives. The triangular markers represent narratives where an identifier was missed.

5.2.3 Quantification of identifying information present

Within the sample of 1,000 consultation narratives, 43% contained at least one identifying word or phrase, 8.8% contained two or more, different, identifiers. In the following descriptions, identifiers occurring more than once, in the same format, within an individual narrative were counted only once (Figure 5.2.d).

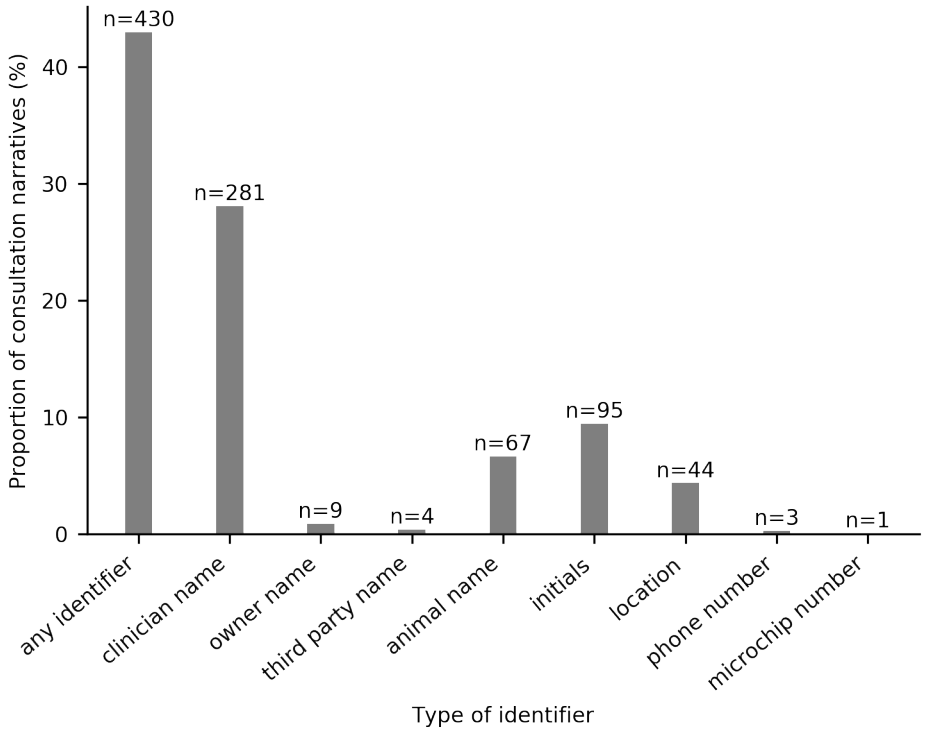

Figure 5.2.d: Nature and quantity of identifiers present within a sample of 1000 veterinary consultation narratives drawn at random from the SAVSNET dataset. Visualised as the proportion of consultations containing at least one identifier of given type

Of the 293 human names present, 286 (97.6%) included at least an initial and surname, 279 (92.8%) included both first and second names. The majority of names identified (98.0%) were that of the treating clinicians.

Initials or title and initial pairs, without a name, were not counted as names. Ninety-five (9.5%) narratives contained initials, these appeared to be those of treating clinicians and colleagues, there were four occurrences of title and single initial, usually appearing to refer to the owner. In addition, 67 narratives contained the name of an animal, with 70 animal names identified. All but one of these names were a lone first name. Locations were present in 44 (4.4%)

narratives, these were a combination of travel destinations and other veterinary practice names or their geographic location.

5.2.4 Inter-operator validation of manual de-identification

The two samples of 100 consultation narratives contained 50 and 55 identifiers respectively. On first reading 12%, 7.2% and 4.8% of identifiers were missed by each of the three human readers, giving a mean sensitivity of 92%. On second reading of those narratives where inter-person discrepancies had occurred, there was agreement between readers on all identifiers except one that was recognised but considered a clinical acronym by one reader and a set of initials by the other. With the exception of this latter string, whose true meaning could not be verified, there were no apparent false positive manual de-identifications.

5.2.5 Verification of principle that words with meaning in the veterinary sublanguage are unlikely to be paired within a name

On examining the intersection of the de-identifier dictionaries and the RCVS register, of 30,990 entries in the register, representing 29,661 unique names, 46 were composed of both a first name and a last name that had been annotated as a non-safe word (see section 5.3.4.2). These 46 names were composed of 17 different first names and 42 different last names. Thus within this list of names the occurrence of name pairs where both carried meaning important within the veterinary sublanguage was 0.15 (95% CI: 0.11, 0.19)%. Context related techniques would be used to augment redaction of such names within the final developed software.

5.2.6 Effect on words found of dictionary word length order

The only whole word present in both the experimental corpus and the dictionaries was the inserted word ‘Frederick’, however where words were permitted to match part words, as with the findall() simple join method, several other words were identified (Table 5.2.b & 5.2.c). This had implications for the precision of any de-identification method relying on this method of word matching.

The least specific method was findall() simple join with a dictionary in ascending word length order. This method found eleven incorrect words within the eleven word phrase, whilst also failing to extract the inserted name. Using a look

around join the problematic lack of precision of the findall() method was resolved, with the exception of where words matched a stem followed by an 's' as demonstrated with 'conte' in Table 5.2.b. The Pandas based methods, phrase.isin(dict) and dict.isin(phrase), both identified only the inserted name, and this was identified regardless of dictionary word length order.

Table 5.2.b: Comparison of words identified by each of four methods using ascending and descending dictionary word length order. Phrase generated from the experimental corpus: 'Animations to theatrical or a would innovative the television Frederick contest'. Dictionary was the full de-identifier dictionary.

	Words identified using full identifier dictionary	
Word identification method	**Ascending sort order**	**Descending sort order**
findall() simple join	['anim', 'thea', 'ric', 'ould', 'innova', 'el', 'ev', 'isio', 'Fred', 'eri', 'conte']	['anima', 'thea', 'rica', 'ould', 'innova', 'ele', 'isio', 'Frederick', 'conte']
findall() look around join	['Frederick', 'conte']	['Frederick', 'conte']
phrase.isin(dict)	['frederick']	['frederick']
dict.isin(phrase)	['frederick']	['frederick']

Table 5.2.c: Comparison of words identified by each of four methods using ascending and descending dictionary word length order. Phrase generated from the experimental corpus: 'Animations to theatrical or a would innovative the television Frederick contest'. Dictionary containing 'fred', 'rick', 'derick', 'frederick'

	Words identified using short identifier dictionary	
Word identification method	**Ascending sort order**	**Descending sort order**
findall() simple join	['Fred', 'rick']	['Frederick']
findall() look around join	['Frederick']	['Frederick']
phrase.isin(dict)	['frederick']	['frederick']
dict.isin(phrase)	['frederick']	['frederick']

5.2.7 Comparison of processing time for each of four methods using ascending and descending dictionary word length order.

The processing time for regular expression methods was shortest where the dictionary was small and string short. When the dictionary was large there was a notable improvement in processing time with the dictionary in descending word-length order (Table 5.2.d, Figure 5.2.e).

Table 5.2.d: Comparison of processing time for regular expression and Pandas word-finding methods. Using a large dictionary in ascending and descending word length order and a string of 1 and 51 words.

Experiment	Processing time (seconds) Ascending word length (95% CI)			Descending word length (95% CI)		
Single word string (multiplier 0)						
findall() simple join	0.0066	(0.0063,	0.0070)	0.0006	(0.0005,	0.0006)
findall() look around join	0.0092	(0.0086,	0.0098)	0.0022	(0.0020,	0.0024)
phrase.isin(dict)	0.0185	(0.0170,	0.0201)	0.0183	(0.0166,	0.0199)
dict.isin(phrase)	0.0110	(0.0104,	0.0115)	0.0109	(0.0104,	0.0114)
51 word string (multiplier 5)						
findall() simple join	0.5843	(0.4366,	0.7320)	0.6549	(0.5163,	0.7935)
findall() look around join	3.7096	(2.6818,	4.7375)	3.7201	(2.6970,	4.7432)
phrase.isin(dict)	0.0184	(0.0169,	0.0199)	0.0189	(0.0170,	0.0208)
dict.isin(phrase)	0.0117	(0.0111,	0.0124)	0.0116	(0.0110,	0.0123)

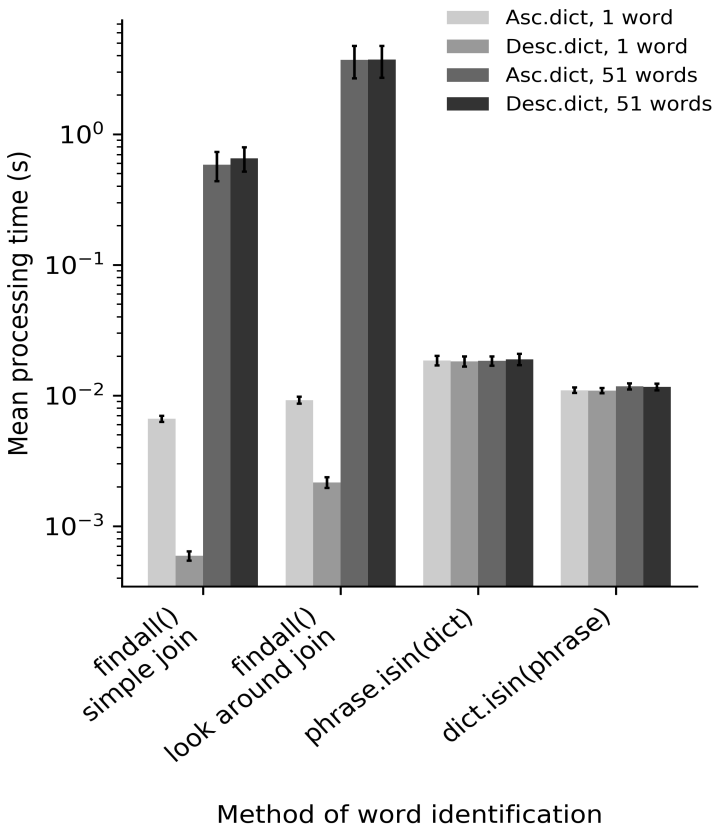

Figure 5.2-e: Comparison of processing time for each of four methods using ascending and descending dictionary word length order. Full identifier word dictionary with phrase composed of a single identifier word, 'Frederick', and the same identifier with 50 non-identifier words.

Preliminary observations with results directing methodology development

Where the dictionary was large and string long (51 words) look-around joins, i.e. joins containing lookaround assertions, increased the processing times to 6.35 times that of a simple join regular expression method (Table 5.2.d). In this situation, the processing time of the Pandas methods, 0.0184 (95% CI: 0.0169, 0.0199) seconds, was 200-fold shorter than that of the comparable regular expression method, 3.7096 (95% CI 2.6818,4.7375) seconds. The effect of look-around joins was less pronounced where the string was short and not appreciable where the dictionary was small (Table 5.2.e, Figure 5.2.f).

Where the dictionary was small regular expression methods consistently outperformed Pandas methods (Figure 5.2.e), the opposite was true with the large full dictionary, except where the string was a single word (Figure 5.2.f).

The Pandas methods were less affected by the length of string being searched, dictionary order and size of dictionary than regular expression methods. Where the dictionary was large, processing time was improved by using the `phrase.isin(dict)` method rather than its converse, `dict.isin(phrase)` (Figure 5.2.g). Where the dictionary was small this appeared no longer relevant.

Table 5.2.e: Comparison of processing time for regular expression and Pandas word-finding methods using a dictionary containing 4 words.

Experiment	Processing time (seconds)			
	Ascending word length (95% CI)		Descending word length (95% CI)	
Single word string (multiplier 0)				
findall() simple join	1.36×10^{-6}	(9.58×10^{-7}, 1.77×10^{-6})	9.11×10^{-7}	(5.64×10^{-7}, 1.26×10^{-6})
findall() look around join	1.08×10^{-6}	(4.77×10^{-7}, 1.68×10^{-6})	9.61×10^{-7}	(4.95×10^{-7}, 1.43×10^{-6})
phrase.isin(dict)	1.01×10^{-3}	(8.75×10^{-4}, 1.15×10^{-3})	1.00×10^{-3}	(8.82×10^{-4}, 1.12×10^{-3})
dict.isin(phrase)	1.07×10^{-3}	(9.16×10^{-4}, 1.22×10^{-3})	1.07×10^{-3}	(8.85×10^{-4}, 1.26×10^{-3})
51 word string (multiplier 5)				
findall() simple join	2.29×10^{-5}	(1.59×10^{-5}, 2.98×10^{-5})	2.20×10^{-5}	(1.56×10^{-5}, 2.85×10^{-5})
findall() look around join	4.10×10^{-5}	(1.23×10^{-5}, 6.98×10^{-5})	3.66×10^{-5}	(2.54×10^{-5}, 4.78×10^{-5})
phrase.isin(dict)	1.04×10^{-3}	(8.83×10^{-4}, 1.20×10^{-3})	1.02×10^{-3}	(8.18×10^{-4}, 1.22×10^{-3})
dict.isin(phrase)	1.13×10^{-3}	(8.49×10^{-4}, 1.41×10^{-3})	1.09×10^{-3}	(9.37×10^{-4}, 1.25×10^{-3})

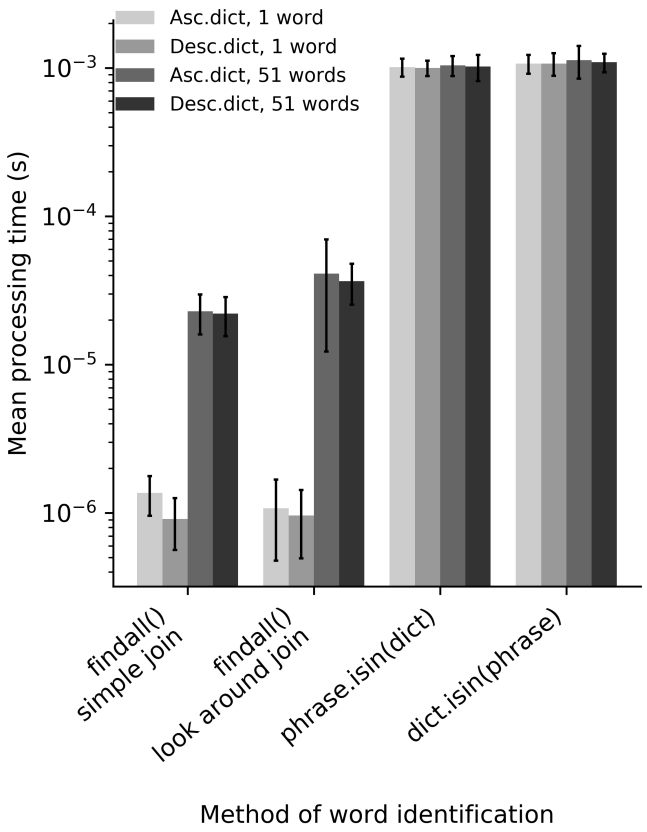

Figure 5.2-f: Comparison of processing time for each of four methods using ascending and descending dictionary word length order. Four words identifier dictionary, [‘fred’, ’rick’, ‘derick’, ‘frederick’] with phrase composed of a single identifier word, ‘Frederick’, and the same identifier with 50 non-identifier words.

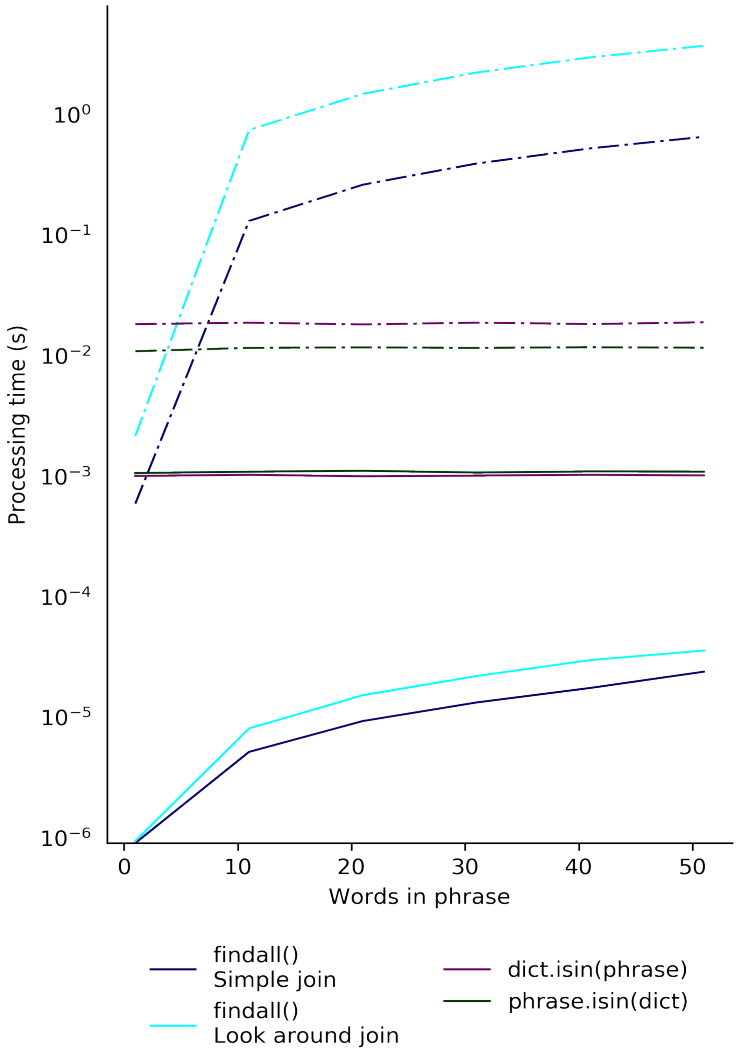

Figure 5.2.g: Comparison of processing time for regular expression and Pandas based identification methods with increasing phrase length. Dotted lines represent processing time when the dictionary consists of the full identifier word dictionary, solid line the four word dictionary. Word variability is controlled for by randomly sampling to generate ten words and repeating the same ten words with the insertion of a single identifier to build longer strings. Plot represents 100 iterations with each method being run ten times per iteration. Error bars not visible when plotted (process outlined in Figure 4.2.a

5.2.8 Summary of outcomes from method evaluation and optimization

The manual de-identification assessment demonstrated that the reliability of manual de-identification, even when augmented by software, began to decline after approximately fifty narratives had been processed. This reiterated the need for an automated process, on the basis of speed, endurance, and efficacy. Comparison of de-identification between 3 parties confirmed that the author's identification of identifiers was in keeping with that of others, and suggested that repeating manual de-identification twice, in small batches, may avoid missing identifiers.

The experiments comparing word matching efficacy demonstrated the shortfalls of simple join regular expression matching, with the identification of several part-word matches and only part match of the inserted identifier where the word length order was ascending. Regardless of speed this limited their use within a de-identification system.

Although look around joins reduce the processing speed they vastly improve the precision of regular expression based matching. The ideal appears to be to use Pandas methods where a large dictionary is mandated and regular expression methods, with look around joins, where a small dictionary is feasible.

5.3 Development of Clancularius, the de-identifier

5.3.1 Software

The Python programming language was used as the basis of all software developed, combining a number of techniques. Web scraping (data extraction) used the Requests and BeautifulSoup libraries. PubMed queries were facilitated by the Biopython Entrez module, permitting direct query of the PubMed database (National Library of Medicine 2016). Where records were available for download as a csv file, these were loaded into a Pandas dataframe for processing. All extracted text was processed using a combination of Pandas, Numpy and Regular Expression based tools (Table 5.3.a for versions used).

Table 5.3.a: Libraries utilised during the development of de-identification software. Versions are those used during most recent development.

Library	Version	Available at	Description of application during de-identifier development
Python	3.6.1	www.python.org/	Programming language
Regular Expression	2.2.1	Integral to Python	Pattern matching, extraction and substitution
Pandas	0.19.2	pandas.pydata.org/	Provided the main data handling and analysis backbone
Numpy	1.12.0	www.numpy.org/	Scientific computing library, primarily used for its np.where function which behaves more intuitively than that of Pandas
BeautifulSoup	4.5.3	pypi.python.org/pypi/beautifulsoup4	HTML and XML parsing
Requests	2.13.0	docs.python-requests.org	Send HTTP requests to retrieve website data
BioPython Entrez	1.65	biopython.org/	Query PubMed and parse XML retrieved

5.3.2 Generating name dictionaries

Openly available data sources (Office For National Statistics 2015a; Royal College of Veterinary Surgeons 2015; Scottish Government 2015; National Library of Medicine 2016) were used to identify names likely to be present within the United Kingdom general and veterinary population (Table 5.3.b). The dictionary was fluid and could be amended with addition of names from other sources in future where this was required, or desired.

First names were retrieved from downloadable text (.csv) files openly available from the Scottish Government and Office of National Statistics websites. The Scottish data contained all registered first forenames, however the England and

Wales data listed the most frequently occurring names, which was two pronged. Most frequently occurring first names were the most likely to be found, they were however also the least individually identifiable. Additionally, these were baby names and thus a generation ahead of most owners and clinicians. Further sources of names were therefore required.

Table 5.3.b: Data sources utilised in creating name-word dictionaries

Data	Organisation	URL	Extraction method
Register of Veterinary Surgeons & Veterinary Nurses	Royal College of Veterinary Surgeons	www.rcvs.org.uk	Web scraping
Baby Names, England and Wales, 2014	Office for National Statistics	ons.gov.uk.	Direct download
National Records of Scotland	Scottish Government	gro-scotland.gov.uk	Direct download
PubMed	National Library of Medicine	www.ncbi.nlm.nih.gov	Bio Entrez query

5.3.2.1 Retrieving names of clinicians registered with the RCVS

In the UK, the Royal College of Veterinary Surgeons (RCVS) is responsible for regulation of the veterinary profession. The RCVS website hosts a searchable database of all registered Veterinary Surgeons and Veterinary Nurses. These provide a source of clinician names likely to appear within UK clinical records. A basic script utilising the BeautifulSoup and Requests libraries was used to retrieve a list of the full names of each clinician within these registers (Figure 5.3.a).

```
import requests, re
from bs4 import BeautifulSoup
import pandas as pd

def scrapeRCVSNames():
    ''' scrapes names from the veterinary surgeon and veterinary nurse
        registers published on the RCVS website
        review the URL and the pattern passed to find_all as page structure
        may change with time. Returns a list of full names'''
    urls = ['http://www.rcvs.org.uk/find-a-nurse/search/'+
            'filter-keyword=&sortby=&filter-choice=name&sortby=name&p=',
            'http://www.rcvs.org.uk/find-a-surgeon/search/'+
            '?filter-keyword=&sortby=&sortby=name&p=']
    soupNameRegex = re.compile('">(.*)</strong>')
    nameList = []
    for num in range(1,3100):
        for url in urls:
            try:
                res = requests.get(url+str(num))
                soup = BeautifulSoup(res.text, 'lxml')
                for thing in soup.find_all("h3", class_ ="subjectNames"):
                    thing = re.sub('(<strong>)', '', str(thing))
                    nameList.append(soupNameRegex.search(str(thing)).group(1))
            except Exception as e:
                print(e)
    return nameList


def splitNames(row, nameRegex):
    ''' Extract title and last name as a string, first names as a list
        This may capture some last names as first names, however the manner
        of name pattern extraction renders this non-problematic'''
    row['title'] = nameRegex.search(row.fullName).group(1)
    row['firstName'] = re.findall('\w+',
                                  nameRegex.search(row.fullName).group(2))
    row['lastName'] = nameRegex.search(row.fullName).group(3)
    return row

def extractAllNames():
    ''' generates arrays of unique first name, last name and title words
        using scrapeRCVSNames() '''
    nameList = scrapeRCVSNames()
    nameDf = pd.DataFrame(index = pd.Series(nameList).str.lower(),
                              columns = ['title', 'firstName', 'lastName'])
    nameDf.index.name = 'fullName'
    nameDf = nameDf.reset_index()
    nameRegex = re.compile('^\W*(?:\s?(\w+)\W+)?(\w+(?:\W*\w+)*)\W+(\w+)\W*$')
    nameDf = nameDf.apply(lambda row:splitNames(row, nameRegex), axis = 1)
    titles = nameDf.title.str.lower().unique()
    firstNames = pd.Series(pd.Series([leaf for tree in
                        nameDf.firstName.tolist() for leaf in tree]).unique())
    lastNames = pd.Series(nameDf.lastName.unique())
    return titles, firstNames, lastNames

titles, firstNames, lastNames = extractAllNames()
```

Figure 5.3.a: Code used to extract the names of veterinary surgeons and nurses from the RCVS

This script iterates over the pages generated by a complete wildcard search, equivalent to searching with an empty search box, and locates the string within the HTML containing clinician names. The list of names, nameList, returned by this function consists of names in the format: *title first name middle names last name,* for example *Dr Jenny Ann Newman*. For reasons of functionality, and data protection, the de-identification process required individual name words, and did not at any point store full names.

The list output by scrapeRCVSNames was passed directly to the extractAllNames function shown in Figure 5.3.a. This utilised components of the Pandas and Regular Expression libraries to parse the full names into discrete name word entities, no longer linked in their original sequence.

5.3.2.2 Names of authors cited within PubMed database

Author names were chosen as they represent the current diverse adult population and thus, it was anticipated, would augment the completeness of the de-identifier dictionaries. There are many citation indices, PubMed was chosen as it has open access and can be queried directly using the Entrez library of BioPython.

Bespoke code (Figure 5.3.b) was written to retrieve lists of first and last names found within the author fields of citations published between 2000 and 2016 from an institution with an affiliation within the United Kingdom or Eire. Names extracted from PubMed introduced a large number of words that were in common general usage, and not used as names, in the UK. In an attempt to reduce the impact of these extraneous words, whose presence increase the risk of data quality degradation, only words present at least three times, within the citation subset returned by the query, were added from the PubMed dataset.

```
from Bio import Entrez, Medline
import pandas as pd
Entrez.email="A.N.Other@example.com"     # Always tell NCBI who you are

def retrieveXML(minDate = 2000, maxDate = 2015):
    ''' Generates a list of names of authors with UK affiliations
        Uses Bio Entrez search to retrieve a list of UIDs from PubMed for
        articles published between 2000 and 2016 with a UK or Eire affiliation
        retrieves the XML entry for each article and extracts the contents of
        the FAU, author information, field. Returns a list of author names '''
    handle = Entrez.esearch(db='pubmed',
                                        term='''(((((UK[Affiliation])
                                        OR United Kingdom[Affiliation])
                                        OR England[Affiliation])
                                        OR Scotland[Affiliation])
                                        OR Wales[Affiliation])
                                        OR Ireland[Affiliation]''',
                                        datetype='pdat',
                                        mindate=(minDate),
                                        maxdate=(maxDate),
                                        retmax = 10000000)
    idlist = Entrez.read(handle)["IdList"]
    handle = Entrez.efetch("pubmed",
                           id=idlist,
                           rettype="medline",
                           retmode="text",
                           retmax = 10000000)
    records = Medline.parse(handle)
    nameList = []
    for record in records:
        nameList.extend(record['FAU'])
    return nameList

def extractNameWords(minNum = 1):
        ''' uses nameList to generate arrays of unique first and last names
        extracts the first word of each name entry as the last name
        extracts the first word after a comma as the first name entry
        provided it is more than minNum characters long '''
    nameList = retrieveXML()
    df = pd.DataFrame()
    df['fullName'] = nameList
    df['lastName'] = df['fullName'].str.extract('^(\w+)(?=,)', expand = False)
    df['firstName'] = df['fullName'].str.extract(',\s*(\w{2,})', expand =
False)
    if minNum == 1:
        lastNames = df.lastName.str.strip().str.lower().unique()
        firstNames = df.firstName.str.strip().str.lower().unique()
    else:
        counts = pd.DataFrame(df.lastName.value_counts()>minNum)
        lastNames = df[df.lastName.isin(counts[counts.lastName==True
                                               ].index)].lastName.unique()
        counts = pd.DataFrame(df.firstName.value_counts()>minNum)
        firstNames = df[df.firstName.isin(counts[counts.firstName==True
                                               ].index)].firstName.unique()
    return lastNames, firstNames

lastNames, firstNames = extractNameWords()
```

Figure 5.3.b: scrapePubMed.py code generated to produce lists of first and last names within citations, published between 2000 and 2016 with a UK or Eire affiliation, held by PubMed.

5.3.2.3 Pet names from online sources

Many animals are named with human first names and it was anticipated that the human name lists would capture the majority of animal names. In an attempt to augment this and capture more animal specific names, online repositories of pet names were accessed via BeautifulSoup. Names not identified by preceding methods within these lists were extracted and processed in the same manner as human first names.

5.3.3 Location words and strings

A location gazetteer (Office For National Statistics 2015b) was used to generate a dictionary of place name words and strings. Address components were handled as individual words (sequences of characters) except where these comprised multi-word locations or road names which were handled as multi-word strings (sequences of words).

The Royal College of Veterinary Surgeons website (Royal College of Veterinary Surgeons 2015) was used as a source of neologisms used to name veterinary practices. These were retrieved using code analogous to that shown in Figure 5.3.a. Words and bigrams were extracted from each veterinary practice name.

Country and capital city locations were acquired from the GeoNames geographical database (Wick 2017) and added to the location dictionary, in an attempt to redact travel plans and further minimise the likelihood of identifying any individual based on prior knowledge.

5.3.4 Converting name lists to a functional identifier dictionary

5.3.4.1 Diacritic letters

Names containing accents were problematic as encoding of narratives during and prior to de-identification was likely to cause false negatives, or de-identification failure, where an encoded byte was no longer recognised as a string. To address this, any diacritic letter in the dictionaries was normalised to the nearest plain ASCII letter, as is common practice (Holzinger et al. 2014).

This step was chosen because diacritic letters cause issues when reading and parsing datasets. In usual UK veterinary practice, at whose clinical records this software was targeted, the same approach would be likely for UK clinicians

entering the clinical narrative. Accents were only considered likely to be present where they were contained within automated or pasted text, as may occur if the clinician or practice name contained an accent.

As an adjunct to this decision, any word within narrative strings passed to the software for de-identification containing an accent was considered an identifier.

5.3.4.2 Examining the intersection of vocabulary and identifiers

English language and veterinary language vocabulary was generated utilising data extracted from online sources including the webpages of the British Broadcasting Corporation (British Broadcasting Corporation 2018) Merck Veterinary Manual (Merck Sharp & Dohme Corporation 2018), multiple online discussion fora and abbreviation lists, later augmented with the vocabulary found within the narratives of the SAVSNET corpus. The intersection of the identifier and vocabulary collections, i.e. words appearing in both, was examined manually and divided into words considered ‘safe’, all of which would ultimately be redacted from the clinical narratives and those considered ‘non-safe’ which would only be redacted if present within the constraints of specified patterns. This process was somewhat subjective and relied on domain knowledge for the use of words and context.

Words were tagged as first names, last names, and place names according to their source, with overlap where they fell into more than one category. Any word that was a first name and had been classified ‘non-safe’ was further screened and if considered likely to represent a first name if the first letter was capitalized was assigned as ‘secondary safe’, in addition to ‘non-safe’ (Table 5.3.c). The word dictionaries were sorted by word length in descending order. The resultant dictionary, stored as a text document, is exemplified in Table 5.3.d.

Place and road names consisting of more than one word were identified and their constituent words screened in the same manner. The first word of these strings was used as its index, unless this was a common English language word when the second or last word was used. Words that occur at such high frequency within the English language so as not to be of classification value (stopwords) were removed from the individual word lists, but not from the multiword place names.

This process generated dictionaries of first, last and place name words, which were each subdivided into ‘safe’, ‘nonSafe’ and ‘secondarySafe’ words, and in addition road and settlement names consisting of multiple words.

Table 5.3.c: Meaning of terms used in dictionary and de-identifier development

Term	Meaning	Application	Examples
Safe word	A word considered to always represent a name or location	All occurrences redacted from narratives	Fitzgerald Jemima
Non-safe word	A word considered to represent both an identifier and a word of relevance within the narrative	Only redacted where present within a name or location pattern	Ball Farrier
Secondary safe word	A first name word considered likely to represent a first name if begins with a capital letter	Redacted regardless of pattern if begins with capital letter	Bill Coco Blue
Multi-word place	A place name consisting of several words	Indexed by first and last word and redacted as a string	Chester High Road South Normanton

Table 5.3.d: Example of initial word dictionary partition into safe and non-safe words

word	wordLen	safeWord	nonSafe	secondarySafe	firstName	lastName	placeName
eastergate	10	1	0	0	0	0	1
shanilka	8	1	0	0	1	0	0
balding	7	0	1	0	0	1	1
alijah	6	1	0	0	1	0	0
bill	4	0	1	1	1	1	1

5.3.5 Simple pattern matching

Regular expressions, search phrases matching patterns of characters and character ranges, were developed to identify strings likely to represent telephone, microchip and passport numbers, postcodes and email addresses (Table 5.3.e). The postcode regular expression was evaluated against the Office of National Statistics Postcode Directory (Office For National Statistics 2015b) and matched all UK postcodes that had been active since the year 2000. In order to distinguish postcodes from other alphanumeric codes, such as vaccination batch numbers, this regular expression required the two-part UK postcode to contain a space, as is convention, and included basic contextual cues. The requirement for a space to be present could be amended dependent on the de-identification requirements of a given dataset.

Table 5.3.e: Regular expressions utilised for simple pattern matching within Clancularius de-identification process

Identifier type	Regular expression
Phone number	(?<!\w)(?:(?:(?:[+]?[0-9]{2}\W{,2}0\W{,2})|0)\d{9,11}| (?:(?:[+]?[0-9]{2}\W{,2}0\W{,2})|0)\d{2,4}\s?\d{6,7}| (?:(?:[+]?[0-9]{2}\W{,2}0\W{,2})| 0)\d{2,4}\s?\d{3,4}\s?\d{3,4})(?![0-9])
Microchip number	(\d{12,17})| (?=(?:chip|mc))[a-z\W]{,20}([\d\s]{24,})| ((?:\d{3,4}(?:\s|\-)*){5,6})
Passport number	(?:(?<!\w)(?:gb|irl|ie)\s?\d{6,8})|(p\w*port\s*\w{2,3}\W?\d+)
Postcode	(?<!\w)((([A-PR-UWYZ][0-9])|([A-PR-UWYZ][0-9][0-9])| ([A-PR-UWYZ][A-HK-Y][0-9])|([A-PR-UWYZ][A-HK-Y][0-9][0-9])| ([A-PR-UWYZ][0-9][A-Z])|([A-PR-UWYZ][A-HK-Y][0-9][ABEHMNPRVWXY])) \s+([0-9][ABD-HJLNP-UW-Z]{2})|GIR\s*0AA) (?!\sdose)(?!dose)(?!\svac)(?!vac)(?!\sboos)(?!boos)(?!\w)
Email	([a-z0-9_\.+\-]+@[a-z0-9\-_]+\.[a-z0-9\-_]+(?:\.[a-z0-9\-_]+)?)

5.3.6 Domain specific preservation of information

Once simple alphanumeric patterns had been redacted, a mask was applied to the narrative. This preserved clinically important and research specific terms by masking specific words and phrases from the de-identification process, and thus ensuring they were not incorrectly redacted in the de-identified output (Figure 5.3.c).

A generic mask was used at all times, to preserve stopwords such as *at* and *to*, and abbreviations that may inadvertently form a name pattern and clinical strings such as *x ray*. This surrounded each letter of the masked word with two angle brackets and the letter z, rendering 'at' as '<<zaz>><<ztz>>' for example, and thereby creating nonsense strings which would not match dictionary words or context patterns, whilst being visible to the developer during fine tuning of pattern recognition. Although this did risk leaving residual identifiers where they formed a clinical abbreviation, for example r hind, it seemed unlikely that such name forms would be used where they could be misinterpreted as having their clinical meaning. The same process was used to mask the word preceding any occurrence of the words 'disease' or 'syndrome' (allowing for considerable abbreviation and misspelling) to protect from the inadvertent redaction of eponymous syndromes.

Figure 5.3.c: Example of masking process, cloaking words and phrases likely to cause false positive de-identification. In example 1, without masking, *r hind* and potentially *to r hind* would be recognised as a name pattern. In Example 2, *rt horners* would be redacted as a name pattern, *underlying disease* is cloaked to prevent eponymous names not recognised during the dictionary building process, where *cushing* and *horner* were both tagged as 'unsafe', being redacted during de-identification.

Example 1	Masking process
Original	Inj to r hind
Masked	inj <<ztz>><<zoz>> <<zrz>> <<zhz>><<ziz>><<znz>><<zdz>>
De-identified	Inj to r hind
Example 2	
Original	Wilf Cushing, rt horners ?underlying disease
Masked	Wilf Cushing, <<zrz>><<ztz>> horners ?<<zuz>><<znz>><<zdz>><<zez>><<zrz>><<zlz>><<zyz>><<ziz>><<znz>><<zgz>> <<zdz>><<ziz>><<zsz>><<zez>><<zaz>><<zsz>><<zez>>
De-identified	<<name>>, rt horners ?underlying disease

5.3.7 Research specific preservation of information

In addition to the division of words as safe or non-safe, a method was built into the infrastructure of Clancularius to facilitate the application of a research-specific mask, tailored to ensure that terms of importance to a specific data application were preserved. For example, if tumour size was of specific importance to a piece of research, common entities often used as real-world

size comparators would be preserved. As food imagery is common place in clinical language (*'grape sized lump', 'caseating granuloma', pear-shaped'*) this would include words such as *plum, grape* and *melon*. Likewise, if diet was of research importance food stuffs and brands could be preserved.

This process relies on the relative commonality of words in subsets of the data, and the redaction of sufficient other identifiers, where a masked term occurs as an identifier, to impair linkage to an individual. The purpose of the masking process was to minimise research critical false positives, this was not permitted to take precedence over the need for highly sensitive identifier redaction.

It was important that if this adjunct method was used an extensive list of terms was acquired, this could be generated either from proprietary lists of known words or via information extraction from online sources to ensure inclusion of all likely terminology and avoidance of introducing familiarity bias.

The software was designed to render it feasible to process a large corpus using the standard mask, with subsequent reprocessing of only those consultations containing the research critical terms where a specific additional mask was later required. In this manner, production of a research-tailored identity-protected corpus could be more responsive.

5.3.8 Resolution of interaction with brand and breed names

Pet food and pharmaceutical brands are commonly formed from words present within person or place names. These are, for most purposes, not relevant to the data user. However, redaction of these words may detract from the ease of understanding the narrative, reducing data quality.

A dictionary of brand names was created. Where brand names created name or location patterns unlikely to be genuine names or locations these were masked prior to de-identification and thus the information preserved. Where individual words such as *'Hills'* were potentially used as both a brand name and a human identifier, these were redacted if they remained after name and place pattern redaction, in the same way as a 'safe name' word. The context preserver *'<<brand name or identifier>>'* was used to avoid data quality degradation by suggesting that a brand was an identifier. A similar process was undertaken with

breed names, which often contained place and occasionally human names, for example *Gordon Setter* and *Norfolk Terrier*, where a contextual marker indicated that the identifier was a breed the <<*breed name or identifier* >> tag was not applied, but if *Gordon Setter* appeared as *Gordon* it would be redacted.

5.3.9 Identifier context preservation

Where an identifier was redacted it was replaced with a context preserving placeholder, for example '<<name>>', '<<location>>'. As there is considerable overlap between place and person names; where these occurred outside a text pattern that was able to reliably discriminate between the two, the placeholder '<<identifier>>' was used. This reduced the likelihood of misidentifying a person as a place and vice versa within a clinical narrative field where notation was frequently sparse and grammar non-standard. Location was taken to include client addresses, veterinary practices and geographic locations, but excluded retail outlets, animal welfare organisations and statutory bodies. The nature of the latter establishments was likely to be of research importance and their inclusion did not risk identifying an individual nor clinical practice. They could however be readily added to the dictionary if required.

5.3.10 Name pattern redaction

The dictionaries of name and address components were loaded into data arrays. The sequence of individual words, non-word characters and spaces within an individual narrative was imported into a dataframe where each word and whitespace was represented by a row. Each word within this narrative dataframe was classified as to whether it was also present within the arrays of person and place name words, and whether it was a number or postcode.

Narrative words recognised as place or person names were used to create subsets of the identifier word arrays (dictNamesPresent{} in Figure 5.3.d). Regular expressions for name pattern identification were generated using only the words known to be present in the individual narrative being processed. This had considerable processing speed advantage over a regular expression based search for patterns utilising the several hundred thousand potential identifier words. Person name patterns and multi-word place names were identified using regular expression based pattern matching. The key principle underlying a number of these rules was that two words that also confer a non-name meaning

are unlikely (although not impossible) to be used as a person's name, as verified in the preliminary work (Section 5.2). Place name patterns were identified within the Pandas dataframe.

Once potential name and place name patterns had been identified and redacted, any remaining place or name words that were annotated as ‘safe words’ were redacted (Figure 5.3.e)

```
def wordTypeInit(self,narr):
    if verbose:
        print(time.time() - verboseTime, 'deIdentify.wordTypeInit()')
    ''' uses the splittingRegex to generate a dataframe where each word and
        whitespace is represented by a row
        original case in the narrative column
        lower case in copyNarr columns
        uses the copyNarr column and the name and place dictionaries to
        create booleans indicating whether each word is found in each
        dictionary series.
        generates name and place dictionaries for this particular string
        returns the narrDf dataframe'''
    self.dictNamesPresent = {}
    self.dictPlacesPresent = {}
    narrDf = pd.DataFrame({'narrative':
                        dictMgmtVars.splittingRegex.findall(narr)})
    narrDf['copyNarr'] = narrDf.narrative.str.lower()
    narrDf['wordType'] = 'other'
    for component, listOfThings in dictMgmtVars.dictNames.items():
        narrDf[component] = np.where((
                        narrDf.copyNarr.isin(listOfThings)),1,0)
        self.dictNamesPresent[component] = narrDf[
                        narrDf[component]==1]['copyNarr'].unique().tolist()
    for component, listOfThings in dictMgmtVars.dictPlaces.items():
        if component != 'multiplaceNames':
            narrDf[component] = narrDf.copyNarr.isin(
                        listOfThings).astype(int)
            self.dictPlacesPresent[component] = narrDf[
                        narrDf[component]==1]['copyNarr'].unique().tolist()
    narrDf = self.wordTypeExtraction(narrDf)
    if verbose:
        narrDf.to_csv('wordTypeInit.csv', index = False)
    return narrDf

def wordTypeExtraction(self, narrDf):
    if verbose:
        print(time.time() - verboseTime, 'deIdentify.wordTypeExtraction()')
    ''' receives the individual string dataframe narrDf
        identifies those words already tagged with a wordType
        returns narrDf'''
    narrDf['wordType'] = np.where(narrDf['copyNarr'].str.contains(
                            '<<[a-y]+>>', na = False),
                            narrDf['copyNarr'].str.extract(
                            '<<([a-z]+)>>', expand = False), 'other')
    narrDf['numPlace'] = narrDf.copyNarr.str.contains(
                            '^[1-9][0-9]{0,3}[a-h]?$').astype(int)
    narrDf['postcode'] = narrDf.copyNarr.str.contains(
                            '<<postcode>>').astype(int)
    narrDf['capitalised'] = narrDf.narrative.str.contains(
                        '''^[A-Z](?=[a-z])''').astype(int)
    if verbose:
        narrDf['wordType'] = np.where(narrDf['copyNarr'].str.contains(
                            '<<z.*z>>', na = False), 'masked',
                            narrDf['wordType'])
    return narrDf
```

Figure 5.3.d: Method used to split a narrative string into a dataframe, known as narrDf within the code of Clancularius, to permit rule-based matching and minimisation of active dictionary size. The splittingRegex mentioned in `wordTypeInit()` is `[^,\-\.:;\(\)\s\'&"/?]+|[^\w]`. A slightly amended method is used when the narrative is subsequently split again later in the de-identification sequence to preserve word type assignments.

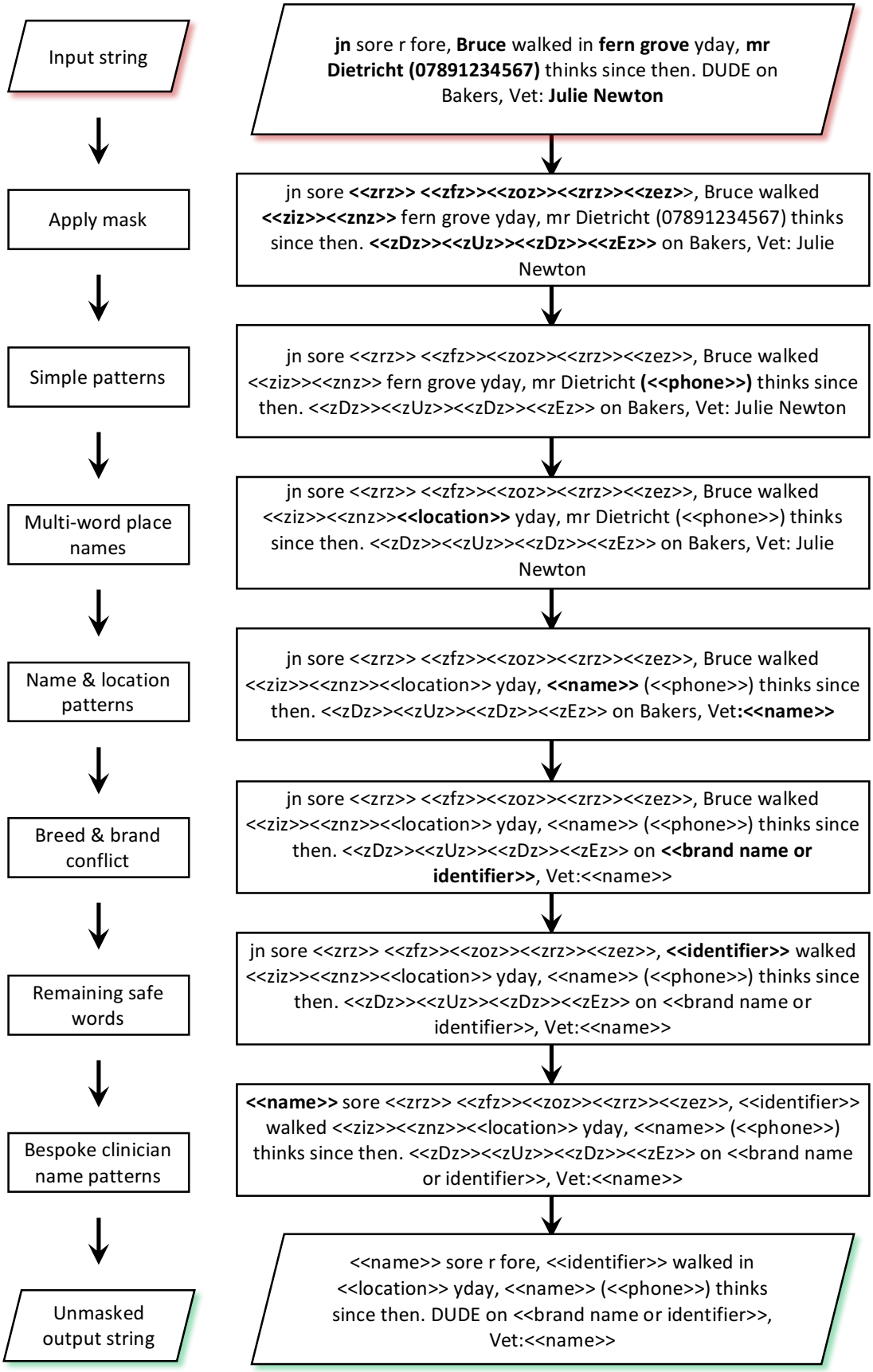

Figure 5.3.e: Outline of the Clancularius de-identification process

5.3.11 Corpus-specific refinement

The core de-identifier functionality was built in a manner intended to render the system capable of de-identification within any English language veterinary clinical narrative system in the UK. The prime reason for development, however, was to enable rapid and adaptable redaction of incidental identifiers occurring within the clinical narratives collated by SAVSNET.

It was recognised that whilst many identifiers occur within the unstructured field, a common location for full name identification of the treating veterinary surgeon was within automatically appended dispensing labels, inserted within the narrative. This provided an opportunity for the system to acquire knowledge of names that could be expected within this dataset, and to ensure that names occurring within the dispensing labels were redacted. The ease of appropriate redaction was frustrated by the marker ‘vet:’, intended to indicate the prescribing vet within the standard template dispensing label, not infrequently, occurring in the absence of a name entry. Unconstrained automated redaction of all subsequent strings would have removed potentially important non-sensitive information.

Once non-corpus specific de-identification was complete, the four words following any occurrence of the word ‘vet’ were further scrutinised. In most cases, the actual veterinarian name had been redacted by this stage. Remaining words were checked against a list of known non-name words occurring at this contextual location. Any previously unseen words were redacted.

A similar process was used for the identification of clinician initials, the first letters of any name following a clinician marker were acquired at the beginning of processing and used during the adjunct clinician name pattern scrubbing stage (Figure 5.3.e). Other contextual markers, such as two and three letter words at the beginning or end of a string, were also used to augment redaction of clinician initials. The optimal location of corpus-specific redaction within the de-identification sequence was assessed by comparing the outcome of these redaction methods whilst the generic mask was applied and following its removal.

5.3.12 Estimating efficacy of Clancularius

5.3.12.1 Efficacy of de-identification within the SAVSNET dataset

A sample of 1000 consultation narratives was drawn at random from the consultations collated within the SAVSNET dataset during the week commencing 1st April 2017. These narratives had not been previously seen by the author and were collated after corpus specific refinement had been completed, they formed an unseen validation set.

The validation set contained a total of 72,672 alphanumeric words, with a mean of 73 words per narrative, range 1 to 365 words. The process described in section 5.2.1.1 was used again to semi-manually de-identify the validation sample, in batches of 50 consultations at a time, and it was then processed using Clancularius. The outcome of manual and software de-identification were compared and the sensitivity (recall) and precision (positive predictive value) of Clancularius calculated.

The author undertook this manual coding. Her manual coding had in itself been validated by comparison to that of two colleagues over two different sets of 100 consultation narratives (Section 5.2.4).

5.3.12.2 Efficacy of de-identification within the Bristol Cats Study dataset

The Bristol Cats Study is a longitudinal study of privately-owned cats being undertaken by the University of Bristol, UK (Murray et al. 2017). Veterinary clinical records of participating animals were submitted to the study by their treating veterinary surgeon in a range of formats. Where it was possible, text was extracted from the submitted files and processed within a Python-based framework such that a dataset of narrative, consultation date and unique identifier was generated. This data had not been used at any point during the development of Clancularius and had been recorded via multiple practice management systems prior to transmission to Bristol and extraction by the first author for de-identification and subsequent text-mining within Bristol Cats Study. As the cats were actively enrolled in the longitudinal study, a record of owner, animal and clinician identifiers was available.

A random sample of 100 narrative fields was taken from the Bristol corpus and processed by Clancularius. This second validation corpus consisted of 9,316 words with a mean of 93 words in each narrative field. The de-identified output from Clancularius was compared to manual de-identification and errors classified.

5.3.12.3 Processing speed assessment

The speed of redaction by Clancularius was assessed using three machine configurations (Table 5.3.f). A testing set of 10,000 narratives was selected at random and the narratives were de-identified with mean time taken to complete the process recorded using Python's time module over 10 processing cycles. Processing speed on machine 1 was further assessed using the same method whilst running 4 instantiations of Clancularius simultaneously.

Table 5.3.f: Operating environments used to assess Clancularius processing speed.

Computer description	Operating system	Processor	RAM (Gb)	Build year
1. Macintosh iMac	macOS Sierra	3.4GHz quad core	32	2013
2. Mac mini	macOS Yosemite	2.6Ghz dual core	8	2014
3. Middle range laptop	64 bit Windows 8.1	1.6Ghz quad core	8	2014

5.4 Results

5.4.1 Dictionary

The vast majority of words within the name and place lists were classified as 'safe words'. The dictionary is readily updated in response to the corpus being used, and observed omitted or over-zealous de-identification. At the time of validation, the dictionary contained 177,621 name and location words, 2.25% of words were classified as non-safe. The multi-word location dictionary contained 328,006 location strings.

5.4.2 Sensitivity within SAVSNET validation corpus

The overall efficacy of Clancularius for redacting the information it was essential to redact (sensitivity 99 (97.6, 99.6)%) compared favourably with human de-identification (sensitivity 97.4 (95.3, 98.5)%) (Table 5.4.a). A single human name

was missed by Clancularius, this name was the only word in that particular narrative and was misspelled.

Three locations were missed, one of these was misspelled, forming a different word, one was Bath, a city in the South West of England, without other location information. In response to this omission the 'secondary safe' system was adapted to include locations as well as names. The third missed location was an acronym of a veterinary clinic.

Table 5.4.a: Efficacy of redacting identifiers within the validation set, a random sample drawn from the SAVSNET corpus. Numbers are counts of identifiers unique within a narrative. Human Missed refers to the number missed on first reading and found on second or in combination with Clancularius

		Software			Human		
	Total present	Found (tp)	Missed (fn)	Sensitivity (%)	Found (tp)	Missed (fn)	Sensitivity
Essential							
Human name	352	351	1	99.7(98.4, 100)%	349	3	99.2(97.5,99.7)%
location	57	54	3	94.7(85.6, 98.2)%	49	8	86(74.7,92.7)%
Microchip	9	9	0	100(70.1,100)%	9	0	100(70.1,100)%
Total essential	418	414	4	99(97.6,99.6)%	407	11	97.4(95.3, 98.5)%
Desirable							
Pet name	70	66	4	94.3(86.2,97.8)%	58	12	82.9(72.4,89.9)%
initials	98	51	47	53.1(42.1,61.9)%	87	11	88.8(81,93.6)%

There were 9 microchip numbers and a single passport number recorded in the validation corpus, all were redacted by Clancularius. There were no telephone numbers, email addresses or postcodes within the validation set. These were assessed separately by searching the wider corpus for likely identifiers, this was not intended to provide a valid measure of efficacy for redaction of these identifiers. In regular use of Clancularius across the full SAVSNET corpus, where approximately 20,000 consultations were de-identified each week, there were minimal observed errors.

Sensitivity for pet names, all of which in this corpus were single first names, was 94.3 (86.2,97.8)%. One of the four missed pet names was an extensive misspelling, within that narrative the correctly spelled occurrences of the same name were redacted.

Identifiers were wholly redacted in all cases of recognition by Clancularius. Where a veterinary practice name was redacted the identifier, but not the fact it was a practice was redacted. For example, *Newman vets*, was output as <<*identifier*>> *vets.* This was desirable as it preserved the contextual information whilst redacting the identifying information.

5.4.3 Specificity within SAVSNET validation corpus

There were 39 software de-identification false positives in the validation set, that is words or phrases redacted that were not identifiers, these occurred in 37 narratives, with two narratives containing two different false positives each. This is a false positive rate of 1 in 1,863 words, or 1 incorrectly redacted word per 26 narratives. Overall precision, the positive predictive value of an redacted identifier actually being an identifier, was 94.9(93.1,96.3)%.

Of these false positives, 25 resulted from typographic or spelling errors inadvertently generating a word correctly recognised as a name by the software, for example 'ever' being mistyped as 'eve r'. Multiple word place names were responsible for five false positives, with a further two created by single word location false positives. Uncommon abbreviation accounted for two of these with *'lower st'* recognised as an abbreviation for 'Lower street' when it appeared to mean 'lower stifle', and 'ness' meaning 'necessary'.

Overzealous matching accounted for three false positives; with 'non-safe' name words occurring immediately after a word designated a safe name forming a recognised name pattern. Two of these instances involved the problematic word 'will'. There were 175 occurrences of this word in the validation set, 21 of them beginning with a capital 'W', none of them were being used as an identifier.

Three sets of apparent initials were wrongly redacted, two of these were generated by typographic errors (and were considered such), and the third *'ds'* was being used as an uncommon abbreviation for days. The optimal location of the corpus-specific redaction method in the de-identification sequence was prior to mask removal, where it occurred after removal of the generic mask, clinical abbreviations were exposed and an additional 25 incorrect redactions occurred.

The secondary safe name rule was responsible for two false positives. ‘Tim’ and ‘Tom’ were erroneously redacted where they were being used, with capitalisation, to mean Timothy hay and tomorrow respectively. There were 6 occurrences of the word ‘tom’ within the validation set, five of these were names, all of them written as ‘Tom’, one as a lone word and the others within name patterns. Likewise the other 4 uses of ‘Tim’ within the validation set were names.

5.4.4 Efficacy on application to different corpus

Within the sample of 100 consultations from the Bristol Cats Study corpus there were 97 identifiers in the essential group; all were redacted (sensitivity 100(96.2,100)%), in addition all 15 pet names were redacted. However one narrative contained both an email address and an HTML encoded URL link to the same, the latter had not been anticipated and could be deciphered to reveal the corporate email address, reducing the sensitivity for all essential identifiers to 99(94.4, 99.8)% or 99.1(95.2, 99.8)% if pet names were included.

The narratives of this small corpus contained a more diverse range of identifiers including 7 postcodes, 5 (plus the missed URL mentioned above) email addresses and 6 phone numbers; all were correctly redacted. Precision within this sample of the Bristol Cats corpus, using Clancularius in its native form, refined to the SAVSNET corpus, was 91.4 (86.4, 94.6)%.

5.4.5 Processing time

Clancualrius was able to process 100 words in 1.55, 1.61 and 2.26 seconds on systems 1, 2 and 3 respectively. Using machine 1 running 4 instantiations of Clancularius simultaneously, 100 words could be processed in 0.43 seconds (Table 5.4.b).

Table 5.4.b: Clancularius processing time, comparison in different operating environments.

Computer description	Operating system	Processor	RAM (Gb)	Build year	Processing time per 100 words (s)
Macintosh iMac	macOS Sierra	3.4GHz quad core	32	2013	1.55
Macintosh iMac with 4 simultaneous instantiations	macOS Sierra	3.4GHz quad core	32	2013	0.43
Mac mini	macOS Yosemite	2.6Ghz dual core	8	2014	1.61
Middle range laptop	64 bit Windows 8.1	1.6Ghz quad core	8	2014	2.26

5.5 Discussion

The efficacy of Clancularius in redacting identifiers within the SAVSNET Corpus compared well to other published systems (Table 5.5.a). It is inappropriate to draw direct comparison to other de-identification software as each system is designed to its own specification, Clancularius for example was not required to redact dates. This system is unique, as far as the author can ascertain, in the published data in that its target data is first opinion small animal veterinary clinical narratives.

In designing this system, a balance was achieved by the use of a research adaptable masking process. This proved a valuable feature in facilitating maximal sensitivity at the expense of occasional false positive redaction, provided those false positives were not of specific research importance. An alternative would have been to mask all potentially research important terms, in anticipation of future need. The latter approach would have risked impairing sensitivity and slowing processing time for little gain in data quality.

Key to the ability to use the masking process was the understanding that its purpose was to reduce research critical false positives but that this was not permitted to be at the expense of impaired sensitivity for genuine identifiers. This required careful mask design, and where necessary dictionary adaptation.

Table 5.5.a: Comparison of the efficacy of Clancularius to other published rule-based de-identification systems. Based on the sensitivity and precision for target data for each system.

De-identification system	Target data type	Sensitivity	Precision
Clancularius	Narrative data		
	SAVSNET corpus essential ids	99(97.6,99.6)	94.9(93.1,96.3)
	SAVSNET corpus inc pet ids	98.4(96.8,99.2)	
	Bristol corpus inc pet names	99.1(95.2,99.8)	91.4(86.4, 94.6)
HMS Scrubber (Beckwith et al. 2006)	Pathology reports	98.3	42.4
Medical De-identification System (MeDS) **(Friedlin and McDonald 2008)**	HL7 messages		
	Initial evaluation all types		~93
	Pathology reports HIPPA ids	99.5	
	Pathology reports All ids	99.1	
	Clinical narrative HIPPA ids	98.9	
	Clinical narrative All ids	95.7	
MIT's system (Neamatullah et al. 2008)	Nursing notes		
	Development corpus	96.7	74.9
	Test corpus	~94	not measured

The redaction of identifier patterns (full names and detailed location information) prior to lone identifiers ensured that where false positives occurred they were lone words which had been considered 'non-safe' during dictionary development, as a result the very occasional residual identifiers posed little risk to disclosing the identity of any party.

Initials were the least well redacted identifiers. As these had not been considered true identifiers the aim was to reduce the bulk of initials present, in order to reduce their likelihood of contributing to contextual identifying information. Approximately half of the initials were identified by Clancularius. If the vet initial redaction function was permitted to process the unmasked narrative this marginally improved sensitivity (61.2 (95%CI51.6, 70.9)%), at the expense of an increase in aberrant redactions which impaired the research quality of the narrative data.

One approach to redacting all clinician initials would be to redact all two and three letter words that are not used as accepted abbreviations. However, this would risk an over fitted system and with the extensive overloading and

idiosyncratic use of abbreviations would be likely to result in the redacting of clinically relevant words for minimal added identity security. Of the 46 sets of initials unique-within-a-narrative that were missed by Clancularius in the SAVSNET validation set 26 (18 unique within the corpus) were commonly used with clinical meaning within the SAVSNET corpus. Including one instance were a set of initials appeared to be being used as both a clinician's initials, which was clear because their full name was also present, and an anatomical abbreviation within the same narrative. If redaction of initials became desirable this could likely be improved with further work.

Clancularius does not contain wild card or spelling correction features within the identifiers themselves. The name dictionaries were generated from a range of sources and will include some misspelled names, the dictionary volume is such that where a misspelling creates another name it will be redacted.

As the majority of identifiers were clinician names this is less problematic as these were redacted based on contextual information in addition to dictionary matching. The greater issue was misspelled words forming names where the correct spelling would not have done. This accounted for 25 of the 39 false positive identifier redactions.

Clancularius was able to de-identify narratives at a rate of 3,870 words (approximately 70 clinical narratives) a minute, making it eminently suitable for integration into a system receiving 20,000 narratives a week in real-time or as a daily batch process. Processing time will vary dependent on the operating environment. Several instantiations of Clancularius can be run at once and in practice, where time was at a premium, a dataset was run as four simultaneous batches. This only marginally reduced the processing speed of individual instantiations and reduced the overall run time to 0.43 seconds per 100 words (Table 5.4.b). This observation may be exploited by adoption of a multithreading approach in future versions of Clancularius.

5.5.1 Further improvements

The 'secondary safe' system of recognising names where they had been considered non-safe but occurred with capitalisation was adapted to include

locations, this addressed the issue of the word bath being used to describe husbandry and management of skin problems and also the location Bath.

The email address rule was adapted to incorporate URL encoded email addresses in addition to their plain text notation.

In order to facilitate the application of the system to other datasets an optional white space cleaning method was incorporated. This had not been added to the original software as parsing and white space cleaning is likely to need tailoring to an individual dataset and its intended use. However, the option was added to increase the breadth of those able to use Clancularius within the research community.

5.5.2 Conclusion

Whilst the main aim was to be able to generate a de-identification system for the SAVSNET corpus it was important that it could be applied to other small animal veterinary datasets. It was anticipated that this would require some adaptation to the language and structure of each dataset. Validation on the Bristol Cats dataset was reassuring in this regard. The nature of Clancularius is such that the dictionaries, rules and type of data redacted by the simple pattern rules can be readily amended with minimal coding knowledge.

The use of this thorough and effective de-identification tool prior to human reading of narrative consultation records within the SAVSNET dataset further minimises the small risk of any party being identified by the researcher, even where there is prior knowledge of a case, whilst having minimal impact on the quality of the data for research purposes. As such, this adds a further tool to SAVSNET's armoury in responsively generating a large volume of research and surveillance ready data whilst maintaining high ethical standards.

Chapter Six Development of classifiers for the identification of clinical signs in the veterinary clinical narrative to support syndromic surveillance

6.1 Introduction

Syndromic surveillance has typically utilised information from accumulated diagnostic reports of notifiable diseases or clinician and laboratory coded data (Wu et al. 2008; Dórea and Vial 2016). Surveillance systems that rely on definitive diagnoses and laboratory investigation are hindered by their dependence on clinician suspicion of a diagnosis (Greene et al. 2012). Attempts to constrain free-text, for example through menu driven data entry or pre-defined mandatory terms, especially within the specialised and complex clinical narrative, risk the introduction of errors, omission and delay (Hall and Lemoine 1986; Penz, Wilcox, and Hurdle 2007), all of which impair the efficacy of surveillance systems that rely on clinically-coded data.

Increasingly, the wealth of information within narrative clinical records is being utilised in surveillance systems (Chapman et al. 2005; Travers et al. 2013). The efficacy of text-mining systems is dependent on the quality and breadth of free text available to the system, and the ability of classifiers to accurately extract information (Johnson and Friedman 1996).

An automated system able to classify consultation records for the presence of clinical signs, would facilitate timely detection of spatio-temporal trends in clinical signs (Conway, Dowling, and Chapman 2013), in turn facilitating observance of variation outside background levels, and thus potentially the emergence of illness independent of a diagnosis being made (Paterson and Durrheim 2013).

Where syndromic data can be automatically extracted the potential for real-time surveillance is more readily realised, because the need for time-consuming manual coding and reporting is avoided. Within companion animal health care a wealth of information is gathered during preventive health care visits (Shaw et al. 2008; N. J. Robinson et al. 2015), automated extraction from all routinely documented clinical narrative would also ensure capture of this information with no additional demands on the practitioner.

Historical challenges to the collation of animal health data are being overcome as individual animal health records become digitised, making them more readily available for research and surveillance. A survey of veterinary surgeons in the

UK found that only 6% did not use a computer system for client records (D. Robinson and Hooker 2006), other work has found that the majority of dogs in the UK visit a veterinary surgeon (Asher et al. 2011). It is likely that the majority of pet animals in the UK and similarly developed countries now have an electronic health record (EHR). This has led to a rapidly growing interest in filling surveillance gaps in these populations by utilising novel, technology driven, solutions based around the collection of large volumes of individual animal EHRs (VetCompass 2017; O'Neill 2013; University of Liverpool 2017).

With over a quarter of UK households hosting a dog or cat (Murray et al. 2010), a surveillance system able to detect and highlight increases in presentations for clinical signs influenced by environmental factors (factors extrinsic to the animal, such as pathogens or weather extremes) offers potential population health benefits both within the small animal population and the human population with whom they share many of those environmental factors.

Several syndromic surveillance systems in human medicine have utilised the chief or presenting complaint field of Emergency Department triage records (Tsui et al. 2003; Ansaldi et al. 2008; Aronsky et al. 2001; P. Brown et al. 2010). This unstructured but concise field in some respects lends itself well to classification with high specificity; it contains pertinent signs and is unlikely to contain negative findings. However, as with much of the clinical narrative, non-standard and over-loaded acronyms and abbreviations, misspellings, local and personal colloquialisms are commonly used. These factors pose challenges to the successful use of many natural language processing tools in developing classifiers of the clinical narrative (Chapman 2006). Notably, spelling correction algorithms are of limited value (Chapman et al. 2005) despite the high rate of spelling errors (Ruch, Baud, and Geissbühler 2003).

Text mining methods have been used to classify small animal veterinary consultation narratives with high sensitivity but these have been prone to poor sensitivity (Anholt et al. 2014). The SAVSNET dataset has already been analysed using simple word-searches combined with manual reading of records, to overcome the poor specificity of such word searches, providing insights in to parasite risk and antimicrobial use (Radford et al. 2011; Tulloch et al. 2017). Here we expand this earlier work to describe the development of rule-based

free-text classifiers able to detect documented clinical signs with high sensitivity and specificity within routinely recorded veterinary clinical narratives.

6.2 Methods

6.2.1 Software

All code was written in Python version 3.6.0 (Python Software Foundation 2016) with text mining methodology developed using the integral re (regular expression) module version 2.2.1. Data handling functionality was provided by Pandas version 0.20.3 (PyData Development Team 2017) and Numpy version 1.13.3 (Numpy developers 2017) and the bespoke regular expression concordance tool, `regexConcordance()` described in Chapter four, was used extensively to explore the language used in describing clinical signs and their context.

6.2.2 Data

The narrative records of consultations collated in near real-time by SAVSNET were the intended target of the developed classifiers, the previously collated SAVSNET dataset was utilised in classifier development. At the time of writing, the dataset contained 3 million narrative consultation records, these had been contributed by 404 veterinary clinics across the UK between November 2013 and November 2017 and approximately 20,000 new narrative records were being collated each week.

6.2.3 Pre-processing

Text was pre-processed to generate a dataset suitable for human reading, with redaction of identifiers using Clancularius as described in Chapter five. The additional needs of software processing were also met during this stage with white space normalisation and proxy sentence creation: single or multiple line breaks were replaced with a single full stop, all whitespace was converted to spaces and multiple contiguous spaces replaced with a single space.

The narrative field was split to separate template-based text representing prescription and dispensing labels, which included warnings of potential adverse drug effects, from the true free-text clinical record. (Figure 2.1.c).

6.2.4 Clinical signs

A group of clinical signs and parameters were selected for analysis. These were chosen on the basis that they were likely to be documented if noted during the history or examination of an animal seen within first-opinion practice, and their presence was likely to be associated with, or in the case of parameters affected by, infectious disease or the effects of other environmental factors (Table 6.2.a).

Signs were considered to have been documented as present where: i. they were recorded as being present at the time of the consultation; ii. they were not subsequently discounted by examination or more in-depth history taking; iii. documentation related to the presenting animal.

6.2.5 Training set

To ensure that syntax, language and text structures from different practices were evenly represented, a sample of 100 consultation records were selected at random (without replacement), using the Pandas `sample()` function, from each clinic to form a new dataset. This dataset was sampled again to produce a subset of 10,000 consultation records regarding dogs and cats (subsequently referred to as the “training dataset”). The consultations within the training dataset were manually coded, by the author, for the presence of clinical signs or parameters documented within the consultation narrative.

Table 6.2.a: Clinical signs and parameters for which classifiers were developed and validated. Rationale for inclusion of these signs in a classification system intended to identify alteration in presentation rates attributable to environmental factors.

Clinical sign	Rationale for use of this group of clinical signs
Upper aero-digestive	
Inflamed conjunctiva Nasal discharge Oropharyngeal inflammation or ulceration Sneeze	Markers of upper respiratory tract infection or inflammation. Described in canine distemper, feline herpes, Calici virus and *Chlamydophila* infections. Oral inflammation also seen with toxin and physical irritant exposure
Lower respiratory	
Cough Crackles Increased respiratory effort Respiratory rate Wheeze	Indicators of upper (cough) and lower respiratory pathology. Described with *Bordetella* infection, distemper, parainfluenza and adenovirus. Wheeze included as a marker of airway irritation.
Gastrointestinal	
Abdominal pain	Inflammation or distension of the abdominal viscera or peritoneum. May be present in hepatopathies prior to clinically evident icterus, but also in acute enteropathies, pancreatic disease and acute nephropathies (leptospirosis)
Diarrhoea	Indicator of impaired intestinal fluid and electrolyte transport. Feature of canine parvo virus, distemper, infectious canine hepatitis and feline enteric coronavirus.
Haematochezia	Commonly ascribed to canine parvovirus and *Clostridium spp* (also acute haemorrhagic diarrhea syndrome). May also occur with anticoagulant toxin ingestion.
Hypersalivation	Multiple associations including oral lesions secondary to infective or toxic insult, nausea, and following exposure to a range of toxins.
Vomit	Indicator of upper GI irritation, neurological and systemic disturbances. Described in toxin ingestion, including ethylene glycol, canine parvo virus, feline panleucopaenia, canine distemper and infectious hepatitis.
Neurological	
Ataxia Nystagmus Seizure	Although primarily of non-environmental aetiology; these signs may indicate ingestion of neuro or hepatotoxins including bromethalin rodenticides, toxic mushrooms, heavy metals and ethylene glycol.
Systemic	
Body temperature Lymphadenopathy Heart rate Jaundice	Infection and other insults generating an inflammatory response. Increase in rate secondary to pyrexia and toxic haemodynamic compromise. Seen in canine infectious hepatitis and *Leptospira spp*. infections, also in hepatotoxin ingestion.

6.2.6 Rule development

6.2.6.1 Individual clinical sign recognition

An initial exploratory examination of the phraseology used to describe each sign of interest, within the coded training dataset, was performed to allow development of a preliminary regular expression-based classifier for each sign. These search phrases combined a sequence of symbols, characters and character ranges to delineate string patterns likely to match the clinical sign of interest. For example, the initial regular expression to identify consultations where cough was documented consisted of the word *cough* and the abbreviations *c+* and *co+.* Refinement, independent of context, generated a more complex search phrase able to identify the wide range of misspelling and phraseology used to denote the clinical sign. Incremental adjustments were made as part of an iterative process until sensitivity could not be further increased (Figure 6.2.a).

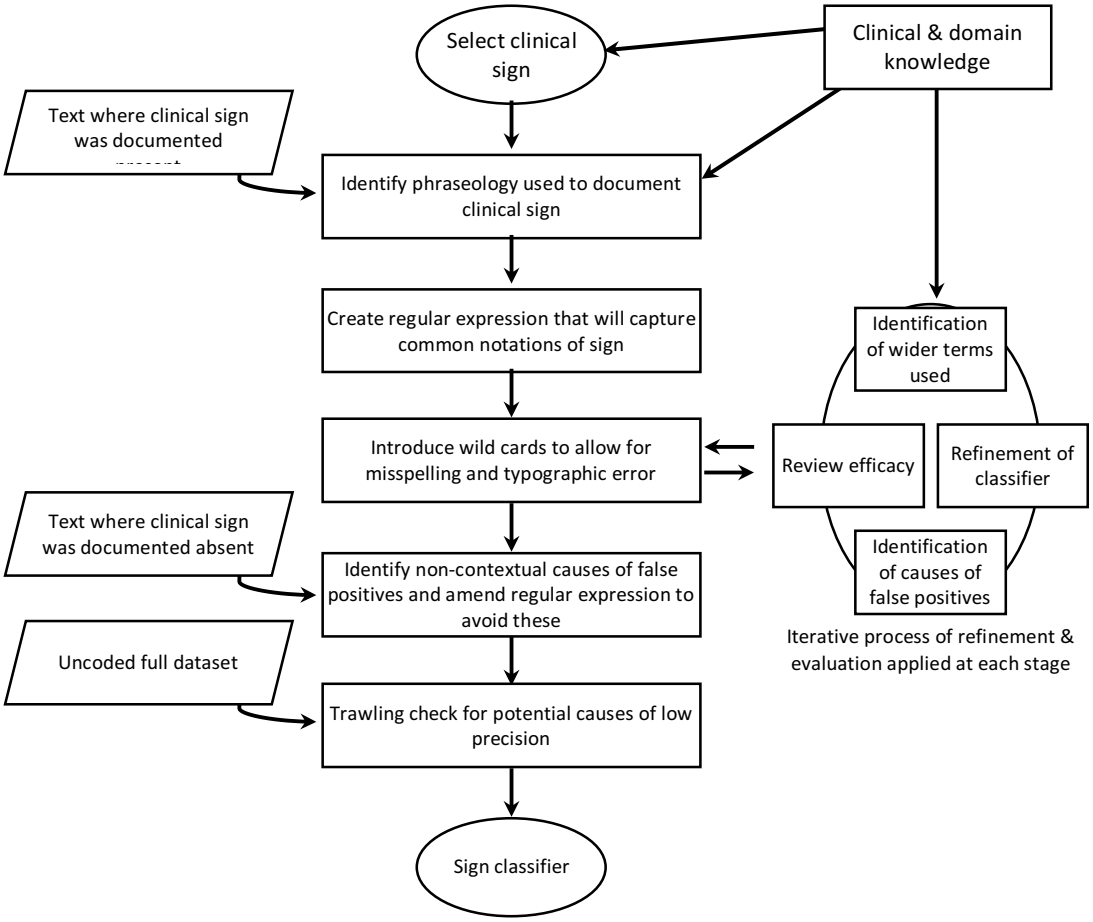

Figure 6.2.a: Process of individual clinical sign classifier development

6.2.7 Identification of contextual features

The `regexConcordance()` method described in Chapter three was used to identify lexical features common to the context of many signs. The individual sign regular expressions were used as search phrases and adjacent phraseology explored. Common characteristics were identified where the clinical record: indicated that a sign was present, for example:

*“been off colour for several days, **episodes of** vomiting after eating”,*

*"v **again**",* and

***"has a** cough";*

excluded the presence of a sign, for example:

*“lost weight but **no** diarrhoea or vomiting”,*

*"cough **no"** and*

***"not** pyrexial";*

or were used where the veterinarian was cautioning to observe for the occurrence of a sign that was not occurring at the time of presentation, for example:

*“**monitor for** vomiting and*

***"if** coughs again" but not*

***"review if** diarrhoea **gets worse** " as the sign is already present.*

Phraseologies or contextual characteristics with similar function within the clinical sublanguage were grouped together as functional synonym entities for reuse throughout the classifiers, this was convenient for notation and permitted ready adaption or addition of patterns. In some instances, these were not strictly synonyms but words used in a similar manner or found in similar contexts.

Functional entities, for example a group of words that meant the sign was being mentioned in the negative would have included *no*, *not*, *none and nad*, were in

turn used to form a series of generic rules encoded within regular expression filters, these filters either permitted or prohibited matching of the clinical sign-specific regular expression, dependent on the context they were designed to match.

An iterative process was used to refine the regular expressions of the generic contextual rules until they facilitated optimal identification of whether each clinical sign was present within the training set (Figure 6.2.b). To pre-empt the impact of the increased variation in phraseology occurring as the number of practices contributing data grows, synonyms and hypernyms (dog is a hypernym of greyhound, all greyhounds being dogs) for each word included in the classifier were incrementally added, if inclusion of a previously unseen term had no bearing on specificity it was were retained. Where there were specific additional phraseologies associated with an individual clinical sign, these were added to a dictionary of sign-specific permissive and prohibitory patterns, developed in an analogous manner to the generic context patterns.

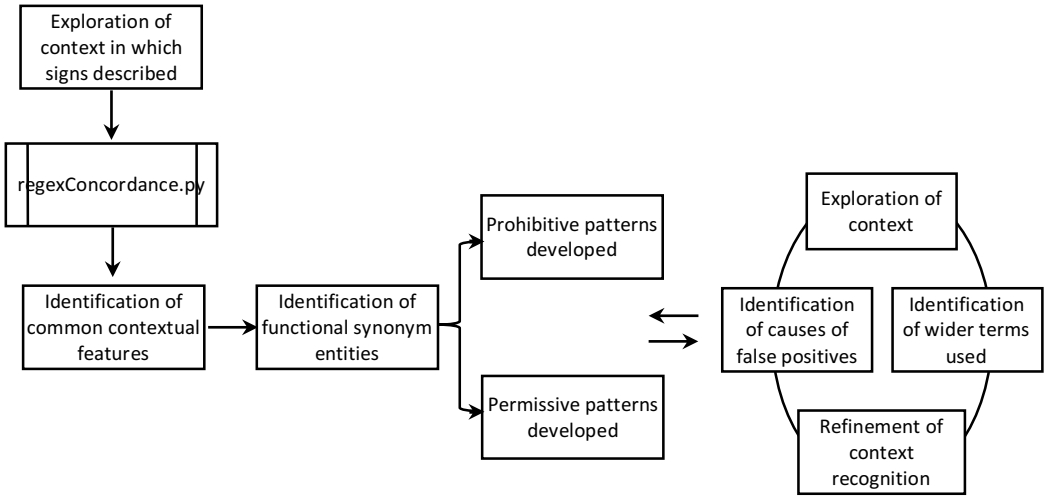

Figure 6.2.b: Incremental process of development and refinement of the contextual recognition framework

6.2.8 Physiological parameters

The training dataset was manually annotated for the presence of numeric physiological parameters including body temperature, heart rate and respiratory rate. A regular expression was used to extract a sample of strings containing a

number between two and three digits in length, with or without a single decimal place, from the broader, uncoded, SAVSNET dataset. These were examined using the `regexConcordance()` method and the characteristics of strings containing each parameter identified.

Regular expression search phrases were then developed to identify numbers representing each of the desired parameters and these were combined and adjusted to form a series of regular expression-based rules that facilitated optimal parameter extraction in the training set. The rules were applied in a specified order to preferentially extract clear statements in preference to less definite parameter notation.

In a small number of narratives, an individual regular expression extracted multiple values. Where this occurred, two options were available, either to ignore both values; or to find the mean of values. If the latter were chosen a lower limit could be set, for example for heart rate a lower limit of 50 was set, to minimise the likelihood of finding the mean of a heart rate and a miss-identified parameter. The manner in which duplicate values are processed can readily be adapted to suit the application of the tool.

Fail-safes were built in to the system, to ensure that any number altered by the extraction tool was manually screened and could be reverted to its original value if necessary. These were applied in circumstances where a parameter extraction had occurred and there was a likelihood that this was a false positive that had not been captured within the hierarchical framework and as a result had been over-ridden, or the extracted value had been calculated as, for example, where more than one value was identified as above.

6.2.9 Inter-person classification validation

Ideally multiple clinicians would have been used to create an external clinical expert-coded validation dataset. Logistics and resources prohibited this and mandated that the author undertook validation. In an effort to provide a measure of the level of agreement between the author and practising clinicians, a sample of consultation narratives was coded manually by four clinicians practising small-animal medicine in a large UK veterinary teaching hospital.

Discrepancies in manual classifications between the author and clinicians were identified within a Pandas dataframe using the `np.where()` method to mark free-text records where there was mismatched coding. This process was undertaken maintaining the author blind to the actual coding of either. The consultation narratives where there was disagreement in manual coding were classified a second time by both of the original manual coders and then re-compared. The *K*appa statistic, a measure of inter-rater reliability (J. Cohen 1960), was calculated for both the initial and recoded sample between each clinician and the author. A *K*appa value in excess of 0.9 is considered to represent near perfect agreement (McHugh 2012) and would be considered to represent an acceptable level of agreement in these circumstances.

6.2.10 Assessment of classifier efficacy

Once the classifiers had been optimised within the training dataset, a second manually-coded dataset was created. This test dataset consisted of 5000 consultation narratives drawn at random from consultations collated within the SAVSNET dataset over the subsequent month, this ensured that the narratives were unseen during classifier development and represented the variety of consultations being collated in real time.

Each consultation was read and coded on two occasions randomly distributed through the coding period to ensure consistency in approach. The same considerations as described previously were used to decide whether a clinical sign was being documented as present in a given consultation.

Manual and software classification was compared within a Pandas dataframe, and recall (sensitivity), specificity, precision (PVP) and the F_1 measure calculated for each clinical sign. The F_1 measure is calculated as the harmonic mean of the precision (analogous to positive predictive value) and the recall (analogous to sensitivity) where precision and recall are equally weighted. It provides a measure of the test's accuracy and is preferred by some data scientists to the other measures of efficacy. The upper and lower bounds of confidence for sensitivity and specificity were calculated using the Wilson Score interval (Wilson 1927) as this permits calculation when the proportion is very

close to 1, and is appropriate for both large and small sample sizes (L. D. Brown, Cai, and DasGupta 2001).

In order to evaluate the potential to use this method to detect syndromes likely to reflect emergence of new disease, a gastrointestinal signs superset was defined which tagged narratives having any or all of the following signs: vomiting, diarrhoea, haematochezia or abdominal pain.

To test potential utility for real-time surveillance the classifier system was deployed on an Apple Mac computer, parsing and processing the data collected nationally by SAVSNET on a daily basis. The resultant dataset was provided as a data feed for a web-based dashboard, available within the University of Liverpool intranet, that demonstrated real-time daily variation and was used routinely by the SAVSNET team.

6.3 Classification framework

The nature of the language used in documenting clinical narratives gave rise to a series of constituent patterns that were used within the individual clinical sign regular expressions and the context-sensitive framework. These patterns took account of the clinical notation style and use of words in describing clinical signs. Notably non-word characters carried clinical meaning in some instances and formed components of words. As a result, a word was denoted as a sequence of any characters except those commonly used as punctuation or carrying meaning in relation to sentence structure, this was encoded within regular expressions (Figure 6.3.a). The endOfWord regular expression used a positive look ahead assertion to ensure that pattern matching moved to the next non-word character.

```
word = '''(?:[^\.\s,:!?]+|\d*[\.]\d+)'''

endOfWord = '''[^\.\s,:!?]*(?=(?:\W|$))'''
```

Figure 6.3.a: Fundamental regular expressions within the classification system. Patterns denoting the characters considered word characters within the clinical context.

6.3.1 Clinical entities & individual clinical signs

Key examples of these functional synonym entities within the individual clinical sign classifiers were the constituents of Celsus' cardinal signs of inflammation (Celsus 25AD): pain (dolor), heat (calor), redness (rubor) and swelling (tumor) with Virchow's addition of loss of function (functio laesa) (Virchow 1858). Each was incorporated as an entity and these applied as a group denoting any feature of inflammation, or individually, as appropriate for classifier purpose.

The patterns used to recognise abnormality and discharge exemplified the formation of functional synonym entities, in that they incorporated a range of words with strictly diverse, but related, meaning used in the description of a clinical event or observation (Figure 6.3.b). Although the functional clinical entities were used within multiple classifiers, by virtue of the pathology that they encoded, language use was such that it was not always effective to incorporate the generic entity despite it meeting the semantic needs of a given classifier.

A prime example of this is the lymphadenopathy classifier, designed to identify documentation of a disease process or abnormality of the lymph nodes, typically enlargement as a result of an immune response to infection. In effect, this classifier looked for notation indicating abnormality in close proximity to notation related to lymph nodes. However, the abnormality entity (Figure 6.3.b) did not encode the words utilised in this context to mean abnormality, which were more akin to the inflammation entity, instead a classifier specific 'opathyWords' entity was used (Figure 6.3.c).

```
abnormalWords= '''(?<![a-z])(?:
                                def[eic]{2,}|
                                delayed|
                                abn|
                                ab+[er]|
                                abse|          #this is a group of
                                anom|          #expressions designed
                                at[iy]p|       #to identify terms for
                                biz+a|         #abnormal with a
                                curi|          #degree of misspelling
                                def+i|         #permitted
                                devi|
                                ec+ent|
                                fun+y|
                                ir+eg|
                                mis+|
                                neg|
                                odd|
                                pec+ul|
                                poor|
                                prob|
                                que+r|
                                strange|
                                unus|
                                un+at|
                                w[ei]+rd
                            )'''+endOfWord

dischargeWords = '''(?<![a-z])(?:
                                obst|
                                blo[ck]|
                                drip|
                                drain|         #these terms are used
                                pour|          #in the veterinary
                                sero|          #clinical sublanguage
                                pur[ul]|       #describedischarging
                                dis(?:ch|hc)|  #lesions or orifices
                                gree|
                                muc+op|
                                muco?us(?!\s?mem)|
                                fl[ui]+d|
                                liq[ui]+d|
                                damp
                            )'''+endOfWord
```

Figure 6.3.b: Examples of clinical entity groups, encoding abnormality and discharge.

```
lnWords = '''(?<![a-z])
                (?:
                    ln+(?![a-z])|
                    ln+(?:[\']|&apos;)?s(?![a-z])|
                    su?b?mln|
                    ln+(?:\W|&apos;)s(?![a-z])|
                    nodes?|
                    ly[mph]{2,}|    #a group of generic terms for
                    subm[ae]n|      #lymph nodes and anatomical
                    tonsi|          #locations of nodes whose
                    prescap|        #condition is commonly documented
                    pop(?=[l\s])|
                    trochl
                )'''+endOfWord

opathyWords = '''(?<![a-z])         #words describing abnormality
                    (?:
                        up|
                        big|
                        larg|
                        en[lar]{1,}g|
                        enl|
                        incre|
                        mass(?:\sin)|
                        rais|
                        huge|
                        hard|
                        hot|
                        and\sin|
                        rais|
                        as\s(?:bef|prev|last)
                    )'''+endOfWord

lymphadenopathy = '''(?:(?:         #structured combination of
    ly[mph]{2,4}ad[a-z]*|           #several entities to form a
    '''+opathyWords+'''             #clinical sign classifier
            [^\.\w,]+(?:
                (?!'''+negativePrior+''')'''+word+'''[^\.\w,]+){0,4}
                    '''+lnWords+'''|
    '''+lnWords+'''[^\.\w,]+
        (?:(?!'''+nadWords+''')
            (?!'''+negativePrior+''')
            (?:'''+word+'''|sl[\.])[^\.\w,]+){0,4}
                '''+opathyWords+'''|
    '''+lnWords+'''[^\.\w,]*(?:
                (?!'''+nadWords+''')
                (?!'''+negativePrior+''')
                (?:'''+word+'''|sl[\.])[^\.\w,]+){,2}
                    (?:[+\^]+|swol|infl|enl)|
    (?:sw[oe]l|infl)'''+endOfWord+'''[^\.\w,]*'''+lnWords+'''
    )(?:[^\.\s]*\son\s(?:l|r))?)'''
```

Figure 6.3.c: The lymphadenopathy classifier

Clinical signs fell broadly into two groups, those that were identified using a single verb, such as 'vomit', and those denoted using a verb and an anatomical location or bodily function, for example 'inflamed conjunctiva'. Colloquialisms, descriptors of clinical signs, and abbreviated forms added complexity to this division. Diarrhoea for example as a lone verb described a clinical sign, however the same sign was also described using phrases such as 'loose stool', 'sloppy poo', 'cow pat' and 'mr whippy'.

Where multiword phrases were used to describe clinical signs, classifiers tended to be composed of two or more clinical entities; these were recognised as the sign descriptor if they were found within the same sentence or other predefined lexical distance. The acronyms used to encode 'no abnormality detected' and its multiple counterparts (Table 4.4.m) were combined with a limited array of words signifying normality to form the `nadWords` entity, this was used extensively within the patterns of verb and location type individual sign classifiers. Incorporation of the `nadWords` entity permitted greater lexical distance between the clinical entities of the sign classifier and so guarded against false positives whilst increasing sensitivity (Figure 6.3.d).

```
nad3Words = '''(?<![a-z])
                    (?:
                        aok|          #series of acronyms for all ok
                        nad|          #no abnormality detected
                        naf|          #no abnormality found etc.
                        nas|
                        nsa|
                        nsf|
                        wnl
                    )'''
nad4Words = '''(?<![a-z])
                    (?:
                        fine(?!\sin)|         #part statements of
                        good|                 #normality
                        norm(?!(?:ote|al+y))|
                        comf|
                        noth|
                        heal(?![is])
                    )'''
nadWords = '''(?<![a-z])                    #combining above
                    (?:
                        n(?:\W|$)|
                        ok|
                        '''+nad3Words+'''(?!\w)|
                        (?:'''+nad4Words+endOfWord+''')
```

Figure 6.3.d: Example of a clinical entity. `nadWords` is a group of acronyms used to indicate that no abnormality was detected.

Although a degree of contextual sensitivity was by necessity incorporated into individual sign classifiers where the sign could be denoted using multiword descriptors, for the most part clinical sign recognition was independent of the context in which the sign was documented. Some exceptions to this context-naivety were however incorporated into the individual sign classifiers, where this assisted in distinguishing between two different uses of a homonym or two different word roots generating the same spelling error.

The cough pattern (Figure 6.3.e) provides an example where a degree of context sensitivity within the sign classifier was necessary. This context sensitivity distinguished between *kennel cough* being used as a description of the presentation of tracheobronchitis and as an abbreviation for the animal having presented for or been administered the *kennel cough vaccination*.

```
cough = '''(?:(?<!need\s)(?<!\Wor\s)(?<![a-z\\/])(?:
    (?:sn?[+]/W*)?co?[+]+(?![rs]|\ss\W)|
    (?<!kennel\s)
        \S*(?:coughing|
        cou[gh]{2,}|
        c[ou]gh[eis])
            [a-z,:]*
                (?!\svac)(?!\safter(?:w|\s[lv]))(?!or\ssn)|
    (?<=^)["?]?kennel\scough(?=[\.])|
    cou[gh]+(?:ed|i[ng]{2})|
    c[a-z]ughi[ng]+[a-z,:]*|
    honk[a-z,:]*|
    2?"?thr\w+\s?clea[a-z,:]*2?"?|2?"?clea\w*\sthro[a-z,:]*2?"?|
    garg+l[a-z,:]*|
    huf+i[ng][a-z,:]*|
    hacki[ng]+[a-z,:]*|
    (?<!owner\s)(?<!o\s)chok[ie][a-z,:]*|
    (?:kc|kennel\scough)\s(?:
            like|
            signs|
            type|
            symptoms|
            again)
    )(?!t\s(?:he|she|his|her|th[ie]r))
    '''+endOfWord+'''
        (?:\s(?:on|a|when))?
        (?:/retching)?
        (?!\sor\ssn)
        (?<!
            (?:hen|
            and|
            \Whe|
            she)
                \scought)
                )'''
```

Figure 6.3.e: The regular expression developed as a classifier for the clinical sign cough.

This cough classifier also provides an example of another instance where a degree of contextual sensitivity was needed within the sign classifier itself. The design of the sign classification regular expressions was such that word endings were captured outwith the sign pattern and within the context sensitive framework. Thus, letters immediately following a word that matched an individual sign classifier were ignored unless otherwise designated by a rule. This was usually the desired effect, however occasionally posed challenges, for example the misspelling of *coughed, cought,* is also seen as a misspelling of *caught*, with overlap between the adjacent phraseology in each instance. Where the word stem *cough* matched its sign classifier the trailing t would by default be captured within the `endOfWord` pattern. Steps were therefore introduced to attempt to differentiate between the two uses of this misspelling.

6.3.2 Contextual entities

Key contextual features were encoded within functional entities. Notably temporal indicators were used to distinguish documentation of previous issues from the current clinical presentation (Figure 6.3.f). The historical entity gained importance because it was common to find record of events occurring at the time of adoption of an animal and at the time of, although not necessarily causally linked to, treatment for worms or with non-steroidal anti-inflammatory drugs (NSAIDS); these acted as markers that an adjacent clinical sign was not being discussed in the present tense. Patterns recognising that a clinical event had resolved also assisted in discerning temporal relationships, these were key within the contextual framework.

```
freqWords = '''(?<![a-z])
                    (?:
                        d[ai]l+y|
                        once|tw[iu]ce|
                        one|t[ow]|thre
                        multip|many|
                        eve?ry|
                        ag[ai]+n|
                        reg|freq
                    )'''

timingWords = '''(?<![a-z])
                    (?:
                        day|
                        week|
                        month|
                        aft|
                        befor|prior|
                        today|toni?[gh]+t|
                        yest|y\W?day|yday|
                        overn|noct|
                        then|this|the|
                        beg[au]n|start|
                        last|
                        since|
                        for
                    )'''

historical = '''(?<![a-z])
                    (?:(?:
                        imp\w*ed|
                        been\sfine|
                        as\sfar|
                        did\shave|
                        after|
                        after\s(?:worm|ns[ai]+d)|
                        with\s(?:worm|ns[ai]d)|
                        when\sfirst|
                        first\sar+ived|
                        since\swas|
                        have\smade|
                        stop+ed|
                        rec\w*ck(?![:])|
                        see\W*note|
                        notes?\W*from|
                        post|
                        ago
                    )(?![,!?]))'''
```

Figure 6.3.f: Examples of temporal entities incorporated as indicators to distinguish historical from current clinical problems.

Recognition of negation was key to deciphering context; regular expressions were built to recognise phraseology adjacent to clinical signs which indicated the absence or resolution of the sign, rather than its presence. The nature of regular expressions is such that it was necessary in some instances to create near duplicate entities to accommodate the fixed width mandated for look-behind assertions in Python and many other programming languages, and to minimise the use of look-around assertions adjacent to whitespace and other extremely common characters as this would have had considerable processing speed implications.

To this end entities specifically recognising negation prior to and after clinical sign notation were formed; the `negativePrior`, `oneWordNegBehind` and `negativeAft` entities respectively. A group of temporal and negation entities were combined within negative look-ahead assertions to form the `negationContext` entity (Figure 6.3.g), a key component of the contextual framework.

```
negationContext =   '''
                        (?!'''+historical+''')
                        (?!'''+prospective+''')
                        (?!'''+negativePrior+''')
                        (?!'''+ifPhrases+''')
                        (?!'''+discursive+''')
                        (?!'''+exampleCaution+''')
                        (?!'''+examDifference+''')
                    '''
```

Figure 6.3.g: Example of a cluster of context conferring entities.

6.3.3 Contextual framework

The context nest was constructed from a series of regular expressions and contextual entities aimed to identify linguistic patterns that had been recognised as signifying the context of an adjacent clinical sign (Figure 6.3.h). Whilst the individual clinical sign classifiers were intended to be highly sensitive for documentation of the sign itself, the context nest conferred specificity for documentation that a clinical sign was noted as clinically present during the history or examination of the animal.

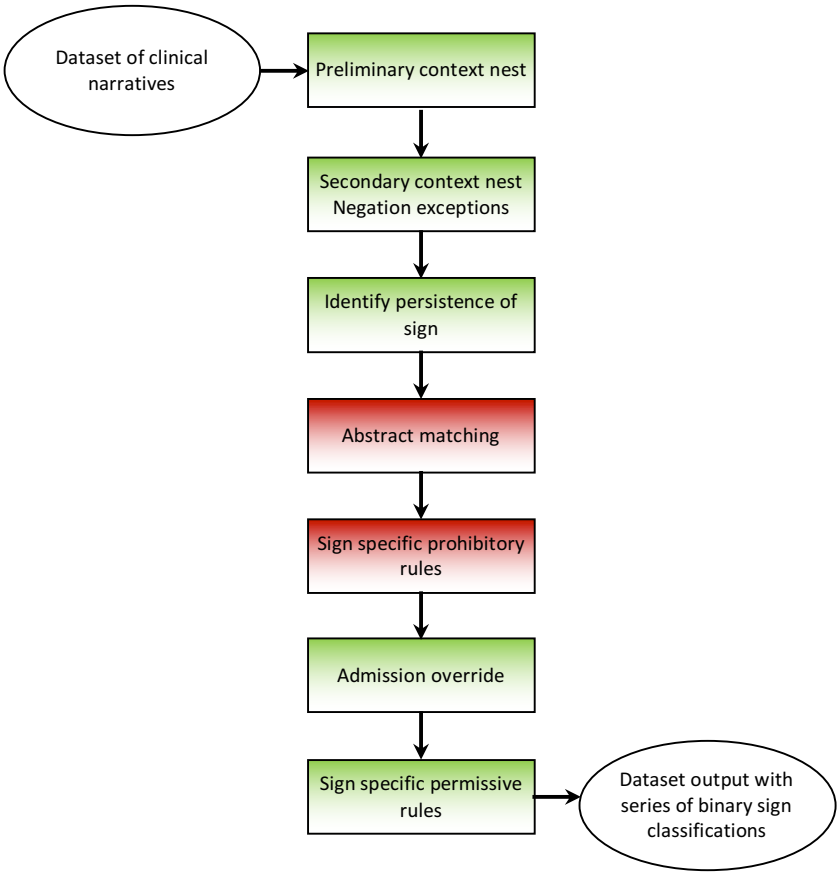

Figure 6.3.h: Flow diagram of the context sensitive classification system. Components with green shading are primarily permissive patterns, those with red shading prohibitory.

Overall, each classifier comprised a clinical sign regular expression filtered hierarchically by a series of filter regular expressions and subsequent application of clinical-sign specific ancillary rules. Classifiers were applied to datasets held in Pandas dataframes with outputs stored as structured text files (csv) for use by downstream consumers.

6.3.3.1 Preliminary context nest, clinical sign matching provided not negated

Lists of negative findings (e.g. “no vomiting, diarrhoea, coughing or sneezing”, "no vdcs") were a common feature and posed a challenge in optimizing specificity. One of the aims of the generic rules was to ignore regular expression matches where they formed a part of these lists. For signs not commonly included in lists of negatives, additional overriding rules were added to ignore matches where a clear statement of the absence of a sign was documented.

Where signs were commonly included in lists of negatives the overriding rule was applied with greater caution as occasionally reflex documentation of negatives occurred, where a sign was also documented as present.

The preliminary context component was the main stage of recognition that a clinical sign was being described as present (Figure 6.3.i). In essence, this rule framework permitted matching of the individual sign classifier if the beginning of the string that it matched was not preceded by a negation phrase, for example "not coughing" would not match whilst "has been coughing" would match. Negation affected the clinical sign match where it was within the same sentence and occurred within a specified number of words prior to the clinical sign match, in most cases this was four words, for some signs five words was required. Addition context sensitivity within this set of rules precluded matching where the clinical sign was noted to have resolved.

6.3.3.2 Secondary context nest, exceptions to negation

By prohibiting matching of a clinical sign pattern adjacent to negating phraseology, phrases such as "vomiting not improved" and "no better still coughing" were potentially prevented from matching. This was in part addressed within the latter part of the preliminary context nest and further by the use of a second permissive pattern that actively screened for phrases suggesting that a sign was present that may have been associated with words found within the negation entity (Figure 6.3.j).

```
regex = re.compile('''
    (?:
        (?:(?:(?:^|\W)
                        (?:
        '''+negationContext+histAdd+'''
        '''+word+ntEnds+'''
                            (?<!no)
                            (?<!not|no\W)
                            (?<!not\W)
                                [^\.\w]+'''+negationContext+''')
                                    {'''+str(num)+'''})|
        (?:(?<!e\.g)
            (?<!eg)
                (?:[\.!?"]+|^)
                        (?:'''+word+ntEnds+'''[^\.\w]+
                            )*)|
        (?:'''+permisAdd+permissiveContext+'''[^\.\w,]+)
        )(?<![a-z])(?<![a-z][+])
            '''+signRegex+''')
                '''+endOfWord+vdAdd+'''(?![?])
                                        (?![^\.\w]+not\snow)\s?
                                        (?!\s'''+historical+''')
                                        (?!'''+historical+''')
    (?:[\.",]|
    $|
    conta|
    \s?\w+\s?(?:times|x)|
    (?:[>]|&gt;)|
    (?:(?!(?<![,;])(?<![,;]\s)not?\W+(?!'''+notResolv2+'''))
        '''+resolvedAdd+notResolved+vdAdd+'''
            (?:[\.",]|
                    $|
            (?:and|&(?!\w)|&amp;|othe)|
            (?:[>]|&gt;)|
                (?:'''+resolvedAdd+notResolved+vdAdd+'''
                    (?!\s'''+historical+''')
                    (?!'''+historical+''')
                    (?:[\.",]|
                    $|
                    (?:and|&(?!\w)|&amp;|othe)|
                    (?:[>]|&gt;)|
                        '''+resolvedAdd+notResolved+vdAdd+''')))))|
    (?:(?![^\.\w,]+'''+negativeAfter+''')
        (?:'''+resolvedAdd+notResolved+''')*
            (?:[\.!?",]|$))
    )
    ''', re.I|re.VERBOSE)
```

Figure 6.3.i: Preliminary context nest designed to only permit matching of a clinical sign classifier where a sign is not being described in the negative or as having resolved.

```
regex = re.compile('''
        (?<!e\.g\.\s)
        (?<!e\.g\s)
        (?<!eg\.\s)
        '''+ntEnds+'''[^\.\w]'''+oneWordNegBehind+'''(?:
        (?:(?<![a-z])(?:
            '''+exceptionsToNegation+'''|
            '''+severity+'''|
            '''+persistence+'''|
            '''+otherIll''')
            '''+endOfWord+'''[^\.\w]+
                (?:(?!'''+negativePrior+''')[^\.!?,]){,20}?
                        '''+histAdd+'''
                        [^\.\w]*'''+signRegex+endOfWord+vdAdd+'''
            (?![?])
            (?:[\.",](?!\s?not\snow)|
            $|
            (?:(?![^\.\w,]+'''+negativeAfter+''')
                '''+resolvedAdd+notResolved+vdAdd+'''
                    (?:[\.",]|
                        $|
                    (?:'''+resolvedAdd+notResolved+vdAdd+'''
                        (?:[\.",]|
                            $|
                        '''+resolvedAdd+notResolved+vdAdd+''')))))|
            (?:(?![^\.\w,]+'''+negativeAfter+''')
                (?:'''+resolvedAdd+notResolved+vdAdd+''')*
                    (?:[\.!?",]|$))
            ))|(?:
            (?:^|\W)
            '''+negationContext+endOfWord+'''[^\.\w]+
            '''+signRegex+endOfWord+'''
                (?![?])[^\.\w]+(?:(?:
                        '''+postSignExceptions+''''')|
                        (?:'''+word+'''\s(?<!no\s)(?<!not\s)){0,2}
                            (?:nearly|
                            almost|
                            begin+ing\sto)))))|
        '''+oneWordNegBehind+'''
            (?:had[^\.\w,]+'''+signRegex+'''[^\.\w,]+(?:also|
                                                    since|
                                                    for|
                                                    past|
                                                    at\shome|
                                                    \d)
            (?!\s\w+\s(?:ago|post|after)))''',
    re.I|re.VERBOSE)
```

Figure 6.3.j: The secondary context nest to capture negation exceptions.

6.3.3.3 Identify persistence context

The third component was permissive and relied on the simple premise that a clinical sign can only persist if it is present. The identify persistence pattern thus identified phraseology suggestive that a sign was still present (Figure 6.3.k).

```
        regex = re.compile(
                oneWordNegBehind+
                exampleWordsNegBehind+
                observeWordsNegBehind+
                signRegex+'''(?:
                    '''+word+'''\s
                        (?:
                    contin|
                    persi[st]|
                    still|
                    int+erm+it|
                    (?:not|'''+endOfWord+'''(?:\Wt))
                        \s(?:
                            st\w*p|
                            res\w*lv|res\w*vl|
                            set+l))))|
                (?:
                    no\sw*\s(?:
                                cha?nge?|
                                impr
                            )'''+endOfWord+''')\W?\s?'''
                        + signRegex,
                re.I|re.VERBOSE)
```

Figure 6.3.k: The contextual framework component providing recognition that a clinical sign continued to be present.

6.3.3.4 Identification of discursive documentation

A primarily prohibitive component of the context sensitive framework identified where a clinician had documented their cautionary advice to an owner. Without this component these warnings would likely be detected as clinical signs as they usually matched the preliminary context nest.

Where this pattern matched for a given sign previous matches of the permissive components were overridden and the sign encoded as absent. This required incorporation of facets to recognise where the clinician's warnings were of deterioration or lack of improvement rather than development or recurrence of a sign (Figure 6.3.l).

```
        regex = re.compile('''
            (?<![a-z])(?:
            '''+warnings+''')
            (?:[^\.,]
                (?!'''+persist+'''|
                '''+deteriorate+'''|
                '''+resolvedNoContext+''')
                )*
                (?<!\sin\s)

                '''+signRegex+'''
                        '''+endOfWord+'''
                            (?:[\.]|
                            [+]{2}|
                            (?:\s(?:and|&|or))|
                            (?!\s(?:
                            get|
                            can|
                            may|
                            '''+negativePrior+'''
                            '''+nearTime+'''|
                            '''+persist+'''|
                            '''+deteriorate+'''|
                            '''+resolvedNoContext+''')
                            '''+endOfWord+''')(?:\W|$)
                                                )|
            (?:no|not)\s
                (?:\w\s)?
                    '''+signRegex+'''\s?since(?![\.])''',
                re.I|re.VERBOSE)
```

Figure 6.3.I: The prohibitory contextual component designed to recognise where documentation is cautioning to be aware of the development of a clinical sign that is not

6.3.3.5 Clinical sign specific contextual rules

Where sequential refinement had identified the need for additional rules for a given sign these were added as components of a prohibitive or permissive method and accessed via a dictionary. The classifier recognised whether additional rules were associated with a given sign during processing, this facilitated mechanisation by the data handling class and provided ready scope for expansion and inclusion of additional clinical signs.

Sign-specific prohibitive rules overrode a previous match, usually where there was clear documentation of the absence of the sign in question, or a specific cause for false positives had been recognised during refinement. The ataxia prohibitive addition for example matched where a wobbly tooth was likely to

have been mis-identified as wobbliness (Figure 6.3.m). Sign-specific permissive rules tended to ensure that very clear notation of the presence of a sign were captured (Figure 6.3.n)

```
    def ataxiaExtra(self):
        # Override a finding of ataxia where the word 'wobbly' and
        # tooth related words occur in the same sentence
        # needs to be separate to allow the nesting to work
        toothRelWords = '''(?:
                dent|
                t[eo]+th|
                cani|
                mola|
                inci|
                prem|
                lo+se|
                e[xtra]+ct|
                rem\w*v|
                take\so|
                pul|
                lowe|
                up+er|
                gum|
                ging|
                gr\w*e|
                gde|
                pain|
                disco|
                \d{3}
                    )'''
        regex = '''
            (?:[\.?!"]|^)
                [^\.]*
                    (?<![a-z])'''+toothRelWords+'''
                        [^\.]*
                            wob+l|
            wobbl[^\.]*
                (?<![a-z])'''+toothRelWords+'''
                    [^\.]*
                        (?:[\.?!"]|$)'''
        return regex
```

Figure 6.3.m: Example of overriding prohibitory sign specific exception, this forms one of the methods of Brian.py's additionalProhibitoryFilters class

```
    def dyspnoeaExtra(self):
        regex = '''
                (?:[\.,?!"]|^)
                    (?:
                        [^\.,]
                        (?<!whimper)
                        (?<!whin))*
                            '''+oneWordNegBehind+'''
                                puffing|
                '''+oneWordNegBehind+'''
                (?<![a-z])
                    v(?:ery)?
                        (?![a-z])
                        \W*
                        dy[ps]+n[oei]'''
        return regex
```

Figure 6.3.n: Example of an overriding permissive sign specific exception. This forms a method within the additionalPermissiveFilters class of Brian.py. This particular permissive exception ensures that unambiguous notation of severe dyspnoea is captured.

6.3.4 Physiological parameters

A parameter extraction class managed extraction of numeric parameters, this incorporated patterns for the extraction of body temperature, heart rate and respiratory rate and was readily expandable where other measures were desired. Patterns recognising the nature of a numeric value were similar to the narrative clinical signs with incorporation of contextual cues and grouped words as entities with prohibitive and permissive effect.

As an example of the nature of the hierarchical extraction system there were four pattern components to the respiratory rate extractor, the main entities incorporated were `nonRespUnits` which acted as a filter for numbers representing measures other than respiratory rate and `nonRespNegBehind` which was a fixed width negative look behind to exclude matches and thus extraction where specified matches occurred (Figure 6.3.o).

```
nonRespUnits = '''(?:
                ui|
                unit|     #a group of terms that occur adjacent
                [ky%]|    #to numbers and assist in differentiation
                gram|     #between respiratory rate and other
                [ckg]m|   #measurements
                m[gml](?!pm)(?!pink)(?!\spink)|
                mile|
                [dm]|
                k[mg]
                day|
                week
                wk|
                month|
                year|
                yr|
                compr|
                st|nd|rd|th|
                @
                )'''

nonRespNegBehind = '''
            (?<!hr|pr|on)
            (?<!stt|
                eye|        #a series of fixed width look behind
                lid|        #assertions to differentiate
                cor|        #between the meaning of numeric content
                scl|
                ulc|
                bra|
                ta[ckh]|
                r[hythm]{2}|
                gal|
                car|
                vhs|
                crt|
                bcs|
                mcs|
                aim|
                hip)
            (?<!mmhg|temp|conj|card|need|conj|fore|hind)
            (?<!grade|needs|axila)
            (?<!target|cornea|cleara|nctiva|axilla)
            '''
```

Figure 6.3.o: Example of entities incorporated into the parameter extraction method. Here the patterns used to exclude extraction where an integer is a non-respiratory rate measure.

The primary respiratory rate extraction pattern focused on key indicators that a nearby integer represented a measured respiratory rate. There were positive recognition components, such as a number that followed the phrase "respiratory rate" encoded as `(?<!\w)(?:r(?:e?sp(?!o)[a-z]*)(?:\s?rate)?` and negative components to prevent extraction of other values. With respect to respiratory

rate the main values that interacted with the positive patterns related to the digits (toes), and findings of the Schirmer test which measures tear production in the eye, these as a result formed the mainstay of the prohibitive components (Figure 6.3.p).

```
respRate1 = re.compile('''
        (?<!(?:if|be|is)\s)
        (?<!(?:aim|for)\s)
        (?<!target\s)
            (?:
            (?<!\w)(?:r(?:e?sp(?!o)[a-z]*)(?:\s?rate)?|
            [rs]rr|
            [br][:;./]?r[:;./]?))
            (?![a-z])(?!\W[cva]\W)(?!dig)
            [^\.a-z0-9\s]*\s?
            (?:[a-z\-_,\(=]+
            '''+nonRespNegBehind+'''\s){,3}
                \W*
                (?:\s?(?:&gt;|&lt;|<|>)\s?)?
                (?<!ex\s)
                (?<!dig\s)
                (?<!digit\s)
                (?<![#x;])
                (?<![#x;]\s)
                ((?:[1-9][0-9]{1,2}|[5-9])(?:\s?[*x\-]\s?\d+)?)
                    (?![\./][0-9])
                    (?![0-9])
                    (?!'''+nonRespUnits+''')
                    (?!\s'''+nonRespUnits+''')
                    (?!\sa\W)
                    ''',
    re.I|re.VERBOSE)
```

Figure 6.3.p: The preliminary respiratory rate extraction pattern.

6.3.4.1 Parameter extraction calculated values and fail safes

In addition to reliance on pattern recognition, parameter extraction was limited to physiologically feasible values. These were predefined by default as the values that were physiologically likely to occur in a dog or cat, that is not the normal ranges of the parameters but the values outside which a clinician would consider it unlikely to represent the parameter of interest. As the focus of parameter extraction was population wide surveillance, where these values resulted in extreme outliers being excluded this was not problematic.

Where more than one value was extracted by a given parameter pattern, either because the record related to more than one animal (as may occur with a litter, for example) the clinician recorded sequential observations in the same animal, or a target rate had erroneously been extracted. In these cases the mean of the extracted values was calculated.

When a clinician measures a rate, it is common to observe and count for a shorter period than that in which the rate is expressed, and then to multiply by a suitable factor to record the rate. For example, where it is regular, heart rate may be auscultated over 20 seconds and multiplied by three to give the rate as beats per minute. Some clinicians document this as a sum rather than its product; for example, 32 beats measured over 20 seconds may be recorded as "20x32" rather than 96 beats per minute. A method was added to recognise these instances and calculate the product on extraction.

A further method was applied to identify the occasional ‘exclamation typo’ caused by the character ‘1’ and the character ‘!’ being located on the same key on a standard keyboard; where a two-digit number was extracted the immediately preceding character was examined and if it was an exclamation mark 100 was added to the number, correcting a documented heart rate of ‘!45’ to the intended 145. There was concern that this may be erroneously triggered where the exclamation point was in fact the clinician's observation of an extreme bradycardia, however in two years of screening records no instances of this were observed.

6.3.4.2 Frequency of parameter extraction in relation to species

As discussed in Section 4.5 there were differences in the volume of numeric content of consultations regarding dogs and cats, compared to those of other animals. The parameter extraction framework, with an additional algorithm for body weight extraction, was used to examine the frequency with which parameters were identified within consultation narratives for the four species groups used in Chapter four.

6.4 Classification efficacy

6.4.1 Inter-classifier reliability

For all four clinicians, there was good agreement with the classifications assigned by the author, with kappa ≥0.86 for three of the four pairings at initial reading and kappa ≥ 0.95 for all four comparisons after re-reading those consultations where there was disagreement between coders (Table 6.4.a).

Table 6.4.a: Inter observer agreement in clinical sign classification. 1st and 2nd refer to agreement after first reading and after 2nd reading of consultations where coding differed.

Manual coder	Narratives coded	Classifications coded	Coding differs 1st	Coding differs 2nd	kappa 1st	kappa 2nd
a	58	580	4	0	0.86	1
b	67	670	5	2	0.88	0.95
c	40	400	8	1	0.6	0.95
d	163	1630	5	0	0.94	1

6.4.2 Clinical signs

The efficacy of each clinical sign classifier was adequate for its purpose, the mean sensitivity achieved for narrative signs in the unseen test dataset (using classifiers developed in the training dataset) was 98.17 (74.47, 99.9)% and mean specificity 99.94 (77.1, 100.0)% (Table 6.4.b). Inaccuracies primarily resulted from the close juxtaposition of positive and negative findings for different anatomical areas or clinical signs within the same sentence. The context sensitive framework was effective in conferring specificity to the classification system, when the classifier for diarrhoea was permitted to match outside of the context nest the specificity fell to 94.16 (93.46, 94.79) & and precision (PPV) to 40.59 (36.25, 45.09).

When used in combination to identify animals with one of a combination of gastrointestinal clinical signs, the sensitivity achieved was 99.44% (95% CI: 98.57, 99.78)% and specificity 99.74 (95% CI: 99.62, 99.83).

Table 6.4.b: Assessment of classifier efficacy in a random sample of 5,000 consultations regarding a cat or dog (the test dataset).

Clinical sign & system	Manually classified		Software classified					
	Proportion	n=	True positive	False positive	Sensitivity (95%CI)	Specificity (95%CI)	PPV (95%CI)	F_1
Upper aero-digestive								
Conjunctival inflammation	2.14%	107	106	2	99.07 (94.89, 99.83)	99.96 (99.85, 99.99)	98.15 (93.5, 99.49)	0.986
Oropharyngeal inflammation	3.54%	177	175	7	98.87 (95.97, 99.69)	99.85 (99.7, 99.93)	96.15 (92.27, 98.12)	0.9749
Sneeze	0.66%	33	33	1	100.0 (89.57, 100.0)	99.98 (99.89, 100.0)	97.06 (85.08, 99.48)	0.9851
Lower respiratory								
Cough	1.78%	89	88	0	98.88 (93.91, 99.8)	100.0 (99.92, 100.0)	100.0 (95.82, 100.0)	0.9944
Increased respiratory effort	0.52%	26	25	1	96.15 (81.11, 99.32)	99.98 (99.89, 100.0)	96.15 (81.11, 99.32)	0.9615
Gastrointestinal								
Abdominal pain	0.82%	41	40	2	97.56 (87.4, 99.57)	99.96 (99.85, 99.99)	95.24 (84.21, 98.68)	0.9639
Diarrhoea	4.16%	208	206	4	99.04 (96.56, 99.74)	99.92 (99.79, 99.97)	98.1 (95.21, 99.26)	0.9856
Haematochezia	1.18%	59	59	3	100.0 (93.89, 100.0)	99.94 (99.82, 99.98)	95.16 (86.71, 98.34)	0.9752
Vomit: Current not resolved	3.02%	151	149	4	98.68 (95.3, 99.64)	99.92 (99.79, 99.97)	97.39 (93.47, 98.98)	0.9803
Vomit: Recent +/- resolved	4.0%	200	199	2	99.5 (97.22, 99.91)	99.96 (99.85, 99.99)	99.0 (96.45, 99.73)	0.9925
Neurological								
Ataxia	0.68%	34	32	1	94.12 (80.91, 98.37)	99.98 (99.89, 100.0)	96.97 (84.68, 99.46)	0.9552
Systemic								
Lethargy	3.84%	192	186	6	96.88 (93.35, 98.56)	99.88 (99.73, 99.94)	96.88 (93.35, 98.56)	0.9688
Lymphadenopathy	0.80%	40	39	2	97.5 (87.12, 99.56)	99.96 (99.85, 99.99)	95.12 (83.86, 98.65)	0.963

6.4.3 Parameters

Where parameter detection was defined as identifying the presence of a given parameter in the record and correct identification of its value, the numeric value extraction system worked with a mean sensitivity of 99.39% and a mean specificity of 99.99% within the test dataset (Table 6.4.c).

Table 6.4.c: Efficacy of parameter extraction in random sample of 10,000 consultations (the test dataset). TP = True positive, FP = False positive, Resp.rate = respiratory rate

	Manual		Software				
Parameter	%	n=	TP	FP	Sensitivity	Specificity	F-measure
Resp. rate	1.42%	71	70	1	98.59	99.96	98.6%
Temperature	8.14%	407	405	2	99.51	99.98	99.51%
Heart rate	10.84%	542	540	1	99.63	100	99.82%

Body temperature had the lowest sensitivity with 2 of 407 temperatures not identified (false negatives), this was due to a number appearing in the text with no contextual information that could be robustly identified by the classifier but that was identifiable on manual reading.

The duplicate detection system was effective, finding the mean of duplicate entries within the constraints of physiologically likely values.

6.4.4 Prototype deployment of classification framework

The signals generated by the classification framework deployed to classify consultation records collated within the SAVSNET dataset, on a nightly basis, generated datasets of coded data of value for syndromic surveillance. Figure 6.4.a illustrates the output from gastrointestinal sign classifiers within consultation records regarding dogs. This depicts temporal trends in individual signs and also a basic syndrome of animals presenting with both diarrhoea and vomiting.

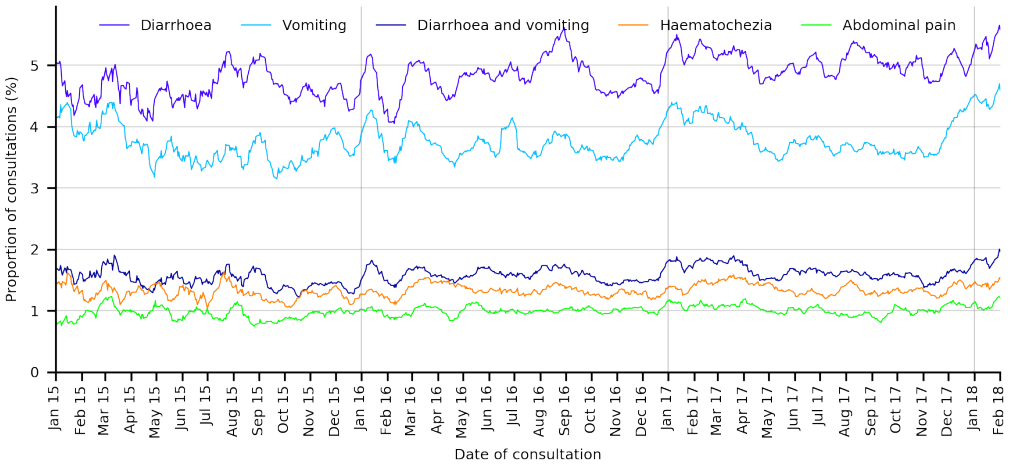

Figure 6.4.a: Temporal trend in gastrointestinal signs identified in the free-text record of dog consultations within the SAVSNET dataset by the context sensitive framework.

Figure 6.4.c illustrates the temporal trend of a syndrome classifier which identifies consultations where cats were documented to have one of a series of respiratory clinical signs (Figure 6.4.b) and also a temperature of at least 39°c (the upper limit of the normal range in cats). This syndrome classifier was generated by combining the output of classifiers for respiratory signs with the output of parameter extraction.

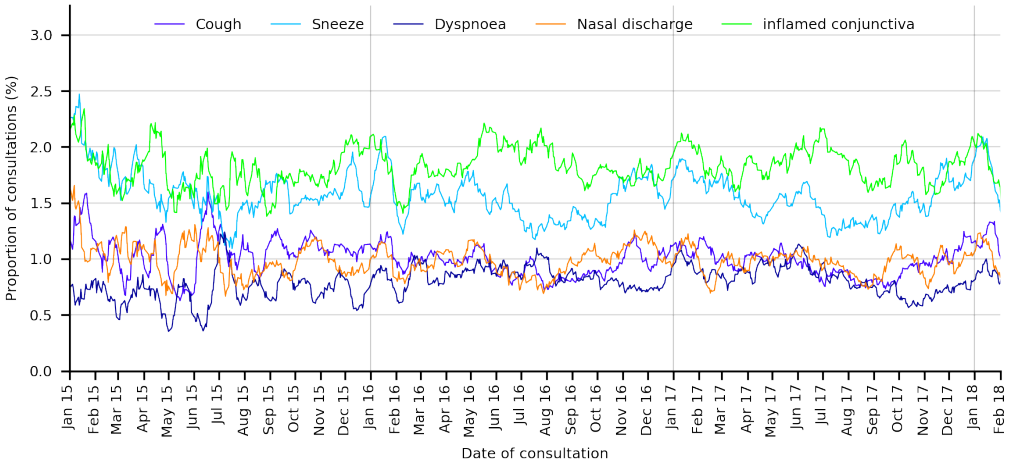

Figure 6.4.b: Temporal trend in respiratory signs identififed in the free-text record of cat consultations within the SAVSNET dataset by the context sensitive framework.

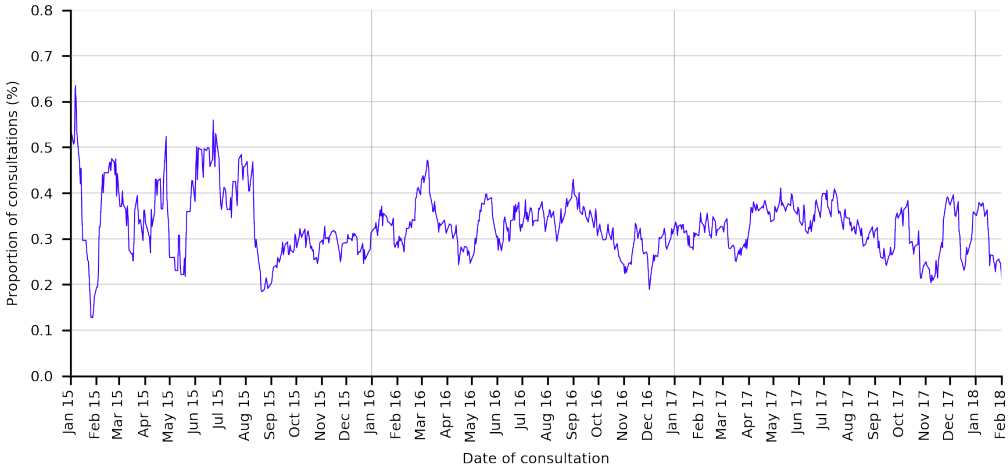

Figure 6.4.c:Temporal trend in the combined signal generated by the presence of a respiratory sign and pyrexia in cat consultations within the SAVSNET dataset.

6.4.5 Differences in frequency of parameter extraction in relation to species

Quantification of the frequency of parameter extraction across the species groups examined in Chapter four demonstrated that a likely contributor to the disparity in numeric content of consultations regarding cats and dogs compared to those of less commonly seen species was the frequency of clinicians recording physiological parameters (Table 6.4.d, Figure 6.4.d).

Table 6.4.d: Comparison of frequency with which parameters were identified within consultation narratives across species groups within the exploratory corpus described in Chapter 4.

		Proportion of consultations from which parameter was extracted			
Species	**n=**	**Heart rate**	**Respiratory rate**	**Temperature**	**Body mass**
Dog	110745	8.86 (8.7, 9.03)	1.1(1.04, 1.16)	8.04 (7.88, 8.2)	3.06 (2.95, 3.16)
Cat	47245	14.34 (14.02, 14.65)	2.01(1.88, 2.13)	9.2 (8.94, 9.46)	4.93 (4.74, 5.13)
Rabbit	2857	3.82 (3.11, 4.52)	0.67(0.37, 0.67)	4.03 (3.3, 4.75)	3.92 (3.21, 4.63)
Uncommon	2393	1.34(0.88, 1.8)	0.71(0.37, 1.05)	1.67 (1.16, 2.19)	2.47 (1.84, 3.09)

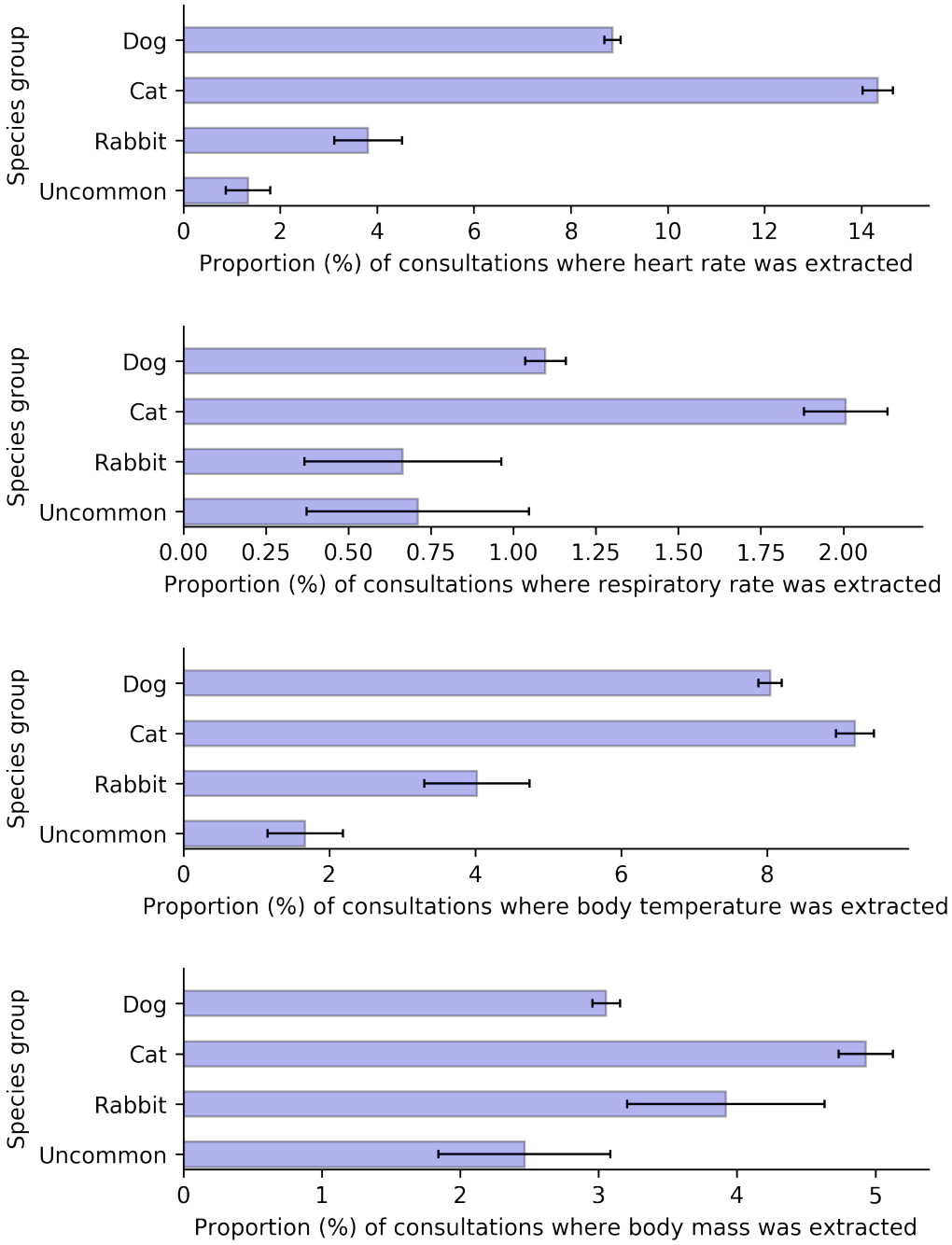

Figure 6.4.d: Comparison of frequency with which parameters were identified within consultation narratives across species groups within the exploratory corpus described in Chapter 4. Error bars reflect 95% confidence interval of the proportion.

6.5 Discussion

The system of clinical sign classifiers designed to augment the real-time syndromic surveillance capabilities of SAVSNET compares favourably with other classifiers developed for information extraction from both veterinary and human clinical narratives. Perhaps the closest comparable system is work that utilised proprietary text mining software to identify veterinary consultation narratives where enteric signs were documented, achieving a sensitivity of 87.6% and a specificity of 99.3% in a dataset of records collated from twelve Canadian veterinary practices (Anholt et al. 2014).

Syndromic surveillance systems utilising classifiers of human health care narrative records often use the Emergency Department presenting complaint field. For detection of documented gastrointestinal syndrome, meeting the definition against which they were designed, sensitivity of the Bayesian chief complaint coder (CoCo) (Chapman, Dowling, and Wagner 2005) compared to manual classification was 69.0% and specificity 95.6%. The classifiers of the Real-time Outbreak and Disease Surveillance (RODS) (University of Pittsburgh, 2016) system was found to have a sensitivity of 63% for its naive Bayes classifier and 38% for its bigram Bayes classifier, both had a specificity of 94% (Ivanov et al. 2002) For comparison, when the classifiers developed here were used to identify animals displaying one or more gastrointestinal clinical signs the sensitivity achieved was 99.44 (95% CI: 98.57, 99.78)% and specificity 99.74 (95% CI: 99.62, 99.83)%. The classifiers described here screen a larger and less concise narrative field than those of the RODS and CoCo systems.

The high sensitivities and specificities achieved render these classifiers ideal for use in a surveillance tool. The complexity of the classifiers required to optimise sign identification is testimony to the substantial lexical diversity present in a dataset contributed by in excess of 1,400 veterinarians.

6.5.1 Challenges to classification of veterinary clinical narrative by free text

The development of these classifiers exposed a number of challenges in the clinical narratives these included:

6.5.1.1 Brevity

The use of abbreviations in describing vomiting and diarrhoea was especially challenging, as has been described by other authors (Anholt et al. 2014). However, this was successfully circumvented in the majority of cases. One of the most challenging aspects of this notation was differentiating 'v' used to indicate vomit from 'v' used as an adverb to describe severity (i.e. an abbreviation of 'very'). It had been intended to use part-of-speech tagging, to augment differentiation of the two contexts on the basis of the type of words adjacent to the abbreviation. However, whilst there were clear differences in the distribution of words types in these contexts, the extensive overlap resulted in this method not being as successful as regular expression based rules utilising a series of words and their synonyms.

The utility of regular expression searching is well illustrated by the ability to detect a consultation where the presence of a sign is mentioned in the affirmative even though the sign has also been mentioned in the negative. This allowed for identification with greater finesse than more rigid Boolean searching. Matching was attempted to every word within the narrative, if a sign was documented as currently being present, having occurred historically and the owner was warned to observe for deterioration, the series of rules within the classifier would recognise that it was currently present.

As has been noted in the development of classification systems targeted to human medical narratives (Chapman et al. 2001), lists of negative findings pose a challenge to accurate classification

6.5.1.2 Verbosity

For some signs, notably haemorrhagic diarrhoea, adjective and noun were occasionally separated by thirty or more words and one or more sentences; enabling matching at such a distance in the narrative risks impairing specificity. This was addressed by matching to other words associated with diarrhoea or faeces, or an adjective used to describe faecal consistency and not commonly used to describe other tissues, discharge or urination, as an indication of the origin of blood. Conversely, the brief phrases 'blood and mucus' and 'mucus and blood' were pathognomonic of haematochezia, and this alone, provided the

other generic negative rules were obeyed, was sufficient to indicated blood being passed per rectum.

Verbosity in specific situations was advantageous to classifier development. There were occasions where a reported sign was refuted by examination, for example where a free-text record stated 'owner reports swollen eyes, ... on examination eyes NAD, no swelling, pain free', and hence was documented in the positive and negative within a single narrative. These were excluded either by the greater distance between antecedent adjective and the noun representing a body area when an owner's description was documented as compared to a veterinarian's observation, or by the presence of a definitive statement of a negative examination finding.

6.5.1.3 Lack of contextual information

Where false negative parameter identification occurred, it was commonly the result of a lack of contextual information. This was most notable for body temperature. Whilst it is eminently feasible to add further rules that can extract numbers from the text where there is no surrounding contextual information and to make an assumption based on the value as to which parameter the value represents, the number of consultations in which such strings occur was low. In contrast, the range of other parameters that lone numbers may represent was large. As a result, such an approach would risk the introduction of an increased margin of error for little benefit. Numbers occurring in the absence of any contextual indication as to their meaning were therefore purposely not sought.

6.5.1.4 Data errors

Human and software errors also posed challenges. Some practices appeared to have used their own auto-complete software with consequent incorrectly autocompleted words. For example, '5 d' might be corrected to '5 diarrhoea' rather than '5 days'. Veterinarian documentation errors accounted for a number of false positives and negatives, notably where there was reference to one body area or parameter where another was clearly intended.

Awareness of the subtle differences in language used was important, for example 'abdo tense' may reflect the response of an animal to pain on

palpation, but was most commonly used to signify a non-painful abdomen and slightly anxious dog resisting palpation. Conversely ‘abdominal guarding’ or ‘boarding’, ostensibly carrying the same meaning as ‘tense abdomen,’ was only used when the abdomen was thought to be painful. For this reason, ‘tense’ was not used as an indicator of pain whereas ‘guard’ and ‘board’ formed a component of the abdominal pain rule, although did not commonly match.

6.5.1.5 Limitations

It must be borne in mind that the classifiers are being assessed using a post-hoc gold standard. It is only possible for a classifier to identify where a sign, or its colloquialisms, is documented in the clinical narrative. Thus, a classifier with 100% sensitivity for the presence of a sign in a clinical narrative, does not necessarily have 100% sensitivity for the presence of that sign in the presenting animal because there are likely to be some animals where the presence of the sign is undocumented. Observation of veterinary consultations with subsequent examination of the associated clinical record has suggested that a third of problems discussed during a consultation are not documented within the electronic record of that consultation (Jones-Diette et al. 2017).

Manual classification was undertaken by a single classifier (JN) with medical training and considerable experience of the diverse phraseology used in consultation narratives in the SAVSNET dataset. Although logistically unavoidable, this was not ideal. However, as classification was for the presence of a series of clinical signs and parameters, and not differentiation into syndromic categories, it did not involve clinical judgment, merely familiarity with the veterinary sublanguage.

The classifiers are closely tailored/fit to the existing format of narratives within the database. As SAVSNET continues to grow and new text patterns appear denoting the presence of clinical signs and modifying context of these, both the sensitivity and specificity might decrease. Therefore, the classifiers will need to be revalidated and adapted periodically to ensure that their efficacy is maintained.

6.5.1.6 Utility of the system

These classifiers could facilitate the regular production of small animal public health reports, directly from the unstructured consultation narrative, providing information similar to the weekly 'GP In Hours Surveillance System Report' produced by Public Health England (2015) and preliminary results for this application are promising (Figures 6.4.b-d.).

The use of clinical signs to monitor the occurrence of individual or grouped clinical signs, rather than clinical diagnoses or laboratory findings, will allow identification of spatio-temporal changes in patterns of syndrome presentation that reflect emergence of disease prior to it formally being diagnosed or its presence being suspected in the population. This may be especially important in the case of emergent diseases where there is a potential for considerable delay whilst awareness evolves and definitive diagnostic tests are developed (Caliendo et al. 2013).

6.6 Conclusion

The ability to monitor individual or combinations of signs utilising the system described here, extends the ability to surveil beyond known diseases and previously identified syndromes. With the SAVSNET system, collating in the region of 20,000 consultations each week, the near real-time syndromic surveillance, directly from the narrative consultation record, made feasible by these classifiers, offers the potential for considerable clinical and public health benefit.

Chapter Seven Application of free-text classifiers in emergent disease surveillance

7.1 Introduction

This chapter describes an approach to the challenge of free-text classification when there is an urgent need for a tool to identify an emergent illness and insufficient clinical data available for extensive exploration of the context of documentation. The development of a tool to identify consultations potentially reflecting an animal with canine idiopathic cutaneous and renal glomerular vasculopathy is used to illustrate the principle.

7.1.1 Emergent diseases

Changes in environmental and socio-demographic factors contribute to situations where patterns of illness are recognised for the first time in a population, or rapidly increase in incidence or geographic range (M. L. Cohen 1998; Morens and Fauci 2013). During the emergent phase of any disease, it is likely that, for a considerable proportion of cases, diagnosis is delayed or unresolved (Caliendo et al. 2013). On a case by case basis, delays in diagnosis may have profound implications on survival outcome, as for example in the case of canine idiopathic cutaneous and renal glomerular vasculopathy (CRGV) where the disease process progresses rapidly to the point where the kidneys are irreparably damaged (Holm et al. 2015).

Early detection of aberration from the previous norm, situational response and awareness are core components of surveillance systems (Paterson and Durrheim 2013), and this can be facilitated by the integration of free-text classifiers (Conway, Dowling, and Chapman 2013). A near real-time surveillance system incorporating a series of context-sensitive free-text classifiers, as described in Chapter six, would permit identification of anomalies in the patterns of combinations of clinical signs, in space and or time. Such a system would however depend on classifiers for the clinical signs forming the pattern of a given syndrome having been developed and deployed, prior to the emergence of illness.

7.1.2 Canine idiopathic cutaneous and renal glomerular vasculopathy

CRGV was first described in the 1980s amongst raced greyhounds in the United States (Carpenter et al. 1988). Affected dogs developed ulcerated skin lesions with or without acute kidney injury (AKI). As a large number of the early cases occurred in dogs raced at Greenetrack Racing Park, Alabama, the disease

became known as Alabama Rot. Over the intervening three decades since Carpenter's description a number of clusters of simililar illness have been described in the United States (Hertzke et al. 1995; Cowan et al. 1997) and isolated cases in Europe (Rotermund et al. 2002).

Histopathologically, dogs with CRGV have thrombotic microangiopathy. This is the same histopathological finding identified in humans with haemolytic uraemic syndrome, which is often caused by shiga toxin-producing bacteria (Carpenter et al. 1988; Hertzke et al. 1995; Mayer et al. 2012); however, shiga toxin has not been identified in UK dogs with CRGV (Holm et al. 2015). The cause of CRGV remains unknown.

Since November 2012, a number of dogs have presented to veterinary practices across the United Kingdom in a similar manner; with cutaneous – and subsequently renal – manifestations of a disease of the small blood vessels (Walker et al. 2015; Walker et al. 2016; Walker 2014; Walker et al. 2014; Holm et al. 2015). By January 2018, clinicians collating data regarding cases of CRGV had identified 122 histologically confirmed cases of CRGV (Woodmansey 2018), based on the finding of thrombotic microangiopathy in renal or cutaneous samples.

In the UK, CRGV was initially thought to only affect dogs from the New Forest, Hampshire (Walker et al. 2015); however, cases have now been identified across the UK (Holm et al. 2015). Although a relatively small number of dogs are known to have been affected by CRGV, the emergence of a rapidly life-threatening illness with acute and unexplained onset, potentially attributable to an environmental toxin or pathogen, has understandably generated considerable public and professional concern. There has been notable social media (@AlabamaRot.CRGV.kills.dogs 2018) and lay press coverage of suspected and confirmed cases and crowd funding for research with the establishment of the New Forest Dog Owners Group Research Fund in 2014 (New Forest Dog Owners Group 2018) and Alabama Rot Research Fund in 2016 (Alabama Rot Research Fund 2018).

It was proposed to create a tool that performed a lexical analysis of the consultation record, creating a classification based on the patterns of words used. This would be capable of identifying cases described using language

similar to that used to describe CRGV and differentiate them from other reasons for presentation of an animal to a veterinary practice. As CRGV is a very rare presentation with manifestations in two individually common affected body systems, skin disease and kidney disease, then a key feature of any algorithm developed for detection of CRGV must be a high specificity i.e. ability to correctly identify non-cases with a very low proportion of false-positives.

To address this challenge of responsively developing and deploying a surveillance system to detect a newly emergent disease for which only a few clinical records of confirmed cases were available, a novel approach was proposed based on lexical analysis of these few confirmed cases. It was hypothesised that the language used in describing the presentation of CRGV in first opinion clincal records would contain sentinel lexical features that could subsequently be used to identify suspect cases of the disease, in the absence of specific diagnostic terms, within a dataset of veterinary consultation narrative records.

7.2 Materials & methods

7.2.1 Free-text records of confirmed cases

In response to a letter in the Veterinary Record (Walker et al. 2014), suspected cases of CRGV or post-mortem tissues from suspected cases were referred to Anderson Moores Veterinary Specialists (AMVS), a referral practice in Winchester, UK, in an attempt to collate data nationally on this illness, and hence try to further the understanding of CRGV. Ethical approval to develop text-mining techniques using these data was gained from the University of Liverpool Veterinary Research Ethics Committee (Reference VREC225) and the clinical records of the first 55 histologically confirmed cases were provided to the author and 45 of them were utilised in the development of a programmatic CRGV surveillance tool, which was named Ping.

The records of ten cases were excluded because they consisted solely of hand-written documentation unable to be reliably transcribed, a referral letter or laboratory results. Referral letters were excluded from the narrative processing because narrative specifically intended for communication to a referral centre is structured very differently to first opinion consultation records, in order to maximise clarity (Meystre, Savova, Kipper-Schuler, and Hurdle 2008b), and thus is not directly comparable to the consultation notes.

7.2.2 Selection of a geographically matched control dataset

It was important to avoid inadvertently building a lexical model of the regional language used in the geographic location of the small number of CRGV cases, rather than the language used to describe the presentation of the illness. The postcodes of the confirmed cases were provided, as a list of postcodes with no link to their respective consultation narratives, two postcodes were not available.

A script, written in Python, was used to identify dog consultations collated within the SAVSNET dataset where the owner's postcode was within a specified radius and had been assigned the same measure of rurality or urbanicity, following the 2011 census (Office For National Statistics 2015b), as one of the case postcodes.

Preliminary scoping identified that using this method a radius of 30 kilometres permited the matching of 1000 controls to the majority of cases, where this was not possible the matching criteria were stepped back to identify controls within the same radius irrespective of rurality measure (Figure 7.2.a).

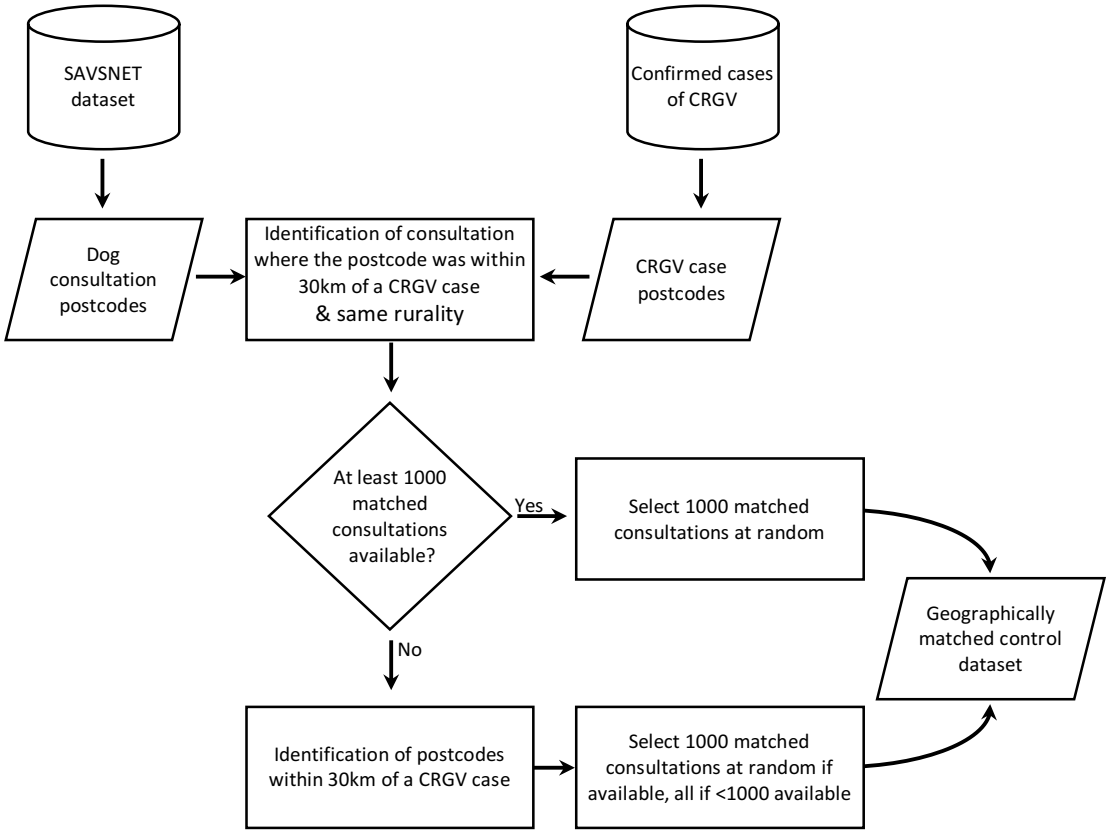

Figure 7.2.a: Selection of a geographically matched control group of consultation narratives.

7.2.3 Text pre-processing

The CRGV consultation records were supplied in a variety of document formats, predominantly scanned image pdfs, and required transcription. Owner identifiers, such as name, address, postcode and telephone numbers, had been redacted in these clinical records. Records were transcribed verbatim, including typographic and spelling errors, with the exception that white space was transcribed as a single space to reflect the white space cleaning process that had been applied to the SAVSNET text data. A text file (CSV), for later upload into a Pandas dataframe, was created containing each consultation narrative record, a unique animal and consultation identifier and how many days from presentation of the animal each consutlation had occured.

Incidental owner and veterinarian identifiers within the consultation narratives of the SAVSNET dataset were redacted using Clancularius, a de-identification tool, as described in Chapter five. Dispensing labels, introduced into the clinical narrative by practice management software, were moved into a separate field using a pattern recognition tool as described in Chapter two. In addition to containing information with the potential to generate false positive classification, such as warnings regarding potential adverse drug reactions, these labels were practice management software specific and had the potential to bias the classification tool given that all SAVSNET consultations used in this work were collated via a single practice management system.

7.2.4 Descriptive lexical analyses

Crude word frequencies were calculated in the narrative records of the case and control datasets using a method that utilised Pandas' string comprehension functions, to identify all words and then count their frequency, in a process analogous to that used in Chapter four to explore word counts (Figure 7.2.b).

```
def generateWordFreqTable(df, narrativeCol):
    #find all words in each narrative field
    df['wordsInNarr'] = df[narrativeCol].str.findall(
                    """(?:(?:[\d]+|(?<=\s))[\.]\d+|
                    [a-z0-9<>\^+~\']+(?:\-(?=\W|$))?)""",
                            flags = re.I|re.VERBOSE)
    #convert the individual narrative word lists
    #into a list of words in whole dataset
    wordList = [word.lower() for wordList in df['wordsInNarr'].tolist()
                    for word in wordList]
    #calculate how often each word occurs
    #and relative frequency in dataset
    wordFreqDf = pd.DataFrame(pd.Series(wordList).value_counts())
    wordFreqDf.index.name = 'word'
    wordFreqDf.columns = ['num']
    wordFreqDf['prop'] = wordFreqDf['num']/wordFreqDf['num'].sum()
    return wordFreqDf
```

Figure 7.2.b: Python function used to extract a word frequency table from the narrative field of a dataset.

Stop-words, those words occurring with such uniform frequency as to be of little value in differentiating text, were ignored in the frequency tables and words that potentially carried a clinical (eg. erythema, ulcer) or temporal (eg. acute, sudden, recently) meaning were manually identified. The frequencies of these words in consultations regarding CRGV were compared to those in the geographically matched control consultations drawn from the SAVSNET dataset. It was acknowledged that the control group may have contained cases of CRGV, and as such was not strictly a non-case, control, group; however, with an extremely low prevalence the effect of any cases within the control group would be diluted by its relatively large size.

Semantic clusters of words, entities, with similar clinical meaning were identified and their strength of association with confirmed cases of CRGV, compared to the control consultations, was established by calculating univariable odds ratios. Regular expression search phrases permitted identification of words in the face of common misspellings, abbreviations and acronyms. Words with more than one meaning were mapped to any clinical or temporal meaning, for example “sore” could be used to describe pain and a wound and so belonged to both semantic clusters (Figure 7.2.c).

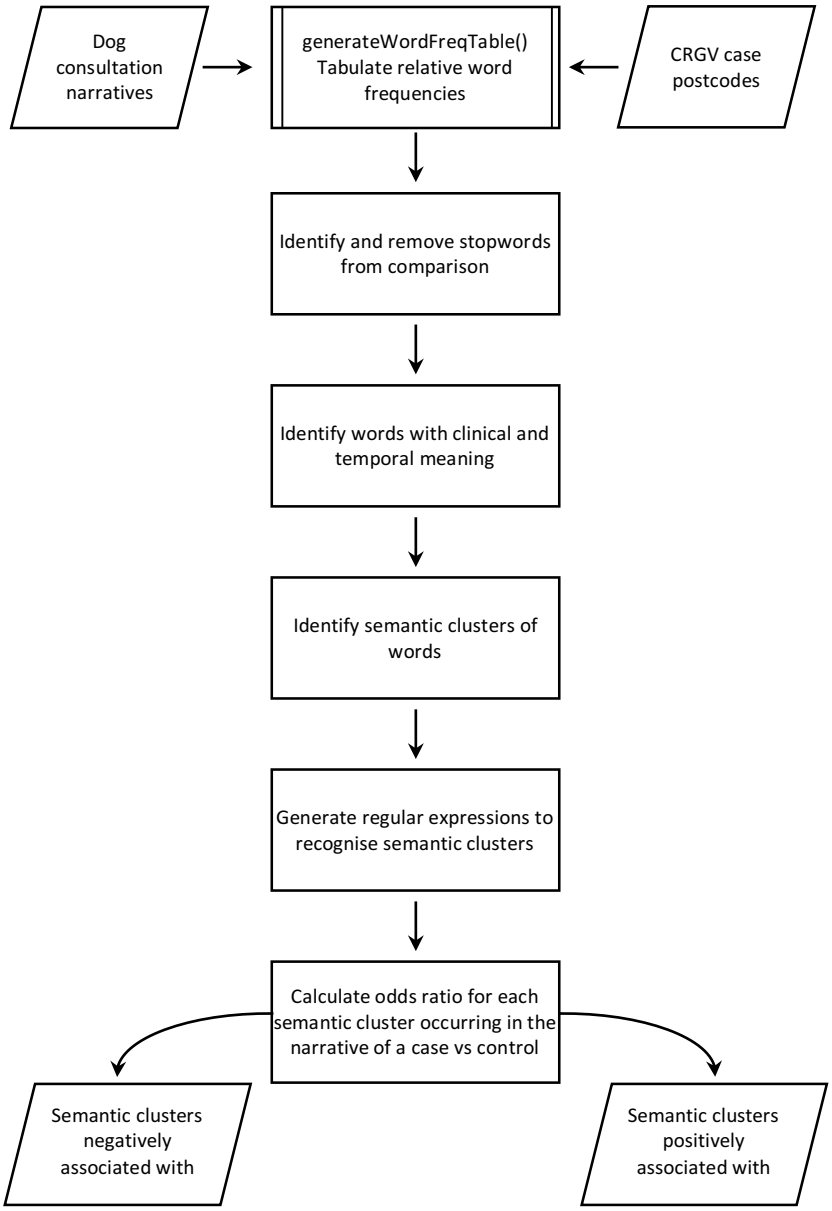

Figure 7.2.c: Flow diagram of the process of identifying candidate semantic clusters from the relative word frequencies in the case and control narratives

Diagnostic words including “Alabama Rot”, “vasculitis” and “CRGV” were excluded from the process as these required the veterinarian to have recognised the illness and would not be used in a substantive fraction of the cases of interest, and thus would bias towards the known cases and later stages of illness. Similarly, words indicating discussion or documentation of euthanasia were excluded, as a tool trained towards identifying animals after death was of less practical application. The dataset was annotated to indicate the presence of text patterns matching each regular expression within each consultation narrative.

7.2.5 Building the predictive model

The binary classifications denoting semantic clusters, groups of words with the same functional meaning, positively associated with confirmed CRGV cases were used to build multivariable logistic regression models. Initially a series of models were built by randomly selecting semantic clusters from those with positive association. The models were then refined sequentially dependent on their multivariable strength of association and efficacy of prediction within the training set.

Each logistic regression model was used to generate a prediction value between zero and one for each consultation within the dataset, and the number of non-case consultations was identified for each new case correctly identified. Having identified the semantic clusters that produced a multivariable model with maximal sensitivity, and minimal compromise to specificity at the lower limit of prediction, those semantic clusters were used in combination with incrementally adjusted semantic clusters with negative association with CRGV cases. This was undertaken in two manners in an attempt to identify the optimal combination; 1) the predictive value at which the positive model began to compromise specificity was identified and only those consultations with this value of prediction or greater were used to build the negative association model; 2) the negative entities were used within a model in combination with the entities of the positive model.

This process generated a series of algorithms whose efficacy could be compared to identify the optimal tool for identification of an extremely rare disease within a large dataset whilst minimising false positives. Specificity was calculated assuming that dogs within the control group were not cases of CRGV.

7.2.6 Evaluation of the multivariable model as a screening tool in syndromic surveillance

Receiver operating characteristic curves were plotted to visualise the efficacy of each model and identify the optimal combination of semantic clusters and the appropriate level at which a signal of high risk consultation should be triggered, this was referred to as the CRGV possibility threshold.

Having built the multivariable model and chosen the CRGV possibility threshold, the earliest consultation identified by the model for each confirmed case was compared to the manually extracted first consultation at which the clinician attending confirmed cases documented recognition that the presentation may represent CRGV.

The optimised model, Ping, was applied to all dog consultations collated within the SAVSNET dataset and the temporal and spatial distribution of consultations meeting the CRGV possibility threshold of Ping's algorithm were plotted. A sample of 100 flagged consultations was selected at random and manually assigned by the author to one of three categories based on the perceived likelihood that the consultation represented a case of CRGV: highly unlikely to represent a case of CRGV, may conceivably be a case of CRGV but this appeared unlikely, and CRGV would have appeared in the author's working differential diagnosis based on the information documented by the attending clinician.

7.3 Results

7.3.1 Description of training dataset

The 45 confirmed cases of CRGV provided 189 consultation narratives relating to their CRGV illness. It was possible to match 1000 control consultations to 48 of the 53 case postcodes, stepping back to match at a radius of 30 kilometres, regardless of rurality, permitted identification of 1000 controls for 4 of the remaining case postcodes, the final postcode was in an isolated location on the Cornish peninsular and only 94 controls were available at a radius of 30 kilometres, these were all selected. The geographically matched control group consisted of 53,094 consultation narratives. The distribution of the consultations within the SAVSNET dataset as a whole and the matched controls is demonstrated in Figure 7.3.a.

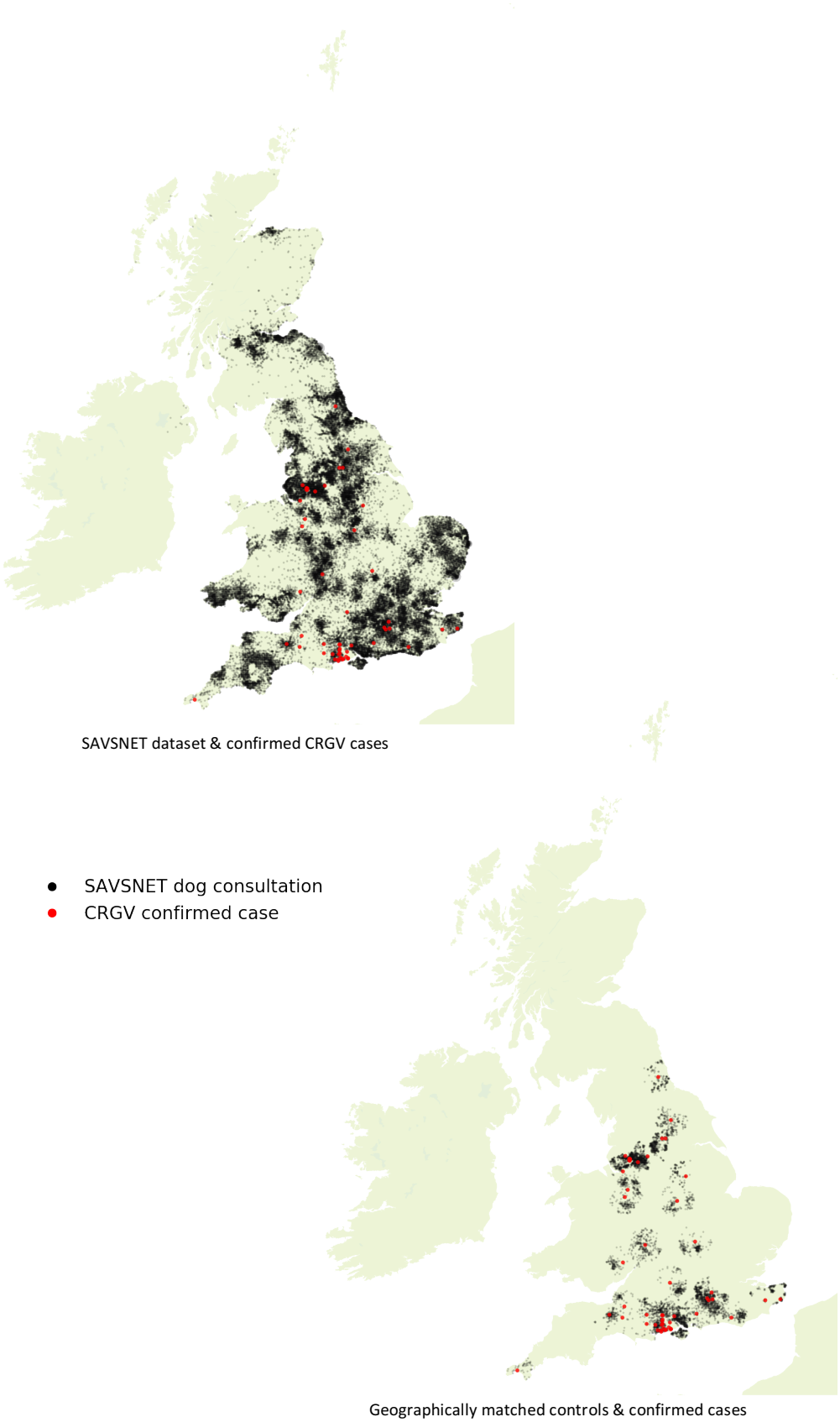

Figure 7.3.a: Spatial distribution of dog consultations collated within the SAVSNET dataset and confirmed cases of CRGV.

7.3.2 The multivariable model

Using semantic clusters with positive association to CRGV cases it was possible to generate a model with a sensitivity 95.56% of and specificity of 98.3% within the training dataset (Figures 7.3.b & 7.3.c).

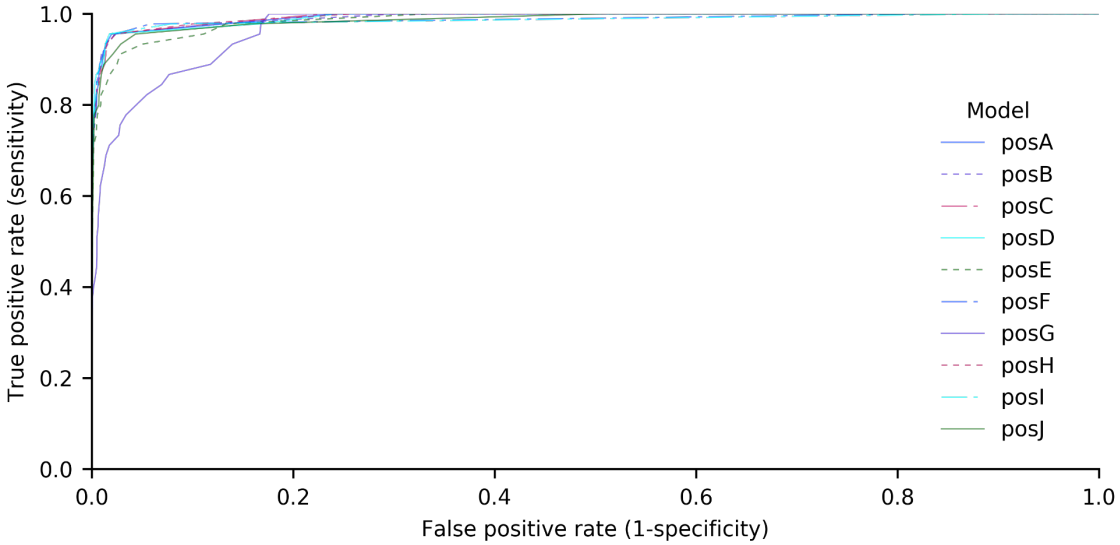

Figure 7.3.c: Initial receiver operating characteristic curves for the series of models (posA - posJ) generated using combinations of semantic clusters with positive association to CRGV.

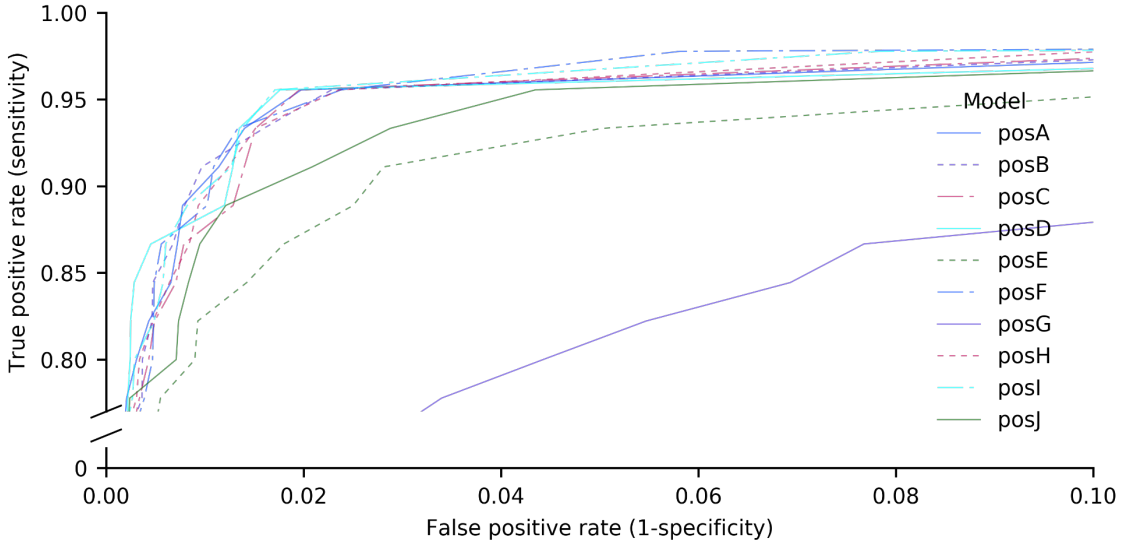

Figure 7.3.b: A focused window of the receiver operating characteristic curves for models generated using semantic clusters with positive association to CRGV, to enable evaluation of the comparative efficacy of the models.

Results

Selecting the model that performed best, the model labelled 'posI' in figures, and sequentially combining it with semantic clusters with negative associations (Figure 7.3.d) improved the efficacy, permitting generation of a model able to improve the specificity to 99.39% at the same sensitivity (Figures 7.3.e & 7.3.f).

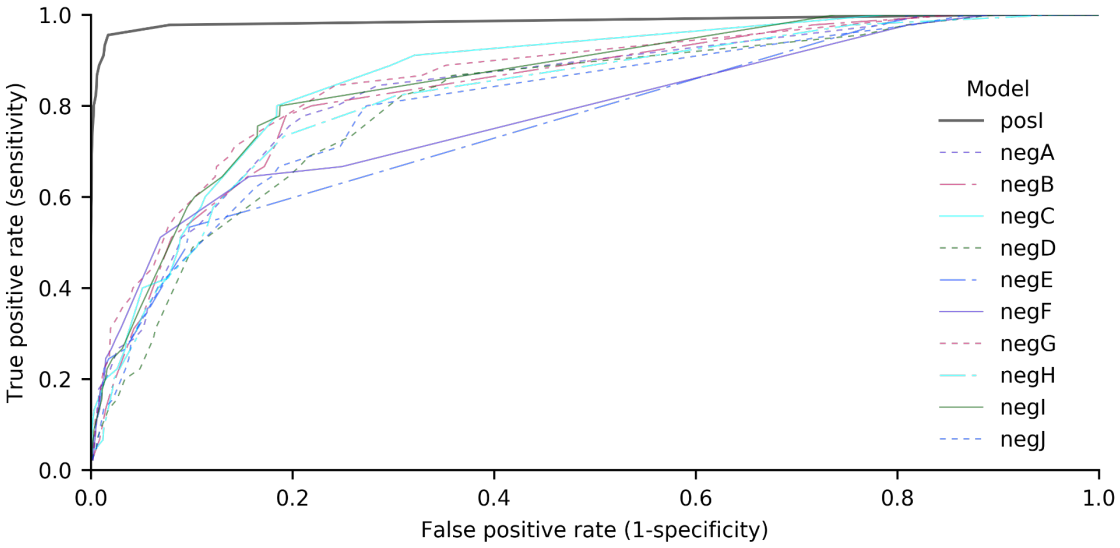

Figure 7.3.d: Receiver operating characteristic curves for the series of models (negA - negJ) built using semantic clusters with negative association with confirmed cases of CRGV. The curve of the optimal model of positive association (posI) is shown for reference.

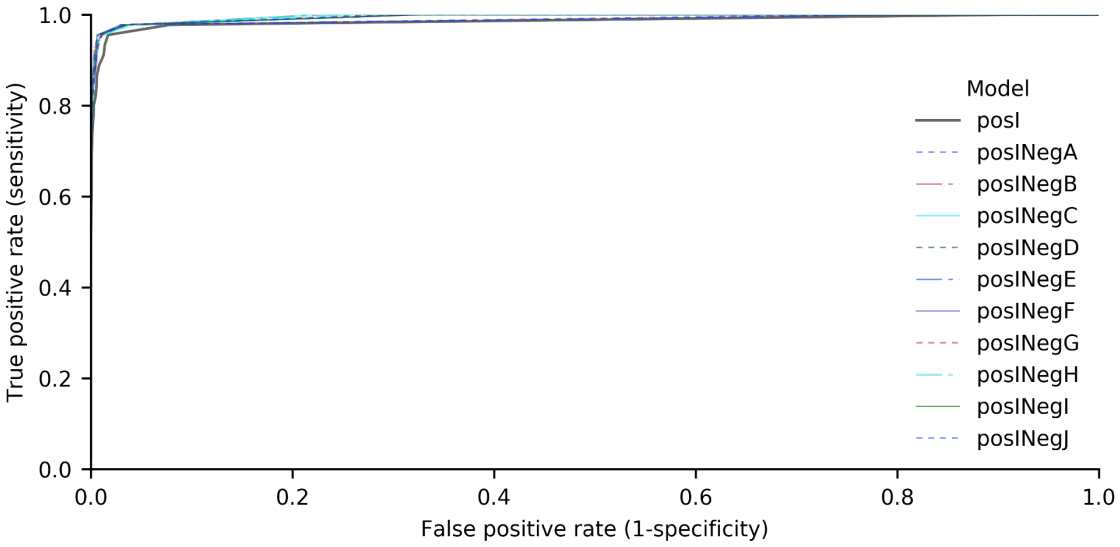

Figure 7.3.f: Receiver operating characteristic curves for the series of combined models of positive and negative associations. Each model combines posI with one of the negative association models, the curve of posI alone is shown for reference.

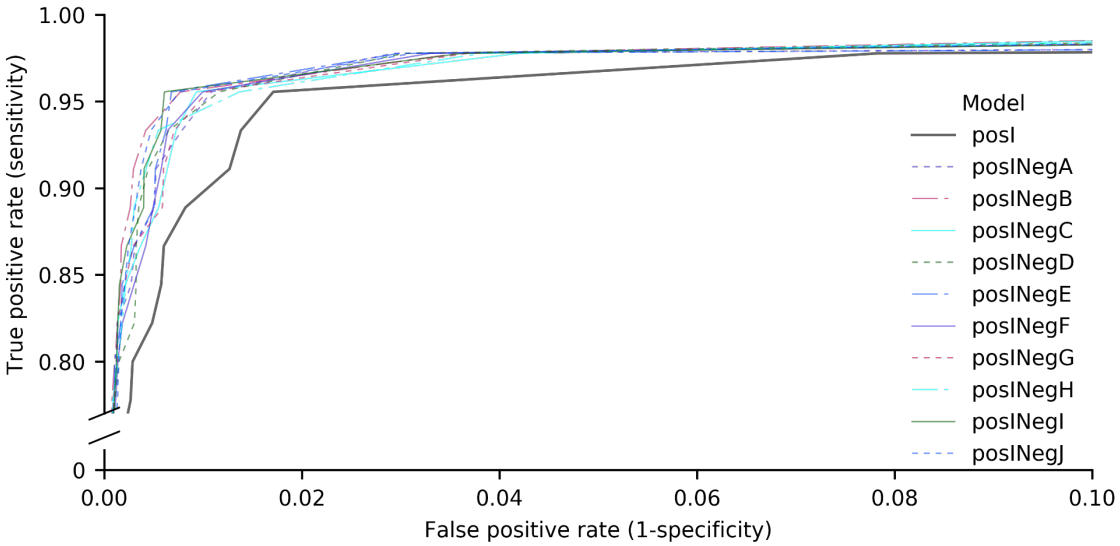

Figure 7.3.e: A focused window of Receiver operating characteristic curves for combined models of positive and negative associations, to enable evaluation of the comparative efficacy of the models. Model posINegI was chosen.

The semantic clusters included in the model chosen as optimal for purpose included those that could have been anticipated, associated with acute renal disease, coagulopathy, skin lesions and an acute onset rapidly progressive illness. Clusters with negative association to CRGV, and thus improving the specificity of the model, were associated with chronicity, management and normality (Table 7.3-a). Although the multivariable strength of association for many of the semantic clusters fell below statistical significance, their inclusion in the model strengthened its predictive capacity.

Table 7.3.a: Semantic clusters incorporated into the final multivariable model (posINegI)

Semantic cluster	Multivariable odds ratio (95% CI)	*p*-value
renal	12.69 (8.20,19.63)	$4.0x10^{-30}$
aki	6.90 (4.57,10.43)	$3.9x10^{-20}$
clotting	5.48 (3.05,9.85)	$1.4x10^{-8}$
prognosis	4.08 (2.57,6.47)	$2.7x10^{-9}$
lesion	3.98 (2.83,5.61)	$2.6x10^{-15}$
referral	2.92 (2.01,4.24)	$1.7x10^{-8}$
depressed	2.87 (1.98,4.16)	$2.9x10^{-8}$
rapid	2.82 (1.44,5.50)	0.002
idiopathic	2.29 (1.05,5.00)	0.037
tablet	1.87 (1.00,3.50)	0.049
shock	1.55 (0.82,2.91)	0.178
exudative	1.44 (0.80,2.59)	0.227
otherOrgan	1.33 (0.93,1.92)	0.123
respiratory	1.27 (0.72,2.24)	0.418
medication	1.21 (0.73,2.02)	0.455
analgesia	0.96 (0.64,1.45)	0.855
diarrhoea	0.81 (0.47,1.41)	0.458
negative	0.75 (0.21,2.77)	0.672
nsaid	0.62 (0.34,1.14)	0.125
think	0.56 (0.27,1.17)	0.121
long	0.41 (0.19,0.87)	0.019
quite	0.41 (0.17,0.97)	0.041
heart	0.41 (0.21, 0.77)	0.006
pale	0.39 (0.15, 1.04)	0.06
issue	0.38 (0.13, 1.05)	0.063
appetite	0.31 (0.10, 0.94)	0.038
qualityOfLife	0.24 (0.05, 1.06)	0.059
generalAnaesthetic	0.22 (0.07, 0.70)	0.011
murmur	0.16 (0.04, 0.66)	0.011
eye	0.10 (0.04, 0.27)	$6.0x10^{-6}$

Table 7.3.b: Example of semantic clusters included in the final model

Semantic cluster	Regular expression	Example of included terms
pupd	pu|pd|p\W*[ud]\W*p\W*[du]|poly\s?[ud]	pu pd pupd pdpu pu/pd polyuria polydipsic
exudative	exud|ooz|pu[rul]+ent|smel+|weep	exudative oozing purulent smelly weeping

7.3.3 Deploying the model as a tool for surveillance of CRGV

During the period January 2016 to January 2018, when the SAVSNET dataset rate of collation was relatively stable with approximately 14,500 dog consultations collected each week, a mean 16 (95% CI: 15.1, 16.8) consultations per week reached the Ping algorithm's CRGV possibility threshold. A formal analysis over a longer time period would be required to interpret, but there appears to be an element of seasonality to the trend of consultations reaching the Ping algorithm threshold, with periods of increased signal during the early months of the year and autumn (Figure 7.3.g) consistent with that observed in CRGV cases in the UK (Holm et al. 2015) and US (Cowan et al. 1997).

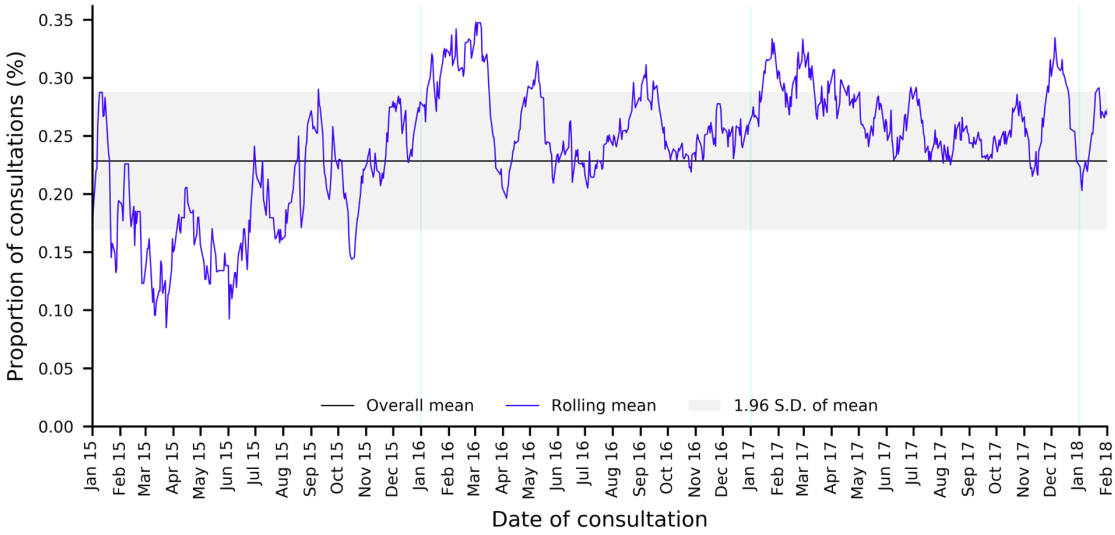

Figure 7.3.g: Temporal trend in the pattern of consultations reaching CRGV possibility threshold using Ping's algorithm

Ping flagged consultations throughout the UK, including in areas such as Scotland, West Wales, East Anglia and Northern Ireland which were not represented in the training set, because at the time there had been no confirmed cases in those areas (Figure 7.3.h).

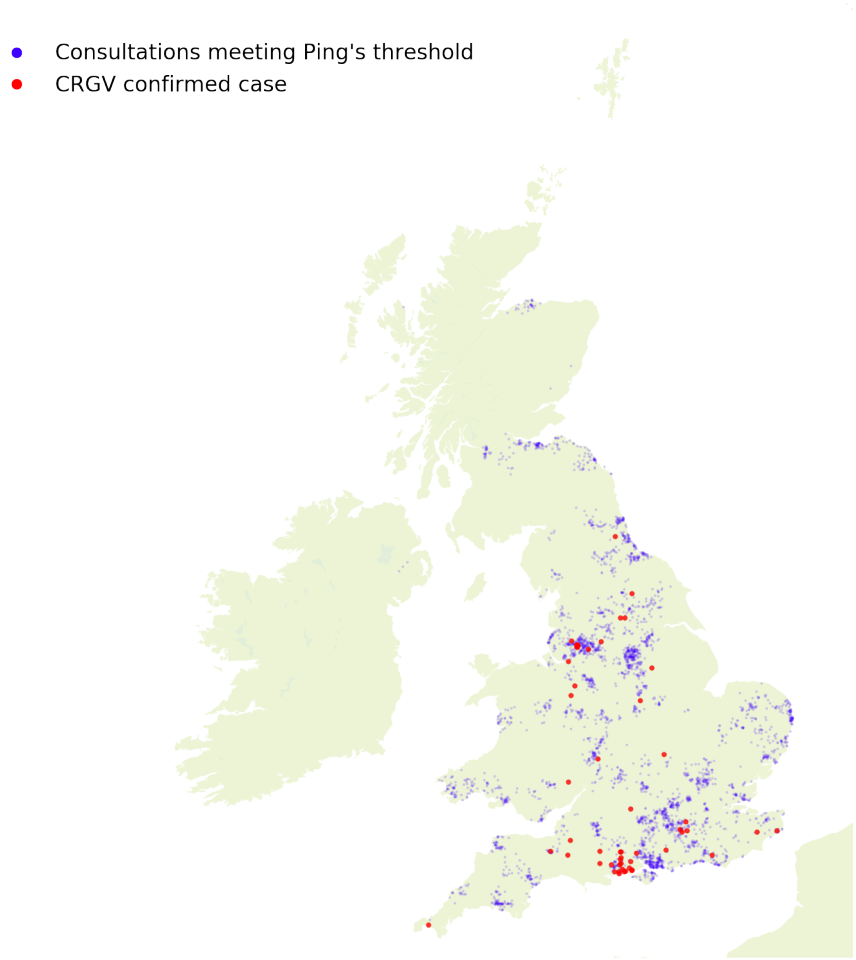

Figure 7.3.h: Spatial distribution of consultations highlighted as matching the lexical pattern of a CRGV consultation by the predictive model of Ping

On manually reading 100 of the flagged consultations the author considered 58 highly unlikely to represent cases of CRGV, 17 may have described a case of CRGV but this appeared unlikely, and for 25 of the consultations CRGV would have appeared in the author's differential diagnosis based on the information documented by the attending clinician.

Of those consultations where CRGV appeared unlikely: 7 had chronic renal disease, 6 had skin lesions for an identifiable reason other than CRGV, including demodectic mange and ulcerated tumours, 5 had an identified mass or

known malignancy. There was one case of grape ingestion and three of ibuprofen ingestion, both of which are nephrotoxins in the dog, and also one case of chocolate ingestion with a concurrent surgical wound. Amongst those consultations where CRGV appeared unlikely but was possible: 5 had clotting, platelet or bleeding issues, 5 had skin issues and 3 were pyrexic. Amongst those consultations where CRGV would be within the differential diagnosis: 16 had acute renal failure, 5 were bleeding, in 4 of those cases the bleeding was into the urine (haematuria), and a further 3 were anaemic.

7.3.4 Evaluation of timeliness of consultation detection

Evaluating the training data set, the Ping algorithm identified consultations consistent with CGRV as early in the clinical course of the disease as the attending clinician, based on their clinical record. The small number of confirmed cases posed a challenge to evaluating the timeliness of case identification, with insufficient numbers to determine whether Ping outperformed the attending clinicians, however it appears to be similarly prompt (Figure 6.3.g).

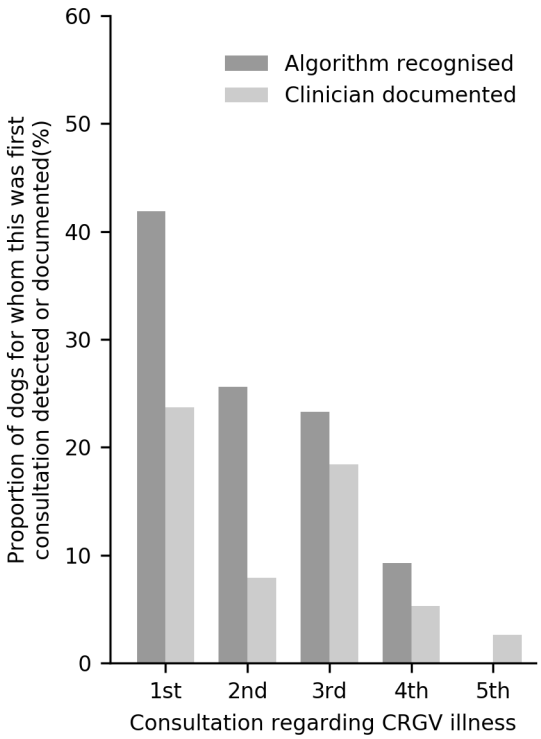

Figure 7.3.i: Efficacy of timely consultation detection

7.4 Discussion

The optimised prediction model identified a small group of dogs in the SAVSNET dataset that might have had CRGV; 58% of these dogs were considered false positives, and a further 17% were considered unlikely to be true positives. However, 25% of the consultations did bear sufficient hallmarks of this life-threatening illness to warrant considering CRGV a possibility. Given there have been 122 cases in a population of several million dogs, over a six year period, this apparently poor precision offers promise of a valuable surveillance and decision support tool.

At an individual level, prompt recognition of a potential disease process is of paramount importance since it permits the early implementation of supportive management. At the population level reliable and timely identification of cases offers scope to identify factors that may have contributed to the emergence of illness and the underlying aetiological agent. The ability for near real time surveillance directly from contemporaneaoulsy recorded clinical records would offer the potential to rapidly identify and monitor patterns of illness within the vet visiting population.

7.4.1 Study limitations

The major limitation of this work is the inability to make a firm diagnosis solely from the clinical narrative and the small number of confirmed cases of the illness. The main diagnostic mechanism for CRGV is histological; given these dogs have AKI, renal biopsy is generally only performed post mortem. Post mortem examinations may not be performed in first opinion practice and histopathology results for possible CRGV cases may therefore be unlikely to appear in the clinical record available via SAVSNET. It is not possible to formally quantify the predictive value of the tool as the gold standard test for CRGV is histological and is only conducted in suspect cases at post-mortem, and there are insufficient cases to divide the training set into a training and test set of a meaningful size.

One potential application of this system might be to provide diagnostic support to clinicians, in that situation multiple false positive advisories might be considered an annoyance. But given these numbers, the false positives would account for only a very small number of advisories to a clinic per year, and those that were clearly false positive to the author would be equally apparent to the

treating clinician. It is intended that, used in the appropriate context, the tool would assist first opinion clinicians in unifying the apparently disparate signs of skin lesions and acute renal impairment. It was clear from the narratives of confirmed cases that this was a particular challenge with many aetiological hypotheses documented.

Exploration of the phraseology responsible for generation of false positives and refinement of the algorithm to better recognise these provide the opportunity to improve specificity. In addition, refinement to the tool's ability to identify chronicity, both within an isolated consultation and across the history of an animal, offer promise to facilitate further improvement. This would require caution however because clinician and owner impressions at first presentation may attribute the skin lesions of CRGV to previous injury or dermatological condition or the signs of renal compromise to pre-existent systemic illness. A tool designed to ignore animals where there was also a chronic condition would by its nature not be able to detect the illness in these animals, who in turn may be equally susceptible to CRGV. This highlights the complex demands of the tool's discriminatory ability. The semantic clusters included within the current tool were intended to preferentially detect severe acute presentations.

7.4.2 Potential application to a real-world dataset

The text-mining tool described demonstrates the potential utility of this technique in addressing real-world challenges in veterinary public health and animal welfare. It has potential in three key areas: facilitating syndromic surveillance, risk factor identification and enabling clinical decision support.

For surveillance purposes or case identification for risk factor study, an additional regular expression designed to recognise the diagnostic phrases excluded from the models of Ping (CRGV, Alabama Rot, vasculopathy), could be combined with the multivariable model, this would generate a detection tool with slightly lower specificity but potentially improved sensitivity. It is important to note that the adjunct phrase would not form part of the multivariable model, as it would bias the model in favour of recognised cases, which was not the purpose of the tool. However, as a bolted-on search phrase, it would further optimise the ability to identify cases, and would be appropriate for detection of cases for risk factor analysis and identification of spatio-temporal trends. As awareness of CRGV rises it is likely that phrases such as CRGV will occur with increasing

frequency as a part of the clinician's differential diagnosis and safety netting advice, this may in turn render an adjunct search phrase obsolete.

An interim step in case identification, of a member of the research team reading the small number of consultations identified, would allow reliable identification of probable cases. This is feasible as the dataset requiring manual reading is reduced from in the region of 14,500 canine consultations each week to a mean of 40 consultations per week.

If the tool were used as a feedback mechanism, to alert clinicians that they may have seen a case of CRGV, in its current form it could be of assistance. It would be important that the tool was used without the bolted on regular expression as this would generate a positive feedback error in that a clinician documenting a suspicion of CRGV would receive a message that this may be CRGV. The algorithm alone would be sufficient to provide feedback based on the description of presenting signs; this avoids the risk of erroneous positive feedback and enables cases bearing the clinical features of CRGV where there was no clinician suspicion of CRGV to be identified. If feedback were combined with advice regarding management and direction towards clinical expertise in management, then this may prove a real asset in both understanding this illness and improving the outcome for affected animals.

The availability of a training dataset, where the presence of CRGV had been confirmed using the diagnostic gold standard, was crucial to the development of this tool. Natural language based surveillance for clinical signs is sometimes feasible without such a dataset, for example surveillance for the presence of diarrhoea or pyrexia, or indeed clusters of signs. However, surveillance at the disease level, especially for an illness that most clinicians will not have seen, is hindered without the availability of a dataset of records where the disease has been confirmed to the best of our ability.

7.4.3 Potential developments

Since this work was undertaken cases of CRGV have continued to be recognised in the UK. It would be feasible to use the more recently collated clinical records of confirmed cases to more formally validate the predictive abilities of Ping. This would be advantageous if it were to be deployed with feedback directly to clinicians.

7.5 Conclusion

It is possible to model the clinical language used to describe consultations for a rare syndrome and subsequently use that model to generate a tool capable of distinguishing between consultations relating to that syndrome and other reasons for consultation. The methods utilised here can be applied to other syndromes. This provides the opportunity for improved syndromic surveillance directly from the clinical narrative, greater understanding of disease by identification of hitherto undiagnosed cases via clinician descriptions of patterns of clinical signs, and where it is desired clinical decision support to clinicians.

Chapter Eight Discussion

8.1 Overview of findings

The work described in previous chapters demonstrates the need for and ability to generate indicators of information within the little explored veterinary clinical sublanguage of free-text clinical records. The techniques presented confer capacity to utilise the large volume of information within the narrative consultation record in the same manner as any other coded field in epidemiological studies and syndromic surveillance, with the advantage of responsive adaptability as to what is encoded. Development of these techniques required the groundwork outlined in Chapter Four and, in order to respectfully handle the data, the de-identification process whose development is described in Chapter Five.

There has been considerable previous work in the field of text-mining human clinical records (Y. Wang et al. 2018; Casey et al. 2016; Birkhead, Klompas, and Shah 2015; Meystre, Savova, Kipper-Schuler, and Hurdle 2008a), but little applied to veterinary clinical records (Dórea, Sanchez, and Revie 2011; Anholt et al. 2014; Lam et al. 2007). The exploration of the veterinary clinical sublanguage, outlined in Chapter four, illustrates the need for development of text mining tools tailored for this sublanguage. This work suggests that techniques designed for application in human healthcare records, and more so those designed for information extraction from grammatically correct standard English are likely not directly applicable to the veterinary clinical narrative, to do so would be akin to utilising an English language textbook to understand conversational Finnish. Similarly, the syntactic and lexical challenges of the veterinary clinical sublanguage and the need for preservation of research valuable data if electronic health records are to be used as a source of information regarding health and disease in the veterinary population, made the development of a de-identification system specifically tailored to the domain paramount, Clancularius serves this purpose well.

The impact of context on the semantics of words and phrases within any language cannot be underestimated (Requejo 2009; Song 2010),where the language is telegraphic and has limited grammatical structure familiarity with context is vital to comprehension by a human reader and equally critical to meaningful programmatic information extraction. The techniques encompassed by the context-sensitive classification framework described in Chapter six

demonstrate the importance of and ability to decipher both context and semantics even where the language is unmapped and syntactically challenging.

On occasion, there may be limited raw material, in the form of the free-text records of animals presented with a syndrome of interest, such atypia of the syndrome to compromise the use of its component signs as documented individually to develop a classifier indirectly, and an urgency to develop a means of case identification prior to the availability of such data. Chapter Seven demonstrates an alternative strategy, which was effective in identifying cases that may have represented the commonly fatal, emergent and poorly understood disease of cutaneous and renal glomerular vasculopathy in dogs.

8.2 Limitations

Any text-mining technique is only able to make use of information provided to it. The lack of documentation that was apparent in the evaluation of SAVSNET's clinician-assigned classification described in Chapter Three, and has been well described elsewhere in regard to human healthcare records (Steindal et al. 2012; Stevenson et al. 2014; De Marinis et al. 2010), will impair the ability of any text-mining dependent surveillance mechanism in identifying the events for which it is attuned. This is balanced however by the large volume of data available to a system extracting information from free-text records, in comparison to coded field or questionnaire responses, neither of which are immune to lack of or aberrant documentation.

The methodologies developed for both de-identification and syndromic surveillance required domain and sublanguage expertise, this was particularly crucial because of the previously limited work involving the veterinary clinical sublanguage. Real-world functional text-mining systems are likely to always require such expertise, either inherent to the developer or within a member of a development team with close communication between members. The capacity to recognise words in the absence of semantic understanding carries little value. The ability to understand and apply the basic principles of text-mining are perhaps more readily learned than the domain knowledge that is so key to their appropriate application. This is a finding supported by the work of Wilcox and Hripcsak (Wilcox and Hripcsak 2003) who, in analysing the effect of domain knowledge on performance and costs of free-text classifiers for clinical data,

found that expert knowledge was the most significant factor affecting classifier performance.

Language is not static and evolves in response to social and cultural drivers (Sapir 1921), as a consequence periodic audit and adaption should form an integral component of any surveillance system designed to extract meaning from clinical records documented in natural language. In addition to the natural linguistic drift of language, where the contributors to a clinical dataset are not a static population, as is likely even where the clinics contributing remain constant, new short hand notations and personal or regional colloquialisms are likely to be introduced. Responsive adaption to linguistic drift and increased entropy carry with them system maintenance and manpower demands, but are vital for the continued efficacy of a text-mining based system. The corollary of the ability and preparedness for adaption is the absence of any demand for uniformity in documentation on the clinicians contributing electronic health records to such a surveillance system.

8.3 Implications for research and syndromic surveillance

The development of veterinary clinical sublanguage specific de-identification software, Clancularius described in Chapter five, brings the ability to responsively generate a large volume of research and surveillance ready data whilst respecting the confidentiality of all parties and maintaining high ethical standards. Whilst Clancularius was tailored to the free-text records within the SAVSNET dataset, it can be applied to any UK small animal clinical records with minor adaptions to account for differences in data structure and language used. Indeed, the technique used in developing Clancularius could be readily adapted to any domain with rebuilding of the dictionaries and domain specific rules.

The ability to extract information regarding the presence of clinical signs within the vet-visiting small animal population renders responsive and adaptive near real-time syndromic surveillance from systems collating large volumes of small animal clinical data a reality, with the potential for considerable clinical and public health benefit.

The same text-mining techniques can also be applied for case identification in epidemiological studies. This has to a degree been used in previous work where text-search strategies without the finesse offered by context-sensitive

techniques have been applied, with the consequent need for considerable manual reading to exclude false positive identifications, whilst risking failure to identify cases documented using less common notation (Tulloch et al. 2017; P. H. Jones et al. 2014; Burke et al. 2017).

The ability to explore phraseology and context via the `regexConcordance()` method described in Chapter four, confers with it the potential to translate the outcomes of this work to other domains, most obviously the wider veterinary domain and human first opinion electronic health records.

Future work should involve the integration of context-sensitive text-mining into multi-source surveillance systems, to provide a reliable picture of spatio-temporal patterns of disease burden in the small-animal population.

8.4 Conclusion

Exploration of the veterinary clinical sublanguage has permitted the development of a process of de-identification in order to facilitate respectful use of veterinary consultation data for research purposes, and a validated series of clinical sign classifiers, with a framework that permits responsive generation of additional classifiers to meet evolving research demands.

The use of clinical signs to monitor the rate of attendance facilitates detection of an excess of presentations above the base-rate prior to clinical diagnosis, or even suspicion. The ability to monitor the rate of documentation of individual, or combinations of, clinical signs within large volumes of consultation records extends surveillance capability beyond known diseases and previously identified syndromes.

.

Bibliography

@AlabamaRot.CRGV.kills.dogs. 2018. “Alabama Rot - UK Mapping & Research.” *Facebook*. https://www.facebook.com/AlabamaRot.CRGV.kills.dogs/.

A M Turing. 1950. “I.—Computing Machinery and Intelligence.” *Mind a Quaterly Review of Psychology and Philosophy* LIX (236). Oxford University Press: 433–60. doi:10.1093/mind/LIX.236.433.

Al-Awqati, Qais. 2006. “How to Write a Case Report: Lessons From 1600 B.C..” *Kidney International* 69 (12): 2113–14. doi:10.1038/sj.ki.5001592.

Alabama Rot Research Fund. 2018. “StopAlabamaRot.” http://www.arrf.co.uk/.

Algeo, John, Robert K Barnhart, and Sol Steinmetz. 1989. “The Barnhart Dictionary of Etymology.” *Language* 65 (4): 848.

Allen, Chris, Ming-Hsiang Tsou, Anoshe Aslam, Anna Nagel, and Jean-Mark Gawron. 2016. “Applying GIS and Machine Learning Methods to Twitter Data for Multiscale Surveillance of Influenza..” Edited by Mansour Ebrahimi. *PLoS ONE* 11 (7): e0157734. doi:10.1371/journal.pone.0157734.

Allen, J F. 1983. “Maintaining Knowledge About Temporal Intervals.” *Commun. ACM* 26 (11): 832–43. doi:10.1145/182.358434.

Alton, Gillian D, David L Pearl, Ken G Bateman, W Bruce McNab, and Olaf Berke. 2010. “Factors Associated with Whole Carcass Condemnation Rates in Provincially-Inspected Abattoirs in Ontario 2001-2007: Implications for Food Animal Syndromic Surveillance..” *BMC Veterinary Research* 6 (1). BioMed Central: 42. doi:10.1186/1746-6148-6-42.

Anholt, R M, J Berezowski, C Robertson, and C Stephen. 2015. “Spatial-Temporal Clustering of Companion Animal Enteric Syndrome: Detection and Investigation Through the Use of Electronic Medical Records From Participating Private Practices..” *Epidemiology and Infection* 143 (12). Cambridge University Press: 2547–58. doi:10.1017/S0950268814003574.

Anholt, R M, J Berezowski, I Jamal, C Ribble, and C Stephen. 2014. “Mining Free-Text Medical Records for Companion Animal Enteric Syndrome Surveillance.” *Preventive Veterinary Medicine* 113 (4): 417–22.

Ansaldi, F, A Orsi, F Altomonte, G Bertone, V Parodi, R Carloni, P Moscatelli, E Pasero, P Oreste, and G Icardi. 2008. “Emergency Department Syndromic Surveillance System for Early Detection of 5 Syndromes: a Pilot Project in a Reference Teaching Hospital in Genoa, Italy..” *J Prev Med Hyg* 49 (4): 131–35.

APHA. 2018. “Animal and Plant Health Agency.” *Gov.Uk*. https://www.gov.uk/government/organisations/animal-and-plant-health-agency/.

Appelt, D, and D Israel. 1999. “Introduction to Information Extraction Technology.” In. Stockholm.

Arcaya, Mariana C, Alyssa L Arcaya, and S V Subramanian. 2015. “Inequalities in Health: Definitions, Concepts, and Theories..” *Global Health Action* 8 (1). Taylor & Francis: 27106. doi:10.3402/gha.v8.27106.

Armbruster, Bonnie B, Thomas H Anderson, and Joyce Ostertag. 1987. “Does

Text Structure/Summarization Instruction Facilitate Learning From Expository Text?." *Reading Research Quarterly* 22 (3): 331. doi:10.2307/747972.

Aronsky, D, D Kendall, K Merkley, B C James, and P J Haug. 2001. "A Comprehensive Set of Coded Chief Complaints for the Emergency Department.." *Acad Emerg Med* 8 (10): 980–89.

Asher, Lucy, Emma L Buckland, C Ianthi Phylactopoulos, Martin C Whiting, Siobhan M Abeyesinghe, and Christopher M Wathes. 2011. "Estimation of the Number and Demographics of Companion Dogs in the UK.." *BMC Veterinary Research* 7 (1). BioMed Central: 74. doi:10.1186/1746-6148-7-74.

Assaf, A R, K L Lapane, J L McKenney, and R A Carleton. 1993. "Possible Influence of the Prospective Payment System on the Assignment of Discharge Diagnoses for Coronary Heart Disease.." *The New England Journal of Medicine* 329 (13). Massachusetts Medical Society: 931–35. doi:10.1056/NEJM199309233291307.

Assareh, Hassan, Helen M Achat, Joanne M Stubbs, Veth M Guevarra, and Kim Hill. 2016. "Incidence and Variation of Discrepancies in Recording Chronic Conditions in Australian Hospital Administrative Data.." Edited by Chiara Lazzeri. *PLoS ONE* 11 (1). Public Library of Science: e0147087. doi:10.1371/journal.pone.0147087.

Augusto, Juan Carlos. 2005. "Temporal Reasoning for Decision Support in Medicine." *Artif Intell Med* 33 (1): 1–24. doi:10.1016/j.artmed.2004.07.006.

Baer, H J, I Cho, R A Walmer, P A Bain, and D W Bates. 2013. "Using Electronic Health Records to Address Overweight and Obesity: a Systematic Review." *American Journal of Preventive Medicine* 45 (4): 494–500.

Baker, Maureen, Gillian E Smith, Duncan Cooper, Neville Q Verlander, Frances Chinemana, Sarafina Cotterill, Vivien Hollyoak, and Rod Griffiths. 2003. "Early Warning and NHS Direct: a Role in Community Surveillance?." *Journal of Public Health Medicine* 25 (4): 362–68.

Beckwith, Bruce A, Rajeshwarri Mahaadevan, Ulysses J Balis, and Frank Kuo. 2006. "Development and Evaluation of an Open Source Software Tool for Deidentification of Pathology Reports." *BMC Medical Informatics and Decision Making* 6 (1). London: BioMed Central: 12–12. doi:10.1186/1472-6947-6-12.

Berman, Jules J. 2003. "Concept-Match Medical Data Scrubbing." *Arch Pathol Lab Med* 127 (6). College of American Pathologists: 680–86. doi:10.1043/1543-2165(2003)127<680:CMDS>2.0.CO;2.

Biber, Douglas. 1988. *Variation Across Speech and Writing*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511621024.

Birkhead, Guthrie S, Michael Klompas, and Nirav R Shah. 2015. "Uses of Electronic Health Records for Public Health Surveillance to Advance Public Health.." *Annual Review of Public Health* 36 (1). Annual Reviews: 345–59. doi:10.1146/annurev-publhealth-031914-122747.

BNC Consortium. 2007. “British National Corpus.” Distributed by Oxford

University Computing Services on behalf of the BNC Consortium. www.natcorp.ox.ac.uk.

Bordonaro, Samantha F, Daniel C McGillicuddy, Francesco Pompei, Dmitriy Burmistrov, Charles Harding, and Leon D Sanchez. 2016. “Human Temperatures for Syndromic Surveillance in the Emergency Department: Data From the Autumn Wave of the 2009 Swine Flu (H1N1) Pandemic and a Seasonal Influenza Outbreak..” *BMC Emergency Medicine* 16 (1). BioMed Central: 16. doi:10.1186/s12873-016-0080-7.

Botsis, T, E J Woo, and R Ball. 2013. “The Contribution of the Vaccine Adverse Event Text Mining System to the Classification of Possible Guillain-Barre Syndrome Reports.” *Appl Clin Inform* 4 (1): 88–99. doi:10.4338/aci-2012-11-ra-0049.

Botsis, T, M D Nguyen, E J Woo, M Markatou, and R Ball. 2011. “Text Mining for the Vaccine Adverse Event Reporting System: Medical Text Classification Using Informative Feature Selection.” *J Am Med Inform Assoc* 18 (5): 631–38. doi:10.1136/amiajnl-2010-000022.

Botsis, T, T Buttolph, M D Nguyen, S Winiecki, E J Woo, and R Ball. 2012. “Vaccine Adverse Event Text Mining System for Extracting Features From Vaccine Safety Reports.” *J Am Med Inform Assoc* 19 (6): 1011–18. doi:10.1136/amiajnl-2012-000881.

Brabazon, E D, M W Carton, C Murray, L Hederman, and D Bedford. 2010. “General Practice Out-of-Hours Service in Ireland Provides a New Source of Syndromic Surveillance Data on Influenza..” *Euro Surveillance : Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin*, January.

British Broadcasting Corporation. 2018. “BBC News.” http://www.bbc.co.uk/news.

Brodbelt, David. 2012. “VeNom Coding.” In. Birmingham. https://www.vin.com/apputil/content/defaultadv1.aspx?id=5328356&pid=11349&.

Brown, L D, T T Cai, and A DasGupta. 2001. “Interval Estimation for a Binomial Proportion.” *Statistical Science* 16 (2): 101–33.

Brown, Phillip, Sylvia Halasz, Colin Goodall, Dennis G Cochrane, Peter Milano, and John R Allegra. 2010. “The Ngram Chief Complaint Classifier: a Novel Method of Automatically Creating Chief Complaint Classifiers Based on International Classification of Diseases Groupings.” *Journal of Biomedical Informatics* 43 (2): 268–72. doi:10.1016/j.jbi.2009.08.015.

Bruns, Brian, James K Lowden, and Frediano Ziglio. 2016. “FreeTDS.” *FreeTDS.org*. http://www.freetds.org/.

Buckland, E L, D O'Neill, J Summers, A Mateus, D Church, L Redmond, and D Brodbelt. 2016. “Characterisation of Antimicrobial Usage in Cats and Dogs Attending UK Primary Care Companion Animal Veterinary Practices..” *The Veterinary Record* 179 (19). British Medical Journal Publishing Group: 489–89. doi:10.1136/vr.103830.

Burke, Sara, Vicki Black, Fernando Sánchez-Vizcaíno, Alan Radford, Angie Hibbert, and Séverine Tasker. 2017. “Use of Cefovecin in a UK Population of Cats Attending First-Opinion Practices as Recorded in Electronic Health Records..” *Journal of Feline Medicine and Surgery* 19 (6): 687–92. doi:10.1177/1098612X16656706.

Caliendo, Angela M, David N Gilbert, Christine C Ginocchio, Kimberly E Hanson, Larissa May, Thomas C Quinn, Fred C Tenover, et al. 2013. “Better Tests, Better Care: Improved Diagnostics for Infectious Diseases.” *Clinical Infectious Diseases* 57 (suppl 3): S139–70. doi:10.1093/cid/cit578.

Campbell, D A, and S B Johnson. 2001. “Comparing Syntactic Complexity in Medical and Non-Medical Corpora..” *Proc AMIA Symp*, 90–94.

Campbell, David A, and Stephen B Johnson. 2002. “A Transformational-Based Learner for Dependency Grammars in Discharge Summaries.” In, 3:37–44. Phildadelphia, Pennsylvania: Association for Computational Linguistics. doi:10.3115/1118149.1118155.

Campbell, James R, Paul Carpenter, Charles Sneiderman, Simon Cohn, Christopher G Chute, and Judith Warren. 1997. “Phase II Evaluation of Clinical Coding SchemesCompleteness, Taxonomy, Mapping, Definitions, and Clarity.” *J Am Med Inform Assoc* 4 (3). Oxford University Press: 238–51. doi:10.1136/jamia.1997.0040238.

Carpenter, J L, N C Andelman, F M Moore, and Jr N W King. 1988. “Idiopathic Cutaneous and Renal Glomerular Vasculopathy of Greyhounds.” *Veterinary Pathology* 25 (6). SAGE PublicationsSage CA: Los Angeles, CA: 401–7. doi:10.1177/030098588802500601.

Carrico, Ruth, and Linda Goss. 2005. “Syndromic Surveillance: Hospital Emergency Department Participation During the Kentucky Derby Festival..” *Disaster Management & Response : DMR : an Official Publication of the Emergency Nurses Association* 3 (3): 73–79. doi:10.1016/j.dmr.2005.04.003.

Casey, Joan A, Brian S Schwartz, Walter F Stewart, and Nancy E Adler. 2016. “Using Electronic Health Records for Population Health Research: a Review of Methods and Applications.” *Dx.Doi.org* 37 (1). Annual Reviews: 61–81. doi:10.1146/annurev-publhealth-032315-021353.

CDC Evaluation Working Group on Public Health Surveillance Systems For Early Detection of Outbreaks. 2004. “Framework for Evaluating Public Health Surveillance Systems for Early Detection of Outbreaks.” *Mmwr* 53 (5).

Celsus, A C. 25AD. *De Medicina*. Self published.

Chapman, W. 2006. “Natural Language Processing for Biosurveillance.” In *Handbook of Biosurveillance*, edited by A W Moore, M M Wagner, and R M Aryel, 255–71. New York: Elsevier.

Chapman, W W, J N Dowling, and M M Wagner. 2005. “Classification of Emergency Department Chief Complaints Into 7 Syndromes: a Retrospective Analysis of 527,228 Patients.” *Ann Emerg Med* 46 (5): 445–55. doi:10.1016/j.annemergmed.2005.04.012.

Chapman, W W, L M Christensen, M M Wagner, P J Haug, O Ivanov, J N Dowling, and R T Olszewski. 2005. “Classifying Free-Text Triage Chief Complaints Into Syndromic Categories with Natural Language Processing.” *Artif Intell Med* 33 (1): 31–40. doi:10.1016/j.artmed.2004.04.001.

Chapman, W W, W Bridewell, P Hanbury, G F Cooper, and B G Buchanan. 2001. “A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries.” *Journal of Biomedical Informatics* 34 (5): 301–10. doi:10.1006/jbin.2001.1029.

Cheng, K, T Baldwin, and K Verspoor. 2017. “Automatic Negation and Speculation Detection in Veterinary Clinical Text.” In, 70–78. Brisbane, Australia.

Cheng, Ping, Annette Gilchrist, Kerin M Robinson, and Lindsay Paul. 2009. “The Risk and Consequences of Clinical Miscoding Due to Inadequate Medical Documentation: a Case Study of the Impact on Health Services Funding:.” *Health Information Management Journal* 38 (1). SAGE PublicationsSage UK: London, England: 35–46. doi:10.1177/183335830903800105.

Chute, C G, S P Cohn, K E Campbell, D E Oliver, J R Campbell, Computer-Based Patient Record Institute's Work Group on Codes & Structures. 1996. “The Content Coverage of Clinical Classifications.” *Journal of the American Medical Informatics Association* 3 (3): 224–33. doi:10.1136/jamia.1996.96310636.

Coden, Anni R, Serguei V Pakhomov, Rie K Ando, Patrick H Duffy, and Christopher G Chute. 2005. “Domain-Specific Language Models and Lexicons for Tagging.” *Journal of Biomedical Informatics* 38 (6): 422–30. doi:10.1016/j.jbi.2005.02.009.

Coe, Jason B, Cindy L Adams, and Brenda N Bonnett. 2008. “A Focus Group Study of Veterinarians‘ and Pet Owners’ Perceptions of Veterinarian-Client Communication in Companion Animal Practice..” *Journal of the American Veterinary Medical Association* 233 (7). American Veterinary Medical Association 1931 North Meacham Road - Suite 100, Schaumburg, IL 60173 USA 847-925-8070 847-925-1329 avmajournals@avma.org: 1072–80. doi:10.2460/javma.233.7.1072.

Cohen, Jacob. 1960. “A Coefficient of Agreement for Nominal Scales.” *Educational and Psychological Measurement* 20 (1). SAGE Publications Inc: 37–46. doi:10.1177/001316446002000104.

Cohen, M L. 1998. “Resurgent and Emergent Disease in a Changing World.” *British Medical Bulletin* 54 (3): 523–32. doi:10.1093/oxfordjournals.bmb.a011707.

Collier, Nigel, Nguyen Truong Son, and Ngoc Mai Nguyen. 2011. “OMG U Got Flu? Analysis of Shared Health Messages for Bio-Surveillance..” *Journal of Biomedical Semantics* 2 Suppl 5 (Suppl 5). BioMed Central: S9. doi:10.1186/2041-1480-2-S5-S9.

Combi, Carlo, and Yuval Shahar. 1997. “Temporal Reasoning and Temporal Data Maintenance in Medicine: Issues and Challenges.” *Computers in Biology and Medicine* 27 (5): 353–68. doi:10.1016/S0010-4825(96)00010-8.

Conway, M, J N Dowling, and W W Chapman. 2013. “Using Chief Complaints

for Syndromic Surveillance: a Review of Chief Complaint Based Classifiers in North America." *Journal of Biomedical Informatics* 46 (4): 734–43.

Coory, M D, H Kelly, and V Tippett. 2009. "Assessment of Ambulance Dispatch Data for Surveillance of Influenza-Like Illness in Melbourne, Australia." *Public Health* 123 (2). W.B. Saunders: 163–68. doi:10.1016/j.puhe.2008.10.027.

Correia-Gomes, Carla, Richard P Smith, Jude I Eze, Madeleine K Henry, George J Gunn, Susanna Williamson, and Sue C Tongue. 2016. "Pig Abattoir Inspection Data: Can It Be Used for Surveillance Purposes?." Edited by Petr Heneberg. *PLoS ONE* 11 (8). Public Library of Science: e0161990. doi:10.1371/journal.pone.0161990.

Cowan, L A, D M Hertzke, B W Fenwick, and C B Andreasen. 1997. "Clinical and Clinicopathologic Abnormalities in Greyhounds with Cutaneous and Renal Glomerular Vasculopathy: 18 Cases (1992-1994)." *Journal of the American Veterinary Medical Association* 210 (6): 789–93.

Dale, E, and J S Chall. 1948. "A Formula for Predicting Readability: Instructions." *Educational Research Bulletin* 27 (2): 37–54. doi:10.2307/1473669.

Damerau, Fred J. 1964. "A Technique for Computer Detection and Correction of Spelling Errors." *Commun. ACM* 7 (3). ACM: 171–76. doi:10.1145/363958.363994.

Danan, C, T Baroukh, F Moury, N Jourdan-Da Silva, A Brisabois, and Y Le Strat. 2010. "Automated Early Warning System for the Surveillance of Salmonella Isolated in the Agro-Food Chain in France." *Epidemiology and Infection* 139 (05): 736–41. doi:10.1017/S0950268810001469.

Day, S, L M Christensen, J Dalto, and P Haug. 2007. "Identification of Trauma Patients at a Level 1 Trauma Center Utilizing Natural Language Processing." *J Trauma Nurs* 14 (2): 79–83. doi:10.1097/01.jtn.0000278792.20913.82.

De Marinis, Maria Grazia, Michela Piredda, Maria Chiara Pascarella, Bruno Vincenzi, Fiorenza Spiga, Daniela Tartaglini, Rosaria Alvaro, and Maria Matarese. 2010. "'If It Is Not Recorded, It Has Not Been Done!'? Consistency Between Nursing Records and Observed Nursing Care in an Italian Hospital." *Journal of Clinical Nursing* 19 (11-12). Wiley/Blackwell (10.1111): 1544–52. doi:10.1111/j.1365-2702.2009.03012.x.

Deiner, Michael S, Thomas M Lietman, Stephen D McLeod, James Chodosh, and Travis C Porco. 2016. "Surveillance Tools Emerging From Search Engines and Social Media Data for Determining Eye Disease Patterns.." *JAMA Ophthalmology* 134 (9). American Medical Association: 1024–30. doi:10.1001/jamaophthalmol.2016.2267.

Delbecque, Thierry, and Pierre Zweigenbaum. 2007. "Artificial Intelligence in Medicine." In *Artificial Intelligence in Medicine*, edited by Riccardo Bellazzi, Ameen Abu-Hanna, and Jim Hunter, 4594:242–46. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-540-73599-1_32.

Deleger, Louise, Katalin Molnar, Guergana Savova, Fei Xia, Todd Lingren, Qi Li,

Keith Marsolo, et al. 2013. “Large-Scale Evaluation of Automated Clinical Note De-Identification and Its Impact on Information Extraction..” *J Am Med Inform Assoc* 20 (1): 84–94. doi:10.1136/amiajnl-2012-001012.

Dick, R S, and E B Steen. 1991. “The Computer-Based Patient Record: an Essential Technology for Health Care.” Committee on Improving the Patient Record, Institute of Medicine, National Academy Press.

Diesel, G, D Pfeiffer, S Crispin, and D Brodbelt. 2010. “Risk Factors for Tail Injuries in Dogs in Great Britain..” *The Veterinary Record* 166 (26): 812–17. doi:10.1136/vr.b4880.

Ding, Haibo, and Ellen Riloff. 2015. “Extracting Information About Medication Use From Veterinary Discussions.” In, 1452–58. Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/v1/N15-1168.

Dinh, Michael M, Christopher Kastelein, Kendall J Bein, Timothy C Green, Tanya Bautovich, and Rebecca Ivers. 2015. “Use of a Syndromic Surveillance System to Describe the Trend in Cycling-Related Presentations to Emergency Departments in Sydney..” *Emergency Medicine Australasia : EMA* 27 (4): 343–47. doi:10.1111/1742-6723.12422.

Dobson, R. 2007. “NHS Direct Could Be a Useful Early Warning System for Respiratory Infections.” *Bmj* 334 (7608): 1343–43. doi:10.1136/bmj.39257.585405.BE.

Dorr, D A, W F Phillips, S Phansalkar, S A Sims, and J F Hurdle. 2006. “Assessing the Difficulty and Time Cost of De-Identification in Clinical Narratives..” *IMIA Yearbook 2008: Access to Health Information* 45 (3): 246–52.

Dórea, F C, A Lindberg, B J McEwen, C W Revie, and J Sanchez. 2014. “Syndromic Surveillance Using Laboratory Test Requests: a Practical Guide Informed by Experience with Two Systems..” *Preventive Veterinary Medicine* 116 (3): 313–24. doi:10.1016/j.prevetmed.2014.04.001.

Dórea, Fernanda C, and Flavie Vial. 2016. “Animal Health Syndromic Surveillance: a Systematic Literature Review of the Progress in the Last 5 Years (2011&Ndash;2016).” *Veterinary Medicine: Research and Reports* Volume 7. Dove Press: 157–70. doi:10.2147/VMRR.S90182.

Dórea, Fernanda C, Javier Sanchez, and Crawford W Revie. 2011. “Veterinary Syndromic Surveillance: Current Initiatives and Potential for Development..” *Preventive Veterinary Medicine* 101 (1-2): 1–17. doi:10.1016/j.prevetmed.2011.05.004.

Dupuy, Céline, Anne Bronner, Eamon Watson, Linda Wuyckhuise-Sjouke, Martin Reist, Anne Fouillet, Didier Calavas, Pascal Hendrikx, and Jean-Baptiste Perrin. 2013. “Inventory of Veterinary Syndromic Surveillance Initiatives in Europe (Triple-S Project): Current Situation and Perspectives..” *Preventive Veterinary Medicine* 111 (3-4): 220–29. doi:10.1016/j.prevetmed.2013.06.005.

Eastwood, K, D N Durrheim, K Main, D Muscatello, W Zheng, T Merritt, K Todd, and K Hope. 2008. “The Public Health Value of Emergency Department Syndromic Surveillance Following a Natural Disaster.” *Communicable Diseases Intelligence Quarterly Report* 32 (1). Department of Health and

Ageing: 92.

Elliot, Alex J, Helen E Hughes, Thomas C Hughes, Thomas E Locker, Tony Shannon, John Heyworth, Andy Wapling, et al. 2012. “Establishing an Emergency Department Syndromic Surveillance System to Support the London 2012 Olympic and Paralympic Games..” *Emergency Medicine Journal : EMJ* 29 (12). BMJ Publishing Group Ltd and the British Association for Accident & Emergency Medicine: 954–60. doi:10.1136/emermed-2011-200684.

European Parliament and of the Council. 2000. *Regulation (EC) No 1760/2000 of the European Parliament and of the Council of 17 July 2000 Establishing a System for the Identification and Registration of Bovine Animals and Regarding the Labelling of Beef and Beef Products and Repealing Council Regulation (EC) No 820/97 (OJ L 204, 11.8.2000, Pp. 1–10)*. http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32000R1760.

European Union. 2016. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC*. *Official Journal of the European Union*. Vol. 119.

Everitt, S, A Pilnick, J Waring, and M Cobb. 2013. “The Structure of the Small Animal Consultation..” *The Journal of Small Animal Practice* 54 (9). Blackwell Publishing Ltd: 453–58. doi:10.1111/jsap.12115.

Evidence-Based Medicine Working Group. 1992. “Evidence-Based Medicine. a New Approach to Teaching the Practice of Medicine..” *Jama* 268 (17): 2420–25.

Eylenbosch, W J, and N D Noah. 1988. “Historical Aspects.” In *Surveillance in Health Disease*, edited by W J Eylenbosch and N D Noah, 166–82.

Fielstein, E M, S H Brown, and T Speroff. 2004. “Algorithmic De-Identification of VA Medical Exam Text for HIPAA Privacy Compliance: Preliminary Findings.” *Medinfo*.

Fischer, M, and G Ereaut. 2012. “When Doctors and Patients Talk.” The Health Foundation. http://www.health.org.uk/sites/health/files/WhenDoctorsAndPatientsTalkMakingSenseOfTheConsultation.pdf.

Flamand, C, S Larrieu, F Couvy, B Jouves, L Josseran, and L Filleul. 2008. “Validation of a Syndromic Surveillance System Using a General Practitioner House Calls Network, Bordeaux, France.” *Euro Surveillance : Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin* 13 (25). European Centre for Disease Prevention and Control: 18905. doi:10.2807/ese.13.25.18905-en.

Flesch, R. 1948. “A New Readability Yardstick..” *The Journal of Applied Psychology* 32 (3): 221–33.

Franz, P, A Zaiss, S Schulz, U Hahn, and R Klar. 2000. “Automated Coding of Diagnoses--Three Methods Compared.” *Proceedings / AMIA . Annual Symposium. AMIA Symposium*, 250–54.

Friedlin, F J, and C J McDonald. 2008. “A Software Tool for Removing Patient Identifying Information From Clinical Documents.” *J Am Med Inform Assoc* 15 (5). American Medical Informatics Association: 601–10.

Friedman, C. 2012. “MedLEE.” *MedLingMap*. http://www.medlingmap.org/taxonomy/term/80.

Friedman, C, P O Alderson, J H Austin, J J Cimino, and S B Johnson. 1994. “A General Natural-Language Text Processor for Clinical Radiology..” *J Am Med Inform Assoc* 1 (2). American Medical Informatics Association: 161–74.

Friedman, Carol, Pauline Kra, and Andrey Rzhetsky. 2002. “Two Biomedical Sublanguages: a Description Based on the Theories of Zellig Harris.” *Journal of Biomedical Informatics* 35 (4): 222–35.

Furrer, L, S Küker, J Berezowski, H Posthaus, and F Vial TIA. 2015. “Constructing a Syndromic Terminology Resource for Veterinary Text Mining..” In. Granada, Spain. http://ceur-ws.org/Vol-1495/paper_8.pdff.

Gaizauskas, R, H Harkema, M Hepple, and A Setzer. 2006. “Task-Oriented Extraction of Temporal Information: the Case of Clinical Narratives.” In, 2006:188–95.

Gal, Tamas S, Thomas C Tucker, Aryya Gangopadhyay, and Zhiyuan Chen. 2014. “A Data Recipient Centered De-Identification Method to Retain Statistical Attributes.” *Journal of Biomedical Informatics* 50 (8//): 32–45. doi:10.1016/j.jbi.2014.01.001.

Galbraith, N S. 1992. “Communicable Disease Surveillance.” In *Recent Advances in Community Medicine*, edited by A Smith, 2nd ed., 127–42.

Gardner, James, and Li Xiong. 2008. “HIDE: an Integrated System for Health Information DE-Identification.” In, 254–59. IEEE. doi:10.1109/CBMS.2008.129.

General Medical Council. 2013. *Good Medical Practice*. https://www.gmc-uk.org/guidance/good_medical_practice/.

Gerbier, S, O Yarovaya, Q Gicquel, A L Millet, V Smaldore, V Pagliaroli, S Darmoni, and M H Metzger. 2011. “Evaluation of Natural Language Processing From Emergency Department Computerized Medical Records for Intra-Hospital Syndromic Surveillance.” *BMC Medical Informatics and Decision Making* 11 (1): 50. doi:10.1186/1472-6947-11-50.

Gilbert, J A. 1998. “Physician Data Entry: Providing Options Is Essential.” *Health Data Management*.

Goldman-Mellor, Sidra, Yusheng Jia, Kevin Kwan, and Jared Rutledge. 2017. “Syndromic Surveillance of Mental and Substance Use Disorders: a Validation Study Using Emergency Department Chief Complaints.” *Psychiatric Services*, September. American Psychiatric AssociationArlington, VA, appi.ps.2017000. doi:10.1176/appi.ps.201700028.

Graunt, John, and William Petty. 1662. *Natural and Political Observations Mentioned in a Following Index, and Made Upon the Bills of Mortality*. Thomas Roycroft, London.

Greene, S K, J Huang, A M Abrams, D Gilliss, M Reed, R Platt, S S Huang, and

M Kulldorff. 2012. “Gastrointestinal Disease Outbreak Detection Using Multiple Data Streams From Electronic Medical Records.” *Foodborne Pathog Dis* 9 (5): 431–41. doi:10.1089/fpd.2011.1036.

Greenhalgh, Trisha. 1999. *Narrative Based Medicine in an Evidence Based World*. *Bmj*. Vol. 318. doi:10.1136/bmj.318.7179.323.

Gregory, Michael. 1967. “Aspects of Varieties Differentiation.” *Journal of Linguistics* 3 (2). Cambridge University Press: 177–98. doi:10.1017/S0022226700016601.

Grishman, R, and B Sundheim. 1996. “Message Understanding Conference-6: a Brief History.” In, 466–71. Copenhagen.

Grishman, R, and L Hirschman. 1978. “Question Answering From Natural Language Medical Data Bases.” *Artificial Intelligence* 11 (1-2): 25–43.

Gunning, R. 1952. *The Technique of Clear Writing*. (1952). doi:10.1234/12345678.

Gupta, Dilip, Melissa Saul, and John Gilbertson. 2004. “Evaluation of a Deidentification (De-Id) Software Engine to Share Pathology Reports and Clinical Documents for Research..” *American Journal of Clinical Pathology* 121 (2): 176–86. doi:10.1309/E6K3-3GBP-E5C2-7FYU.

Hahn, Udo, and Joachim Wermter. 2004. “High-Performance Tagging on Medical Texts.” In, 973–es. Geneva, Switzerland: Association for Computational Linguistics. doi:10.3115/1220355.1220495.

Hall, P A, and N R Lemoine. 1986. “Comparison of Manual Data Coding Errors in Two Hospitals.” *Journal of Clinical Pathology* 39 (6): 622–26. doi:10.1136/jcp.39.6.622.

Haroon, S M M, G P Barbosa, and P J Saunders. 2011. “The Determinants of Health-Seeking Behaviour During the a/H1N1 Influenza Pandemic: an Ecological Study.” *Journal of Public Health* 33 (4): 503–10. doi:10.1093/pubmed/fdr029.

Harris, Zellig. 1991. *Theory of Language and Information: a Mathematical Approach*.

Harris, Zellig S. 1981. “Co-Occurrence and Transformation in Linguistic Structure.” In *Papers on Syntax*, 143–210. Dordrecht: Springer Netherlands. doi:10.1007/978-94-009-8467-7_8.

Hazlehurst, B, A Naleway, and J Mullooly. 2009. “Detecting Possible Vaccine Adverse Events in Clinical Notes of the Electronic Medical Record.” *Vaccine* 27 (14): 2077–83. doi:10.1016/j.vaccine.2009.01.105.

Hazlehurst, Brian, John Mullooly, Allison Naleway, and Brad Crane. 2005. “Detecting Possible Vaccination Reactions in Clinical Notes..” *AMIA Annu Symp Proc*, 306–10.

Health Social Care Information Centre. 2011. “Read Codes.” *Systems.Hscic.Gov.Uk*. http://systems.hscic.gov.uk/data/uktc/readcodes.

Hearst, Marti A. 1999. “Untangling Text Data Mining.” *The 37th Annual Meeting of the Association for Computational Linguistics*. College Park, Maryland:

Association for Computational Linguistics, 3–10. doi:10.3115/1034678.1034679.

Hersh, W R, and E M Campbell. 1997. “Assessing the Feasibility of Large-Scale Natural Language Processing in a Corpus of Ordinary Medical Records: a Lexical Analysis..” *Proceedings of the AMIA …*.

Hertzke, D M, L A Cowan, P Schoning, and B W Fenwick. 1995. “Glomerular Ultrastructural Lesions of Idiopathic Cutaneous and Renal Glomerular Vasculopathy of Greyhounds..” *Veterinary Pathology* 32 (5). SAGE PublicationsSage CA: Los Angeles, CA: 451–59. doi:10.1177/030098589503200501.

Hess, V. 2010. “… Beobachtung. Die Genese Der Modernen Krankenakte Am Beispiel Der Berliner Und Pariser Medizin (1725-1830)/Formalizing Observation: the Emergence of the ….” *Medizinhistoriches Journal*.

Hirschman, L, G Story, E Marsh, M Lyman, and N Sager. 1981. “An Experiment in Automated Health Care Evaluation From Narrative Medical Records.” *Computers and Biomedical Research* 14 (5): 447–63.

Hobbs, J R. 2002. “Information Extraction From Biomedical Text.” *Journal of Biomedical Informatics* 35 (4): 260–64.

Hoffman, Jessica M, Dan G O'Neill, Kate E Creevy, and Steven N Austad. 2017. “Do Female Dogs Age Differently Than Male Dogs?.” *The Journals of Gerontology. Series a, Biological Sciences and Medical Sciences*, May. doi:10.1093/gerona/glx061.

Holm, L P, I Hawkins, C Robin, R J Newton, R Jepson, G Stanzani, L A McMahon, et al. 2015. “Cutaneous and Renal Glomerular Vasculopathy as a Cause of Acute Kidney Injury in Dogs in the UK..” *The Veterinary Record* 176 (15). British Medical Journal Publishing Group: 384–84. doi:10.1136/vr.102892.

Holzinger, Andreas, Johannes Schantl, Miriam Schroettner, Christin Seifert, and Karin Verspoor. 2014. “Biomedical Text Mining: State-of-the-Art, Open Problems and Future Challenges.” In *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*, 8401:271–300. Interactive Knowledge Discovery and Data Mining in Biomedical Informatics: State-of-the-Art and Future Challenges. Berlin, Heidelberg: Springer, Berlin, Heidelberg. doi:10.1007/978-3-662-43968-5_16.

House of Representatives 111th Congress, House of Representatives 111th. 2009. *American Recovery and Reinvestment Act of 2009*. 111.

Hripcsak, G, L Zhou, S Parsons, A K Das, and S B Johnson. 2005. “Modeling Electronic Discharge Summaries as a Simple Temporal Constraint Satisfaction Problem.” *Journal of the American Medical Informatics Association* 12 (1): 55–63.

Hsu, Joy, Jennifer A Pacheco, Whitney W Stevens, Maureen E Smith, and Pedro C Avila. 2014. “Accuracy of Phenotyping Chronic Rhinosinusitis in the Electronic Health Record..” *American Journal of Rhinology & Allergy* 28 (2): 140–44. doi:10.2500/ajra.2014.28.4012.

Huang, Lu-Chou, Huei-Chung Chu, Chung-Yueh Lien, Chia-Hung Hsiao, and

Tsair Kao. 2010. “Embedding a Hiding Function in a Portable Electronic Health Record for Privacy Preservation.” *Journal of Medical Systems* 34 (3). Springer US: 313–20. doi:10.1007/s10916-008-9243-8.

Huang, Lu-Chou, Huei-Chung Chu, Chung-Yueh Lien, Chia-Hung Hsiao, and Tsair Kao. 2009. “Privacy Preservation and Information Security Protection for Patients’ Portable Electronic Health Records.” *Computers in Biology and Medicine* 39 (9): 743–50. doi:10.1016/j.compbiomed.2009.06.004.

Hunter, John, Darren Dale, Eric Firing, Michael Droettboom, Matplotlib development team. 2018. “Matplotlib.” https://matplotlib.org/.

Hyppönen, H, K Saranto, R Vuokko, P Mäkelä-Bengs, P Doupi, M Lindqvist, and M Mäkelä. 2014. “Impacts of Structuring the Electronic Health Record: a Systematic Review Protocol and Results of Previous Reviews.” *International Journal of Medical Informatics* 83 (3): 159–69.

Hyun, S, S Bakken, and S B Johnson. 2006. “Markup of Temporal Information in Electronic Health Records.” In, 122:907–8.

ISO/IEC JTC 1/SC 27 IT Security techniques. 2016. *ISO/IEC CD 20889 Privacy Enhancing Data De-Identification Techniques*. International Organization for Standardization. https://www.iso.org/standard/69373.html.

Ivanov, Oleg, Michael M Wagner, Wendy W Chapman, and Robert T Olszewski. 2002. “Accuracy of Three Classifiers of Acute Gastrointestinal Syndrome for Syndromic Surveillance..” *Proc AMIA Symp*, 345–49.

Jackson, Graham. 2005. “‘Oh … by the Way … ’: Doorknob Syndrome.” *International Journal of Clinical Practice* 59 (8). Blackwell Science Ltd: 869–69. doi:10.1111/j.1368-5031.2005.0599a.x.

Jackson, Richard G, Rashmi Patel, Nishamali Jayatilleke, Anna Kolliakou, Michael Ball, Genevieve Gorrell, Angus Roberts, Richard J Dobson, and Robert Stewart. 2017. “Natural Language Processing to Extract Symptoms of Severe Mental Illness From Clinical Text: the Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) Project..” *BMJ Open* 7 (1). British Medical Journal Publishing Group: e012012. doi:10.1136/bmjopen-2016-012012.

Jäger, Gerhard, and James Rogers. 2012. “Formal Language Theory: Refining the Chomsky Hierarchy.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 367 (1598). The Royal Society: 1956–70. doi:10.1098/rstb.2012.0077.

Johnson, S B, and C Friedman. 1996. “Integrating Data From Natural Language Processing Into a Clinical Information System..” *Proc AMIA Annu Fall Symp*, 537–41.

Jones, Karen Sparck. 1994. “Natural Language Processing: a Historical Review.” In *Current Issues in Computational Linguistics: in Honour of Don Walker*, 3–16. Dordrecht: Springer, Dordrecht. doi:10.1007/978-0-585-35958-8_1.

Jones, P H, S Dawson, R M Gaskell, K P Coyne, Tierney, C Setzkorn, A D Radford, and P J M Noble. 2014. “Surveillance of Diarrhoea in Small Animal Practice Through the Small Animal Veterinary Surveillance Network

(SAVSNET).” *The Veterinary Journal* 201 (3): 412–18. doi:10.1016/j.tvjl.2014.05.044.

Jones-Diette, Julie, Natalie J Robinson, Malcolm Cobb, Marnie L Brennan, and Rachel S Dean. 2017. “Accuracy of the Electronic Patient Record in a First Opinion Veterinary Practice.” *Preventive Veterinary Medicine* 148 (December). Elsevier: 121–26. doi:10.1016/j.prevetmed.2016.11.014.

Josseran, L, Javier Nicolau, N Caillère, P Astagneau, and G Brücker. 2006. “Syndromic Surveillance Based on Emergency Department Activity and Crude Mortality: Two Examples.” *Euro Surveillance : Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin* 11 (12). European Centre for Disease Prevention and Control: 11–12. doi:10.2807/esm.11.12.00668-en.

Kagashe, Ireneus, Zhijun Yan, and Imran Suheryani. 2017. “Enhancing Seasonal Influenza Surveillance: Topic Analysis of Widely Used Medicinal Drugs Using Twitter Data..” *Journal of Medical Internet Research* 19 (9). JMIR Publications Inc., Toronto, Canada: e315. doi:10.2196/jmir.7393.

Kahn, Michael G, Samson Tu, and Lawrence M Fagan. 1991. “TQuery: a Context-Sensitive Temporal Query Language.” *Computers and Biomedical Research* 24 (5): 401–19. doi:10.1016/0010-4809(91)90016-P.

Kajita, Emily, Monica Z Luarca, Han Wu, Bessie Hwang, and Laurene Mascola. 2017. “Harnessing Syndromic Surveillance Emergency Department Data to Monitor Health Impacts During the 2015 Special Olympics World Games..” *Public Health Reports (Washington, D.C. : 1974)* 132 (1_suppl). SAGE PublicationsSage CA: Los Angeles, CA: 99S–105S. doi:10.1177/0033354917706956.

Kass, Philip H, Hsin-Yi Weng, Mark A L Gaona, Amy Hille, Max H Sydow, Elizabeth M Lund, and Peter J Markwell. 2016. “Syndromic Surveillance in Companion Animals Utilizing Electronic Medical Records Data: Development and Proof of Concept..” *PeerJ* 4 (Suppl). PeerJ Inc.: e1940. doi:10.7717/peerj.1940.

Kass-Hout, T A, D Buckeridge, J Brownstein, Z Xu, P McMurray, C K Ishikawa, J Gunn, and B L Massoudi. 2012. “Self-Reported Fever and Measured Temperature in Emergency Department Records Used for Syndromic Surveillance.” *J Am Med Inform Assoc* 19 (5): 775–76. doi:10.1136/amiajnl-2012-000847.

Kavanagh, Kimberley, Chris Robertson, Heather Murdoch, George Crooks, and Jim McMenamin. 2012. “Syndromic Surveillance of Influenza-Like Illness in Scotland During the Influenza a H1N1v Pandemic and Beyond.” *Journal of the Royal Statistical Society: Series a (Statistics in Society)* 175 (4). Blackwell Publishing Ltd: 939–58. doi:10.1111/j.1467-985X.2012.01025.x.

Kay, S, and I N Purves. 1996. “Medical Records and Other Stories: a Narratological Framework.” *IMIA Yearbook 2008: Access to Health Information* 35 (2). Schattauer Publishers: 72–87.

Keravnou, Elpida T. 1996. “Temporal Reasoning in Medicine.” *Artif Intell Med* 8 (3): 187–91. doi:10.1016/0933-3657(95)00032-1.

Kite-Powell, A, and J Livengood. 2006. “Syndromic Surveillance of Emergency

Department Chief Complaints Post-Hurricane Wilma, Broward County, Florida 2005." In. Boston, MA.

Kittredge, Richard I. 1983. "Semantic Processing of Texts in Restricted Sublanguages." *Computers & Mathematics with Applications* 9 (1): 45–58.

Kleene, S C. 1956. *Representation of Events in Nerve Nets and Finite Automata*. Edited by C Shannon, J McCarthy, and undefined author. *Automata Studies*. Princeton, NJ: Princeton University Press.

Kukich, K. 1992. "Techniques for Automatically Correcting Words in Text." *ACM Computing Surveys (CSUR)*.

Kuramoto-Crawford, S Janet, Erica L Spies, and John Davies-Cole. 2017. "Detecting Suicide-Related Emergency Department Visits Among Adults Using the District of Columbia Syndromic Surveillance System." *Public Health Reports* 132 (1_suppl). SAGE PublicationsSage CA: Los Angeles, CA: 88S–94S. doi:10.1177/0033354917706933.

Lai, K.H., Topaz, M., Goss, F.R., Zhou,L., 2015 "Automated misspelling detection and correction in clinical free-text records." Journal of Biomedical Informatics, 55,188-195.

Lall, Ramona, Jasmine Abdelnabi, Stephanie Ngai, Hilary B Parton, Kelly Saunders, Jessica Sell, Amanda Wahnich, Don Weiss, and Robert W Mathes. 2017. "Advancing the Use of Emergency Department Syndromic Surveillance Data, New York City, 2012-2016.." *Public Health Reports (Washington, D.C. : 1974)* 132 (1_suppl). SAGE PublicationsSage CA: Los Angeles, CA: 23S–30S. doi:10.1177/0033354917711183.

Lam, K, T Parkin, C Riggs, and K Morgan. 2007. "Use of Free Text Clinical Records in Identifying Syndromes and Analysing Health Data." *Veterinary Record* 161 (16): 547–51. doi:10.1136/vr.161.16.547.

Langmuir, A D. 1976. *William Farr: Founder of Modern Concepts of Surveillance. International Journal of Epidemiology*. Vol. 5.

Langmuir, Alexander D. 1963. "The Surveillance of Communicable Diseases of National Importance." *The New England Journal of Medicine* 268 (4): 182–92.

Ledley, R S, and L B Lusted. 1959. "Reasoning Foundations of Medical Diagnosis; Symbolic Logic, Probability, and Value Theory Aid Our Understanding of How Physicians Reason.." *Science (New York, N.Y.)* 130 (3366): 9–21.

Lee, Hyukki, Soohyung Kim, Jong Wook Kim, and Yon Dohn Chung. 2017. "Utility-Preserving Anonymization for Health Data Publishing.." *BMC Medical Informatics and Decision Making* 17 (1). BioMed Central: 104. doi:10.1186/s12911-017-0499-0.

Lenat, Douglas B. 1995. "CYC: a Large-Scale Investment in Knowledge Infrastructure." *Commun. ACM* 38 (11). ACM: 33–38. doi:10.1145/219717.219745.

Li, Ninghui, Tiancheng Li, and Suresh Venkatasubramanian. 2007. "T-Closeness: Privacy Beyond K-Anonymity and L-Diversity." In, 106–15. IEEE.

doi:10.1109/ICDE.2007.367856.

Liljeqvist, Henning T G, David Muscatello, Grant Sara, Michael Dinh, and Glenda L Lawrence. 2014. “Accuracy of Automatic Syndromic Classification of Coded Emergency Department Diagnoses in Identifying Mental Health-Related Presentations for Public Health Surveillance..” *BMC Medical Informatics and Decision Making* 14 (1). BioMed Central: 84. doi:10.1186/1472-6947-14-84.

Lin, Chen, Dmitriy Dligach, Timothy A Miller, Steven Bethard, and Guergana K Savova. 2016. “Multilayered Temporal Modeling for the Clinical Domain..” *J Am Med Inform Assoc* 23 (2): 387–95. doi:10.1093/jamia/ocv113.

Liu, Hongfang, and Carol Friedman. 2004. “CliniViewer: a Tool for Viewing Electronic Medical Records Based on Natural Language Processing and XML..” *Stud Health Technol Inform* 107 (Pt 1): 639–43.

Liu, Hongfang, Virginia Teller, and Carol Friedman. 2004. “A Multi-Aspect Comparison Study of Supervised Word Sense Disambiguation.” *J Am Med Inform Assoc* 11 (4). American Medical Informatics Association: 320–31. doi:10.1197/jamia.M1533.

Liu, Hongfang, Yves A Lussier, and Carol Friedman. 2001. “Disambiguating Ambiguous Biomedical Terms in Biomedical Narrative Text: an Unsupervised Method.” *Journal of Biomedical Informatics* 34 (4): 249–61. doi:10.1006/jbin.2001.1023.

Liu, K, W Chapman, R Hwa, and R S Crowley. 2007. “Heuristic Sample Selection to Minimize Reference Standard Training Set for a Part-of-Speech Tagger.” *J Am Med Inform Assoc* 14 (5): 641–50. doi:10.1197/jamia.M2392.

Lloyd, Susan S. 1985. “Physician and Coding Errors in Patient Records.” *Jama* 254 (10): 1330. doi:10.1001/jama.1985.03360100080018.

Lombardo, J, H Burkom, E Elbert, S Magruder, S Happel Lewis, W Loschen, J Sari, C Sniegoski, R Wojcik, and J Pavlin. 2003. “A Systems Overview of the Electronic Surveillance System for the Early Notification of Community-Based Epidemics (ESSENCE II).” *Journal of Urban Health* 80 (2 SUPPL. 1): i32–i42.

Lujic, Sanja, Diane E Watson, Deborah A Randall, Judy M Simpson, and Louisa R Jorm. 2014. “Variation in the Recording of Common Health Conditions in Routine Hospital Data: Study Using Linked Survey and Administrative Data in New South Wales, Australia.” *BMJ Open* 4 (9). British Medical Journal Publishing Group: e005768–68. doi:10.1136/bmjopen-2014-005768.

Lund, E M. 2015. “Power of Practice: Using Clinical Data to Advance Veterinary Medicine.” *Veterinary Record* 176 (2): 46–47. doi:10.1136/vr.g7763.

Lyman, Margaret, Naomi Sager, Emile C Chi, Leo J Tick, Ngo Thanh Nhan, Yun Su, Francois Borst, and Jean-Raoul Scherrer. 1989. “Medical Language Processing for Knowledge Representation and Retrievals.” In, 548–53.

Maas, Heinz-Dieter. 1972. “Über Den Zusammenhang Zwischen Wortschatzumfang Und Länge Eines Textes [the Relationship Between Lexical Diversity and the Length of a Sample].” *Zeitschrift Für Literaturwissenschaft Und Linguistik* 2 (8): 73–79.

Machanavajjhala, A, J Gehrke, D Kifer, and M Venkitasubramaniam. 2006. “L-Diversity: Privacy Beyond K-Anonymity.” In, 24–24. IEEE. doi:10.1109/ICDE.2006.1.

MacIntyre, C Raina, Michael J Ackland, Eugene J Chandraraj, and John E Pilla. 1997. “Accuracy of ICD–9–CM Codes in Hospital Morbidity Data, Victoria: Implications for Public Health Research.” *Australian and New Zealand Journal of Public Health* 21 (5). Wiley/Blackwell (10.1111): 477–82. doi:10.1111/j.1467-842X.1997.tb01738.x.

Madouasse, Aurélien, Alexis Marceau, Anne Lehébel, Henriëtte Brouwer-Middelesch, Gerdien van Schaik, Yves Van der Stede, and Christine Fourichon. 2013. “Evaluation of a Continuous Indicator for Syndromic Surveillance Through Simulation. Application to Vector Borne Disease Emergence Detection in Cattle Using Milk Yield.” Edited by Yung-Fu Chang. *PLoS ONE* 8 (9). Public Library of Science: e73726. doi:10.1371/journal.pone.0073726.

Marceau, Alexis, Aurélien Madouasse, Anne Lehébel, Gerdien van Schaik, Anouk Veldhuis, Yves Van der Stede, and Christine Fourichon. 2014. “Can Routinely Recorded Reproductive Events Be Used as Indicators of Disease Emergence in Dairy Cattle? an Evaluation of 5 Indicators During the Emergence of Bluetongue Virus in France in 2007 and 2008..” *Journal of Dairy Science* 97 (10): 6135–50. doi:10.3168/jds.2013-7346.

Martini, M, Massimo Fenati, Maristella Agosti, Rudi Cassini, Michele Drigo, Nicola Ferro, Carlo Guglielmini, Ivano Masiero, Manuela Signorini, and Roberto Busetto. 2017. “A Surveillance System for Diseases of Companion Animals in the Veneto Region (Italy).” *Revue Scientifique Et Technique International Office of Epizootics* 36 (3).

Mathes, R W, K Ito, and T Matte. 2011. “Assessing Syndromic Surveillance of Cardiovascular Outcomes From Emergency Department Chief Complaint Data in New York City.” Edited by J Jaime Miranda. *PLoS ONE* 6 (2): e14677. doi:10.1371/journal.pone.0014677.

Mayer, Chad L, Caitlin S Leibowitz, Shinichiro Kurosawa, and Deborah J Stearns-Kurosawa. 2012. “Shiga Toxins and the Pathophysiology of Hemolytic Uremic Syndrome in Humans and Animals.” *Toxins* 4 (12). Molecular Diversity Preservation International: 1261–87. doi:10.3390/toxins4111261.

McCulloch, WarrenS, and Walter Pitts. 1943. “A Logical Calculus of the Ideas Immanent in Nervous Activity.” *The Bulletin of Mathematical Biophysics* 5 (4). Kluwer Academic Publishers: 115–33. doi:10.1007/BF02478259.

McDonald, C J. 1997. “The Barriers to Electronic Medical Record Systems and How to Overcome Them..” *J Am Med Inform Assoc* 4 (3). UNITED STATES: Hanley & Belfus: 213–21.

McDonald, J L, I R Cleasby, D C Brodbelt, D B Church, and D G O'Neill. 2017. “Mortality Due to Trauma in Cats Attending Veterinary Practices in Central and South-East England..” *The Journal of Small Animal Practice* 58 (10): 570–76. doi:10.1111/jsap.12716.

McGreevy, Paul, Peter Thomson, Navneet K Dhand, David Raubenheimer, Sophie Masters, Caroline S Mansfield, Timothy Baldwin, et al. 2017.

"VetCompass Australia: a National Big Data Collection System for Veterinary Science.." *Animals : an Open Access Journal From MDPI* 7 (10). Multidisciplinary Digital Publishing Institute: 74. doi:10.3390/ani7100074.

McHugh, Mary L. 2012. "Interrater Reliability: the Kappa Statistic.." *Biochemia Medica* 22 (3). Croatian Society for Medical Biochemistry and Laboratory Medicine: 276–82.

McLaughlin, G Harry. 1969. "SMOG Grading-a New Readability Formula." *Journal of Reading* 12 (8). [Wiley, International Reading Association]: 639–46.

Merck Sharp & Dohme Corporation. 2018. "MSD Veterinary Manual." https://www.msdvetmanual.com.

Meystre, S M, G K Savova, K C Kipper-Schuler, and J F Hurdle. 2008a. "Extracting Information From Textual Documents in the Electronic Health Record: a Review of Recent Research." *IMIA Yearbook 2008: Access to Health Information* 3 (1). Schattauer Publishers: 128–44.

Meystre, S M, O Ferrandez, F J Friedlin, B R South, S Shen, and M H Samore. 2014. "Text De-Identification for Privacy Protection: a Study of Its Impact on Clinical Text Information Content." *Journal of Biomedical Informatics* 50 (August): 142–50. doi:10.1016/j.jbi.2014.01.011.

Meystre, S, and P J Haug. 2005. "Automation of a Problem List Using Natural Language Processing." *BMC Medical Informatics and Decision Making* 5 (1): 30. doi:10.1186/1472-6947-5-30.

Meystre, S, and P J Haug. 2006. "Natural Language Processing to Extract Medical Problems From Electronic Clinical Documents: Performance Evaluation." *Journal of Biomedical Informatics* 39 (6): 589–99. doi:10.1016/j.jbi.2005.11.004.

Meystre, S, G Savova, K Kipper-Schuler, and J Hurdle. 2008b. "Extracting Information From Textual Documents in the Electronic Health Record: a Review of Recent Research." *Methods Inf Med* 47 (2008//).

Meystre, Stephane M, F Jeffrey Friedlin, Brett R South, Shuying Shen, and Matthew H Samore. 2010. "Automatic De-Identification of Textual Documents in the Electronic Health Record: a Review of Recent Research.." *BMC Medical Research Methodology* 10 (1). BioMed Central: 70. doi:10.1186/1471-2288-10-70.

Meystre, Stéphane, Shuying Shen, Deborah Hofmann, and Adi Gundlapalli. 2014. "Can Physicians Recognize Their Own Patients in De-Identified Notes?." *Stud Health Technol Inform* 205: 778–82.

Mitchell, Kevin J, Michael J Becich, Jules J Berman, Wendy W Chapman, John Gilbertson, Dilip Gupta, James Harrison, Elizabeth Legowski, and Rebecca S Crowley. 2004. "Implementation and Evaluation of a Negation Tagger in a Pipeline-Based System for Information Extract From Pathology Reports.." *Stud Health Technol Inform* 107 (Pt 1): 663–67.

Morens, David M, and Anthony S Fauci. 2013. "Emerging Infectious Diseases: Threats to Human Health and Global Stability.." Edited by Joseph Heitman. *PLoS Pathogens* 9 (7). Public Library of Science: e1003467.

doi:10.1371/journal.ppat.1003467.

Morrison, Frances P, Li Li, Albert M Lai, and George Hripcsak. 2009. "Repurposing the Clinical Record: Can an Existing Natural Language Processing System De-Identify Clinical Notes?." *J Am Med Inform Assoc* 16 (1): 37–39. doi:10.1197/jamia.M2862.

Murray, J K, R A Casey, E Gale, C A T Buffington, C Roberts, R H Kinsman, and T J Gruffydd-Jones. 2017. "Cohort Profile: the 'Bristol Cats Study' (BCS)-a Birth Cohort of Kittens Owned by UK Households.." *International Journal of Epidemiology*, June. doi:10.1093/ije/dyx066.

Murray, J K, W J Browne, M A Roberts, A Whitmarsh, and T J Gruffydd-Jones. 2010. "Number and Ownership Profiles of Cats and Dogs in the UK." *Veterinary Record* 166 (6): 163–68. doi:10.1136/vr.b4712.

Mutalik, P G, A Deshpande, and P M Nadkarni. 2001. "Use of General-Purpose Negation Detection to Augment Concept Indexing of Medical Documents: a Quantitative Study Using the UMLS.." *J Am Med Inform Assoc* 8 (6): 598–609.

National Institute for Health Research. 2018. "Clinical Practice Research Datalink." https://www.cprd.com/.

National Library of Medicine. 2009. "Unified Medical Language System (UMLS) Metathesaurus." U.S. National Library of Medicine. https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/.

National Library of Medicine. 2016. "PubMed Database." *Ncbi.Nlm.Nih.Gov*. https://www.ncbi.nlm.nih.gov/pubmed/.

National Library of Medicine. 2017a. "SNOMEDCT_VET (the Veterinary Extension to SNOMED CT) - Synopsis." U.S. National Library of Medicine. https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/SNOMEDCT_VET/.

National Library of Medicine. 2017b. "Snomed." https://www.nlm.nih.gov/healthit/snomedct/.

Neamatullah, Ishna, Margaret M Douglass, Li-wei H Lehman, Andrew Reisner, Mauricio Villarroel, William J Long, Peter Szolovits, George B Moody, Roger G Mark, and Gari D Clifford. 2008. "Automated De-Identification of Free-Text Medical Records." *BMC Medical Informatics and Decision Making* 8 (1): 1–17. doi:10.1186/1472-6947-8-32.

Neighbour, R. 2004. *The Inner Consultation: How to Develop an Effective and Intuitive Consulting Style*. 2nd ed. Radcliffe Medical Press.

Nelson, Gregory S. 2015. "Practical Implications of Sharing Data: a Primer on Data Privacy, Anonymization, and De-Identification." In. Cary, NC: SAS Institute Inc. https://support.sas.com/resources/papers/proceedings15/1884-2015.pdf.

New Forest Dog Owners Group. 2018. "NFDOG Research Fund." *New Forest Dog Owners Group*. https://www.newforestdog.org.uk/nfdog-research-fund.

NHS Employers. 2016. "General Medical Services Contract Qualiy and Outcomes Framework." *NHS Employers*.

http://www.nhsemployers.org/your-workforce/primary-care-contacts/general-medical-services/quality-and-outcomes-framework.

NLTK Project. 2015. “Natural Language Toolkit 3.0 Documentation.” *Nltk.org*. http://www.nltk.org/.

Noble, Peter-John M, Jenny Newman, Alison M Wyatt, Alan D Radford, and Philip H Jones. 2017. “Heightened Risk of Canine Chocolate Exposure at Christmas and Easter..” *The Veterinary Record* 181 (25). British Medical Journal Publishing Group: 684–84. doi:10.1136/vr.104762.

Nouraei, Seyed Ahmad Reza, Jagdeep Singh Virk, Anita Hudovsky, Christopher Wathen, Ara Darzi, and Darren Parsons. 2016. “Accuracy of Clinician-Clinical Coder Information Handover Following Acute Medical Admissions: Implication for Using Administrative Datasets in Clinical Outcomes Management.” *Journal of Public Health* 38 (2): 352–62. doi:10.1093/pubmed/fdv041.

Numpy developers. 2017. “NumPy.” *Numpy.org*. http://www.numpy.org/.

Nusinovici, Simon, Aurélien Madouasse, and Christine Fourichon. 2016. “Quantification of the Increase in the Frequency of Early Calving Associated with Late Exposure to Bluetongue Virus Serotype 8 in Dairy Cows: Implications for Syndromic Surveillance..” *Veterinary Research* 47 (1). BioMed Central: 18. doi:10.1186/s13567-015-0296-7.

Nusinovici, Simon, Pascal Monestiez, Henri Seegers, François Beaudeau, and Christine Fourichon. 2014. “Using Animal Performance Data to Evidence the Under-Reporting of Case Herds During an Epizootic: Application to an Outbreak of Bluetongue in Cattle.” Edited by Houssam Attoui. *PLoS ONE* 9 (6). Public Library of Science: e100137. doi:10.1371/journal.pone.0100137.

O'Malley, Kimberly J, Karon F Cook, Matt D Price, Kimberly Raiford Wildes, John F Hurdle, and Carol M Ashton. 2005. “Measuring Diagnoses: ICD Code Accuracy..” *Health Services Research* 40 (5 Pt 2). Blackwell Publishing: 1620–39. doi:10.1111/j.1475-6773.2005.00444.x.

O'Neill, D. 2013. “Surveillance: Pointing the Way to Improved Welfare for Companion Animals..” *The Veterinary Record* 173 (10). British Medical Journal Publishing Group: 240–42. doi:10.1136/vr.f4519.

O'Neill, D G, A Riddell, D B Church, L Owen, D C Brodbelt, and J L Hall. 2017. “Urinary Incontinence in Bitches Under Primary Veterinary Care in England: Prevalence and Risk Factors..” *The Journal of Small Animal Practice* 28 (September): 153. doi:10.1111/jsap.12731.

O'Neill, D G, C Scudder, J M Faire, D B Church, P D McGreevy, P C Thomson, and D C Brodbelt. 2016. “Epidemiology of Hyperadrenocorticism Among 210,824 Dogs Attending Primary-Care Veterinary Practices in the UK From 2009 to 2014..” *The Journal of Small Animal Practice* 57 (7). Blackwell Publishing Ltd: 365–73. doi:10.1111/jsap.12523.

O'Neill, D G, J Case, A K Boag, D B Church, P D McGreevy, P C Thomson, and D C Brodbelt. 2017. “Gastric Dilation-Volvulus in Dogs Attending UK Emergency-Care Veterinary Practices: Prevalence, Risk Factors and Survival..” *The Journal of Small Animal Practice* 58 (11): 629–38. doi:10.1111/jsap.12723.

O'Neill, D G, R Gostelow, C Orme, D B Church, S J M Niessen, K Verheyen, and D C Brodbelt. 2016. “Epidemiology of Diabetes Mellitus Among 193,435 Cats Attending Primary-Care Veterinary Practices in England..” *Journal of Veterinary Internal Medicine* 30 (4): 964–72. doi:10.1111/jvim.14365.

O'Neill, D, A Hendricks, J Summers, and D Brodbelt. 2012. “Primary Care Veterinary Usage of Systemic Glucocorticoids in Cats and Dogs in Three UK Practices.” *The Journal of Small Animal Practice* 53 (4). Blackwell Publishing Ltd: 217–22. doi:10.1111/j.1748-5827.2011.01190.x.

O'Neill, Dan G, David B Church, Paul D McGreevy, Peter C Thomson, and David C Brodbelt. 2015. “Longevity and Mortality of Cats Attending Primary Care Veterinary Practices in England..” *Journal of Feline Medicine and Surgery* 17 (2). SAGE PublicationsSage UK: London, England: 125–33. doi:10.1177/1098612X14536176.

O'Neill, Dan G, Richard L Meeson, Adam Sheridan, David B Church, and Dave C Brodbelt. 2016. “The Epidemiology of Patellar Luxation in Dogs Attending Primary-Care Veterinary Practices in England..” *Canine Genetics and Epidemiology* 3 (1). BioMed Central: 4. doi:10.1186/s40575-016-0034-0.

OECD. 2010. “Improving Health Sector Efficiency: the Role of Information and Communication Technologies.” Organisation for Economic Co-operation and Development (OECD).

Office For National Statistics. 2015a. “Baby Names, England and Wales, 2014.” *Ons.Gov.Uk*. http://www.ons.gov.uk/ons/publications/.

Office For National Statistics. 2015b. “Geography Portal.” *Geoportal.Statistics.Gov.Uk*. https://geoportal.statistics.gov.uk/geoportal/catalog/main/home.page.

Pakhomov, Serguei V, Anni Coden, and Christopher G Chute. 2006. “Developing a Corpus of Clinical Notes Manually Annotated for Part-of-Speech.” *Int J Med Inform* 75 (6): 418–29. doi:10.1016/j.ijmedinf.2005.08.006.

Paterson, Beverley J, and David N Durrheim. 2013. “The Remarkable Adaptability of Syndromic Surveillance to Meet Public Health Needs.” *Journal of Epidemiology and Global Health* 3 (1): 41–47. doi:10.1016/j.jegh.2012.12.005.

Payne, Thomas H, Don E Detmer, Jeremy C Wyatt, and Iain E Buchan. 2011. “National-Scale Clinical Information Exchange in the United Kingdom: Lessons for the United States..” *J Am Med Inform Assoc* 18 (1): 91–98. doi:10.1136/jamia.2010.005611.

Pearl, Raymond. 1930. “Introduction to Medical Biometry and Statistics.,” no. Second Edition. London : W. B. Saunders Company: 459pp.

Peeraully, R, K Henderson, and B Davies. 2016. “Emergency Readmissions to Paediatric Surgery and Urology: the Impact of Inappropriate Coding..” *Annals of the Royal College of Surgeons of England* 98 (4). Royal College of Surgeons: 250–53. doi:10.1308/rcsann.2016.0067.

Penz, J F, A B Wilcox, and J F Hurdle. 2007. “Automated Identification of Adverse Events Related to Central Venous Catheters.” *Journal of*

Biomedical Informatics 40 (2): 174–82. doi:10.1016/j.jbi.2006.06.003.

Perktold, Josef, Skipper Seabold, Jonathan Taylor, statsmodels-developers. 2017. “StatsModels Statistics in Python.” http://www.statsmodels.org/.

Perrin, Jean-Baptiste, Christian Ducrot, Jean-Luc Vinard, Eric Morignat, Didier Calavas, and Pascal Hendrikx. 2012. “Assessment of the Utility of Routinely Collected Cattle Census and Disposal Data for Syndromic Surveillance..” *Preventive Veterinary Medicine* 105 (3): 244–52. doi:10.1016/j.prevetmed.2011.12.015.

Peters, Tim. 2004. “PEP 20 -- the Zen of Python.” *Python*. https://www.python.org/dev/peps/pep-0020/#id3.

Pogreba-Brown, Kristen, Kyle McKeown, Sarah Santana, Alisa Diggs, Jennifer Stewart, and Robin B Harris. 2013. “Public Health in the Field and the Emergency Operations Center: Methods for Implementing Real-Time Onsite Syndromic Surveillance at Large Public Events..” *Disaster Medicine and Public Health Preparedness* 7 (5): 467–74. doi:10.1017/dmp.2013.83.

Porta, Miquel S, Sander Greenland, Miguel Hernan, Isabel dos Santos Silva, and John M Last. 2014. “A Dictionary of Epidemiology..” 6th ed.

Princeton University. 2018. “WordNet a Lexical Database for English.” *Princeton University*. http://wordnet.princeton.edu/.

Public Health England. 2010. *The Health Protection (Notification) Regulations 2010*. http://www.legislation.gov.uk/uksi/2010/659/pdfs/uksi_20100659_en.pdf.

Public Health England. 2015. “Syndromic Surveillance: Systems and Analyses.” *Gov.Uk*. https://www.gov.uk/government/collections/syndromic-surveillance-systems-and-analyses.

PyData Development Team. 2017. “Python Data Analysis Library.” *Pandas.Pydata.org*. http://pandas.pydata.org/.

pymssql developers. 2016. “Pymssql.” http://www.pymssql.org/.

Python Software Foundation. 2016. “Python.” www.python.org.

Python Software Foundation. 2018. “Regular Expression Operations.” *Python Standard Library*. https://docs.python.org/3.6/library/re.html.

Radford, A D, P J Noble, K P Coyne, R M Gaskell, P H Jones, J G E Bryan, C Setzkorn, Á Tierney, and S Dawson. 2011. “Antibacterial Prescribing Patterns in Small Animal Veterinary Practice Identified via SAVSNET: the Small Animal Veterinary Surveillance Network.” *Veterinary Record* 169 (12): 310–10. doi:10.1136/vr.d5062.

Radford, Alan, Aine Tierney, Karen Coyne, Susan Dawson, P J Noble, and Ros Gaskell. 2010. “National Surveillance of Small Animal Disease in the UK..” *The Veterinary Record* 166 (15). British Medical Journal Publishing Group: 471–72. doi:10.1136/vr.c1676.

Rangachari, Pavani. 2007. “Coding for Quality Measurement: the Relationship Between Hospital Structural Characteristics and Coding Accuracy From the Perspective of Quality Measurement..” *Perspectives in Health Information*

Management 4 (April). American Health Information Management Association: 3.

Rappold, Ana G, Susan L Stone, Wayne E Cascio, Lucas M Neas, Vasu J Kilaru, Martha Sue Carraway, James J Szykman, et al. 2011. “Peat Bog Wildfire Smoke Exposure in Rural North Carolina Is Associated with Cardiopulmonary Emergency Department Visits Assessed Through Syndromic Surveillance.” *Environmental Health Perspectives* 119 (10). National Institute of Environmental Health Science: 1415–20. doi:10.1289/ehp.1003206.

Rees-Miller, Janie, and Mark Aronoff. 2003. “The Handbook of Linguistics.” Wiley-Blackwell. https://www.wiley.com/en-us/The+Handbook+of+Linguistics-p-9781405102520.

Reiser, Stanley J. 1991a. “The Clinical Record in Medicine Part 1: Learning From Cases.” *Annals of Internal Medicine* 114 (10): 902. doi:10.7326/0003-4819-114-10-902.

Reiser, Stanley J. 1991b. “The Clinical Record in Medicine Part 2: Reforming Content and Purpose.” *Annals of Internal Medicine* 114 (11): 980. doi:10.7326/0003-4819-114-11-980.

Requejo, Maria Dolores Porto. 2009. “The Role of Context in Word Meaning Construction: a Case Study.” *International Journal of English Studies* 7 (1): 169–79.

Ribeiro-Neto, B, A H F Laender, and L R S De Lima. 2001. “An Experimental Study in Automatically Categorizing Medical Documents.” *Journal of the American Society for Information Science and Technology* 52 (5): 391–401.

Robertson, Colin, and Lauren Yee. 2016. “Avian Influenza Risk Surveillance in North America with Online Media..” Edited by Dena L Schanzer. *PLoS ONE* 11 (11): e0165688. doi:10.1371/journal.pone.0165688.

Robinson, D, and H Hooker. 2006. “The UK Veterinary Profession in 2006: the Findings of a Survey of the Profession Conducted by the Royal College of Veterinary Surgeons. Royal College of Veterinary Surgeons..”

Robinson, N J, M L Brennan, M Cobb, and R S Dean. 2015. “Capturing the Complexity of First Opinion Small Animal Consultations Using Direct Observation..” *The Veterinary Record* 176 (2). British Medical Journal Publishing Group: 48–48. doi:10.1136/vr.102548.

Robinson, N J, R S Dean, M Cobb, and M L Brennan. 2014. “Consultation Length in First Opinion Small Animal Practice..” *The Veterinary Record* 175 (19): 486–86. doi:10.1136/vr.102713.

Rotermund, A, M Peters, M Hewicker-Trautwein, and I Nolte. 2002. “Cutaneous and Renal Glomerular Vasculopathy in a Great Dane Resembling ‘Alabama Rot’ of Greyhounds..” *Veterinary Record* 151 (17): 510–12.

Royal College of Veterinary Surgeons. 2014. *Code of Professional Conduct for Veterinary Surgeons*.

Royal College of Veterinary Surgeons. 2015. *Rcvs.org.Uk*. https://www.rcvs.org.uk/.

Royal College of Veterinary Surgeons. 2016. *Code of Professional Conduct for Veterinary Surgeons*. www.rcvs.org.uk/.

Ruch, Patrick, Robert Baud, and Antoine Geissbühler. 2003. “Using Lexical Disambiguation and Named-Entity Recognition to Improve Spelling Correction in the Electronic Patient Record..” *Artif Intell Med* 29 (1-2): 169–84.

Ruple-Czerniak, A A, H W Aceto, J B Bender, M R Paradis, S P Shaw, D C Van Metre, J S Weese, D A Wilson, J Wilson, and P S Morley. 2014. “Syndromic Surveillance for Evaluating the Occurrence of Healthcare-Associated Infections in Equine Hospitals..” *Equine Veterinary Journal* 46 (4): 435–40. doi:10.1111/evj.12190.

Ruple-Czerniak, A, H W Aceto, J B Bender, M R Paradis, S P Shaw, D C Van Metre, J S Weese, D A Wilson, J H Wilson, and P S Morley. 2013. “Using Syndromic Surveillance to Estimate Baseline Rates for Healthcare-Associated Infections in Critical Care Units of Small Animal Referral Hospitals..” *Journal of Veterinary Internal Medicine* 27 (6): 1392–99. doi:10.1111/jvim.12190.

Samal, L, A Wright, B T Wong, J A Linder, and D W Bates. 2011. “Leveraging Electronic Health Records to Support Chronic Disease Management: the Need for Temporal Data Views.” *Informatics in Primary Care* 19 (2): 65–74.

Santos, Suong, Gregory Murphy, Kathryn Baxter, and Kerin M Robinson. 2008. “Organisational Factors Affecting the Quality of Hospital Clinical Coding..” *Health Information Management : Journal of the Health Information Management Association of Australia* 37 (1): 25–37.

Sapir, Edward. 1921. “Language as a Historical Product: Drift..” In *Language: an Introduction to the Study of Speech.*, 147–70. New York: Harcourt Brace & Company. doi:10.1037/13026-007.

Savova, Guergana K, Anni R Coden, Igor L Sominsky, Rie Johnson, Philip V Ogren, Piet C de Groen, and Christopher G Chute. 2008. “Word Sense Disambiguation Across Two Domains: Biomedical Literature and Clinical Notes.” *Journal of Biomedical Informatics* 41 (6): 1088–1100. doi:10.1016/j.jbi.2008.02.003.

Sánchez-Vizcaíno, Fernando, David Singleton, Philip H Jones, Bethaney Heayns, Maya Wardeh, Alan D Radford, Vanessa Schmidt, Susan Dawson, Peter J M Noble, and Sally Everitt. 2016. “Small Animal Disease Surveillance: Pruritus, and Coagulase-Positive Staphylococci..” *The Veterinary Record* 179 (14): 352–55. doi:10.1136/vr.i5322.

Sánchez-Vizcaíno, Fernando, Peter-John M Noble, Phil H Jones, Tarek Menacere, Iain Buchan, Suzanna Reynolds, Susan Dawson, Rosalind M Gaskell, Sally Everitt, and Alan D Radford. 2017. “Demographics of Dogs, Cats, and Rabbits Attending Veterinary Practices in Great Britain as Recorded in Their Electronic Health Records..” *BMC Veterinary Research* 13 (1). BioMed Central: 218. doi:10.1186/s12917-017-1138-9.

Sánchez-Vizcaíno, Fernando, Philip H Jones, Tarek Menacere, Bethaney Heayns, Maya Wardeh, Jenny Newman, Alan D Radford, et al. 2015. “Small Animal Disease Surveillance..” *The Veterinary Record* 177 (23). British Medical Journal Publishing Group: 591–94. doi:10.1136/vr.h6174.

Schmidt, Peggy L. 2007. “Evidence-Based Veterinary Medicine: Evolution, Revolution, or Repackaging of Veterinary Practice?.” *The Veterinary Clinics of North America. Small Animal Practice* 37 (3): 409–17. doi:10.1016/j.cvsm.2007.01.001.

Schoen, Cathy, Robin Osborn, David Squires, Michelle Doty, Petra Rasmussen, Roz Pierson, and Sandra Applebaum. 2012. “A Survey of Primary Care Doctors in Ten Countries Shows Progress in Use of Health Information Technology, Less in Other Areas..” *Health Affairs (Project Hope)* 31 (12). Health Affairs: 2805–16. doi:10.1377/hlthaff.2012.0884.

Schoen, Cathy, Robin Osborn, Michelle M Doty, David Squires, Jordon Peugh, and Sandra Applebaum. 2009. “A Survey of Primary Care Physicians in Eleven Countries, 2009: Perspectives on Care, Costs, and Experiences..” *Health Affairs (Project Hope)* 28 (6). Project HOPE - The People-to-People Health Foundation, Inc.: w1171–83. doi:10.1377/hlthaff.28.6.w1171.

Schrader, C D, and L M Lewis. 2013. “Racial Disparity in Emergency Department Triage.” *Journal of Emergency Medicine* 44 (2): 511–18.

SciPy developers. 2018. “SciPy.” https://www.scipy.org/.

Scottish Government. 2015. “National Records of Scotland.” *Gro-Scotland.Gov.Uk*. http://www.gro-scotland.gov.uk/statistics-and-data/statistics/.

Seil, Kacie, Jennifer Marcum, Ramona Lall, and Catherine Stayton. 2015. “Utility of a Near Real-Time Emergency Department Syndromic Surveillance System to Track Injuries in New York City..” *Injury Epidemiology* 2 (1). Nature Publishing Group: 11. doi:10.1186/s40621-015-0044-5.

Seo, Dong-Woo, and Soo-Yong Shin. 2017. “Methods Using Social Media and Search Queries to Predict Infectious Disease Outbreaks..” *Healthcare Informatics Research* 23 (4): 343–48. doi:10.4258/hir.2017.23.4.343.

Shahar, Yuval. 1997. “A Framework for Knowledge-Based Temporal Abstraction.” *Artificial Intelligence* 90 (1–2): 79–133. doi:10.1016/S0004-3702(96)00025-2.

Shapiro, Alan R. 2004. “Taming Variability in Free Text: Application to Health Surveillance” 53: 95–100. doi:10.1037/e307182005-018.

Sharpe, J Danielle, Richard S Hopkins, Robert L Cook, and Catherine W Striley. 2016. “Evaluating Google, Twitter, and Wikipedia as Tools for Influenza Surveillance Using Bayesian Change Point Analysis: a Comparative Analysis..” *JMIR Public Health and Surveillance* 2 (2). JMIR Publications Inc., Toronto, Canada: e161. doi:10.2196/publichealth.5901.

Shaw, Jane R, Cindy L Adams, Brenda N Bonnett, Susan Larson, and Debra L Roter. 2008. “Veterinarian-Client-Patient Communication During Wellness Appointments Versus Appointments Related to a Health Problem in Companion Animal Practice..” *Journal of the American Veterinary Medical Association* 233 (10). American Veterinary Medical Association 1931 North Meacham Road - Suite 100, Schaumburg, IL 60173 USA 847-925-8070 847-925-1329 avmajournals@avma.org: 1576–86. doi:10.2460/javma.233.10.1576.

Shin, Soo-Yong, Taerim Kim, Dong-Woo Seo, Chang Hwan Sohn, Sung-Hoon Kim, Seung Mok Ryoo, Yoon-Seon Lee, Jae Ho Lee, Won Young Kim, and Kyoung Soo Lim. 2016. “Correlation Between National Influenza Surveillance Data and Search Queries From Mobile Devices and Desktops in South Korea.” Edited by Donald R Olson. *PLoS ONE* 11 (7). Public Library of Science: e0158539. doi:10.1371/journal.pone.0158539.

Shoop, Stephanie Jw, Stephanie Marlow, David B Church, Kate English, Paul D McGreevy, Anneliese J Stell, Peter C Thomson, Dan G O'Neill, and David C Brodbelt. 2015. “Prevalence and Risk Factors for Mast Cell Tumours in Dogs in England..” *Canine Genetics and Epidemiology* 2 (1). BioMed Central: 1. doi:10.1186/2052-6687-2-1.

Sidorov, J. 2006. “It Ain't Necessarily So: the Electronic Health Record and the Unlikely Prospect of Reducing Health Care Costs.” *Health Affairs* 25 (4): 1079–85.

Siegler, Eugenia L. 2010. “The Evolving Medical Record.” *Annals of Internal Medicine* 153 (10): 671. doi:10.7326/0003-4819-153-10-201011160-00012.

Singleton, D A, F Sánchez-Vizcaíno, S Dawson, P H Jones, P J M Noble, G L Pinchbeck, N J Williams, and A D Radford. 2017. “Patterns of Antimicrobial Agent Prescription in a Sentinel Population of Canine and Feline Veterinary Practices in the United Kingdom.” *The Veterinary Journal* 224 (June): 18–24. doi:10.1016/j.tvjl.2017.03.010.

Small, Sharon Gower, and Larry Medsker. 2013. “Review of Information Extraction Technologies and Applications.” *Neural Computing and Applications* 25 (3-4). Springer London: 533–48. doi:10.1007/s00521-013-1516-6.

Sohn, Sunghwan, Jean-Pierre A Kocher, Christopher G Chute, and Guergana K Savova. 2011. “Drug Side Effect Extraction From Clinical Narratives of Psychiatry and Psychology Patients.” *J Am Med Inform Assoc* 18 (Supplement_1). Oxford University Press: i144–49. doi:10.1136/amiajnl-2011-000351.

Song, L. 2010. “The Role of Context in Discourse Analysis.” *Journal of Language Teaching and Research* 1 (6): 876–79. http://www.academypublication.com/issues/past/jltr/vol01/06/19.pdf.

Spache, George. 1953. “A New Readability Formula for Primary-Grade Reading Materials.” *The Elementary School Journal* 53 (7). University of Chicago Press: 410–13. doi:10.1086/458513.

Spencer, Stephen Andrew. 2016. “Future of Clinical Coding..” *Bmj* 353 (May): i2875.

Spyridakis, Jan H. 2000. “Guidelines for Authoring Comprehensible Web Pages and Evaluating Their Success.” Society for Technical Communication.

Steinbusch, Paul J M, Jan B Oostenbrink, Joost J Zuurbier, and Frans J M Schaepkens. 2007. “The Risk of Upcoding in Casemix Systems: a Comparative Study.” *Health Policy* 81 (2-3): 289–99. doi:10.1016/j.healthpol.2006.06.002.

Steindal, Simen A, Liv Wergeland Sørbye, Inger Schou Bredal, and Anners

Lerdal. 2012. “Agreement in Documentation of Symptoms, Clinical Signs, and Treatment at the End of Life: a Comparison of Data Retrieved From Nurse Interviews and Electronic Patient Records Using the Resident Assessment Instrument for Palliative Care..” *Journal of Clinical Nursing* 21 (9-10). Wiley/Blackwell (10.1111): 1416–24. doi:10.1111/j.1365-2702.2011.03867.x.

Stephens, M J, D G O'Neill, D B Church, P D McGreevy, P C Thomson, and D C Brodbelt. 2014. “Feline Hyperthyroidism Reported in Primary-Care Veterinary Practices in England: Prevalence, Associated Factors and Spatial Distribution..” *The Veterinary Record* 175 (18): 458–58. doi:10.1136/vr.102431.

Stevenson, Jean E, Johan Israelsson, Gunilla C Nilsson, Göran I Petersson, and Peter A Bath. 2014. “Recording Signs of Deterioration in Acute Patients: the Documentation of Vital Signs Within Electronic Health Records in Patients Who Suffered in-Hospital Cardiac Arrest.” *Health Informatics Journal* 22 (1): 21–33. doi:10.1177/1460458214530136.

Struchen, Rahel, Martin Reist, Jakob Zinsstag, and Flavie Vial. 2015. “Investigating the Potential of Reported Cattle Mortality Data in Switzerland for Syndromic Surveillance.” *Preventive Veterinary Medicine* 121 (1-2). Elsevier: 1–7. doi:10.1016/j.prevetmed.2015.04.012.

Swales, J. 1990. “The Concept of Discourse Community.” *Genre Analysis: English in Academic and ….*

Sweeney, L. 1996. “Replacing Personally-Identifying Information in Medical Records, the Scrub System..” *Proc AMIA Annu Fall Symp*, 333–37.

Szarvas, G, R Farkas, and R Busa-Fekete. 2007a. “State-of-the-Art Anonymization of Medical Records Using an Iterative Machine Learning Framework.” *J Am Med Inform Assoc* 14 (5): 574–80. doi:10.1197/j.jamia.M2441.

Szarvas, György, Richárd Farkas, and Róbert Busa-Fekete. 2007b. “State-of-the-Art Anonymization of Medical Records Using an Iterative Machine Learning Framework.” *J Am Med Inform Assoc* 14 (5). Oxford University Press: 574–80. doi:10.1197/j.jamia.M2441.

Taira, R K, AAT Bui, and H Kangarloo. 2002a. “Identification of Patient Name References Within Medical Documents Using Semantic Selectional Restrictions..” *Proceedings of the AMIA ….*

Taira, Ricky K, Alex A T Bui, and Hooshang Kangarloo. 2002b. “Identification of Patient Name References Within Medical Documents Using Semantic Selectional Restrictions..” *Proc AMIA Symp*. American Medical Informatics Association, 757–61.

Taylor, W A. 2000. “Change-Point Analysis: a Powerful New Tool for Detecting Changes.”

Teutsch, S M, and R E Churchill. 2000. *Principles and Practice of Public Health Surveillance*. 2nd ed. Oxford University Press.

Thacker, S B, and R.L. Berkelman. 1988. “Public Health Surveillance in the United States.” *Epidemiologic Reviews* 10: 164–90.

Thacker, S B, and R.L. Berkelman. 1992. “History of Public Health Surveillance.” In *Public Health Surveillance*, edited by W Halperin and E L Baker, 1–15.

Thomas, Sean M, Burke Mamlin, Gunther Schadow, and Clement McDonald. 2002. “A Successful Technique for Removing Names in Pathology Reports Using an Augmented Search and Replace Method..” *Proc AMIA Symp*, 777–81.

Thomas-Bachli, Andrea L, David L Pearl, Robert M Friendship, and Olaf Berke. 2014. “Exploring Relationships Between Whole Carcass Condemnation Abattoir Data, Non-Disease Factors and Disease Outbreaks in Swine Herds in Ontario (2001–2007).” *BMC Research Notes* 7 (1). BioMed Central: 185. doi:10.1186/1756-0500-7-185.

Thompson, Ken. 1968. “Programming Techniques: Regular Expression Search Algorithm.” *Commun. ACM* 11 (6): 419–22. doi:10.1145/363347.363387.

Thorsen, H, K Witt, H Hollnagel, and K Malterud. 2001. “The Purpose of the General Practice Consultation From the Patient's Perspective--Theoretical Aspects..” *Family Practice* 18 (6): 638–43.

Todd, Stacy, Peter J Diggle, Peter J White, Andrew Fearne, and Jonathan M Read. 2014. “The Spatiotemporal Association of Non-Prescription Retail Sales with Cases During the 2009 Influenza Pandemic in Great Britain..” *BMJ Open* 4 (4). British Medical Journal Publishing Group: e004869. doi:10.1136/bmjopen-2014-004869.

Todkill, Dan, Paul Loveridge, Alex J Elliot, Roger A Morbey, Obaghe Edeghere, Tracy Rayment-Bishop, Chris Rayment-Bishop, John E Thornes, and Gillian Smith. 2017. “Utility of Ambulance Data for Real-Time Syndromic Surveillance: a Pilot in the West Midlands Region, United Kingdom..” *Prehospital and Disaster Medicine* 32 (6). Cambridge University Press: 667–72. doi:10.1017/S1049023X17006690.

Tolentino, H D, M D Matters, W Walop, B Law, W Tong, F Liu, P Fontelo, K Kohl, and D C Payne. 2007. “A UMLS-Based Spell Checker for Natural Language Processing in Vaccine Safety.” *BMC Medical Informatics and Decision Making* 7.

Tomanek, Katrin, Joachim Wermter, and Udo Hahn. 2007. “A Reappraisal of Sentence and Token Splitting for Life Sciences Documents..” *Stud Health Technol Inform* 129 (Pt 1): 524–28.

Torres, G, V Ciaravino, S Ascaso, V Flores, L Romero, and F Simón. 2015. “Syndromic Surveillance System Based on Near Real-Time Cattle Mortality Monitoring..” *Preventive Veterinary Medicine* 119 (3-4): 216–21. doi:10.1016/j.prevetmed.2015.03.003.

Trace, D, F Naeymi-Rad, D Haines, J J Robert, F deSouza Almeida, L Carmony, and M Evans. 1993. “Intelligent Medical Record--Entry (IMR-E)..” *Journal of Medical Systems* 17 (3-4). UNITED STATES: Kluwer Academic/Plenum Publishers: 139–51.

Travers, Debbie, Stephanie W Haas, Anna E Waller, Todd A Schwartz, Javed Mostafa, Nakia C Best, and John Crouch. 2013. “Implementation of Emergency Medical Text Classifier for Syndromic Surveillance..” *AMIA Annu Symp Proc* 2013: 1365–74.

Triple S Project. 2011. “Assessment of Syndromic Surveillance in Europe..” *Lancet (London, England)* 378 (9806): 1833–34. doi:10.1016/S0140-6736(11)60834-9.

Tsui, Fu-Chiang, Jeremy U Espino, Virginia M Dato, Per H Gesteland, Judith Hutman, and Michael M Wagner. 2003. “Technical Description of RODS: a Real-Time Public Health Surveillance System.” *Journal of the American Medical Informatics Association* 10 (5): 399–408. doi:10.1197/jamia.M1345.

Tu, Karen, Julie Klein-Geltink, Tezeta F Mitiku, Chiriac Mihai, and Joel Martin. 2010. “De-Identification of Primary Care Electronic Medical Records Free-Text Data in Ontario, Canada.” *BMC Medical Informatics and Decision Making* 10 (1): 1–7. doi:10.1186/1472-6947-10-35.

Tudor Hart, Julian. 1971. “The Inverse Care Law.” *The Lancet* 297 (7696): 405–12. doi:10.1016/S0140-6736(71)92410-X.

Tulloch, J S P, L McGinley, F Sánchez-Vizcaíno, J M Medlock, and A D Radford. 2017. “The Passive Surveillance of Ticks Using Companion Animal Electronic Health Records..” *Epidemiology and Infection* 145 (10): 2020–29. doi:10.1017/S0950268817000826.

Turchin, A, J T Chu, and M Shubina. 2007. “Identification of Misspelled Words Without a Comprehensive Dictionary Using Prevalence Analysis.” *AMIA Annual Symposium*

U.S. Department of Health and Human Services. 1996. *The Health Insurance Portability and Accountability Act (HIPAA)*. http://purl.fdlp.gov/GPO/gpo10291.

UK Parliament. 1998. *Data Protection Act*.

UK Parliament. 2017a. *Data Protection Bill [HL]*. Edited by Department for Digital, Culture, Media and Sport. https://publications.parliament.uk/pa/bills/cbill/2017-2019/0153/18153.pdf.

UK Parliament. 2017b. *ILR Specification 2017 to 2018 Appendix C - Valid Postcode Format: Version 1 (28 April 2017)*. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/283357/ILRSpecification2013_14Appendix_C_Dec2012_v1.pdf.

University of Liverpool. 2017. “The Small Animal Veterinary Surveillance Network (SAVSNET).”

University of Pittsburgh. 2016. “RODS Laboratory Real-Time Outbreak and Disease Surveillance at the Department of Bioinformatics.” *Rods.Pitt.Edu*. http://www.rods.pitt.edu/.

Uzuner, Özlem, Tawanda C Sibanda, Yuan Luo, and Peter Szolovits. 2008. “A De-Identifier for Medical Discharge Summaries.” *Artif Intell Med* 42 (1): 13–35. doi:10.1016/j.artmed.2007.10.001.

Van Der Zwaan, J, E T K Sang, and M De Rijke. 2007. “An Experiment in Automatic Classification of Pathological Reports.” In, 4594 LNAI:207–16.

Van Ginneken, A M. 2002. “The Computerized Patient Record: Balancing Effort and Benefit.” *International Journal of Medical Informatics* 65 (2): 97–119.

Van Rossum, Guido. 2009. “The History of Python

.” http://python-history.blogspot.co.uk/2009/01/brief-timeline-of-python.html.

Veldhuis, Anouk, Henriëtte Brouwer-Middelesch, Alexis Marceau, Aurélien Madouasse, Yves Van der Stede, Christine Fourichon, Sarah Welby, Paul Wever, and Gerdien van Schaik. 2016. “Application of Syndromic Surveillance on Routinely Collected Cattle Reproduction and Milk Production Data for the Early Detection of Outbreaks of Bluetongue and Schmallenberg Viruses.” *Preventive Veterinary Medicine* 124 (February). Elsevier: 15–24. doi:10.1016/j.prevetmed.2015.12.006.

VetCompass. 2017. “VetCompass: Health Surveillance for UK Companion Animals.” https://www.rvc.ac.uk/vetcompass.

Vial, Flavie, and Martin Reist. 2015. “Comparison of Whole Carcass Condemnation and Partial Carcass Condemnation Data for Integration in a National Syndromic Surveillance System: the Swiss Experience..” *Meat Science* 101 (March): 48–55. doi:10.1016/j.meatsci.2014.11.002.

Vilain, Pascal, Sophie Larrieu, Katia Mougin-Damour, Pierre-Jean Marianne Dit Cassou, Marc Weber, Xavier Combes, and Laurent Filleul. 2017. “Emergency Department Syndromic Surveillance to Investigate the Health Impact and Factors Associated with Alcohol Intoxication in Reunion Island..” *Emergency Medicine Journal : EMJ* 34 (6). BMJ Publishing Group Ltd and the British Association for Accident & Emergency Medicine: 386–90. doi:10.1136/emermed-2015-204987.

Virchow, R. 1858. “Die Pathologische Physiologie Und Die Pathologischen Institute.” *Archiv Für Pathologische Anatomie Und Physiologie Und Für Klinische Medicin* 13 (1). Springer-Verlag: 1–15. doi:10.1007/BF02674509.

Walker, David. 2014. “Idiopathic Cutaneous and Renal Glomerular Vasculopathy (Alabama Rot).” *Companion Animal* 19 (2). MA Healthcare London: 68–68. doi:10.12968/coan.2014.19.2.68.

Walker, David, Laura Holm, Ian Hawkins, and Rachel Cianciolo. 2014. “Suspected Idiopathic Cutaneous and Renal Glomerular Vasculopathy in Dogs..” *The Veterinary Record* 174 (5): 124–24. doi:10.1136/vr.g1174.

Walker, David, Laura Holm, Richard Newton, and Catherine O'Conner. 2015. “CRGV in Dogs Visiting the New Forest..” *The Veterinary Record* 176 (15). British Medical Journal Publishing Group: 392–92. doi:10.1136/vr.h1848.

Walker, David, Laura Holm, Rosanne Jepson, and Ludovic Pelligand. 2016. “Diagnosing CRGV in Dogs with Skin Lesions..” *The Veterinary Record* 178 (3): 74–74. doi:10.1136/vr.i210.

Walsh, Stephen H. 2004. *The Clinician's Perspective on Electronic Health Records and How They Can Affect Patient Care*. *Bmj*. Vol. 328. doi:10.1136/bmj.328.7449.1184.

Wang, Xiaoyan, George Hripcsak, Marianthi Markatou, and Carol Friedman. 2009. “Active Computerized Pharmacovigilance Using Natural Language Processing, Statistics, and Electronic Health Records: a Feasibility Study.” *J Am Med Inform Assoc* 16 (3). Oxford University Press: 328–37. doi:10.1197/jamia.M3028.

Wang, Yanshan, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, et al. 2018. “Clinical Information Extraction Applications: a Literature Review..” *Journal of Biomedical Informatics* 77 (January): 34–49. doi:10.1016/j.jbi.2017.11.011.

Wasserman, Richard C. 2011. “Electronic Medical Records (EMRs), Epidemiology, and Epistemology: Reflections on EMRs and Future Pediatric Clinical Research..” *Academic Pediatrics* 11 (4): 280–87. doi:10.1016/j.acap.2011.02.007.

Weeber, M, J G Mork, and A R Aronson. 2001. “Developing a Test Collection for Biomedical Word Sense Disambiguation..” *Proc AMIA Symp*, 746–50.

Weide, R L. 1998. *Carnegie Mellon Pronouncing Dictionary” Release 0.6*. Online: http://www. speech. cs. cmu. edu.

Wellner, B, M Huyck, S Mardis, J Aberdeen, A Morgan, L Peshkin, A Yeh, J Hitzeman, and L Hirschman. 2007. “Rapidly Retargetable Approaches to De-Identification in Medical Records.” *J Am Med Inform Assoc* 14 (5): 564–73. doi:10.1197/jamia.M2435.

Westheimer, Emily, Marc Paladini, Sharon Balter, Don Weiss, Anne Fine, and Trang Quyen Nguyen. 2012. “Evaluating the New York City Emergency Department Syndromic Surveillance for Monitoring Influenza Activity During the 2009-10 Influenza Season..” *PLoS Currents* 4 (August). Public Library of Science: e500563f3ea181. doi:10.1371/500563f3ea181.

WHO Global Observatory for eHealth. 2006. “Report of the WHO Global Observatory for eHealth: Building Foundations for eHealth.”

Wick, Marc. 2017. “GeoNames Geographic Database.” *GeoNames*.

Wilcox, Adam B, and George Hripcsak. 2003. “The Role of Domain Knowledge in Automating Medical Text Report Classification..” *Journal of the American Medical Informatics Association* 10 (4): 330–38. doi:10.1197/jamia.M1157.

Wilson, E B. 1927. “Probable Inference, the Law of Succession, and Statistical Inference.” *Journal of the American Statistical Association* 22 (158): 209–12. doi:10.1080/01621459.1927.10502953.

Witten, I H. 2005. “Text Mining.” In *Practical Handbook of Internet Computing*, edited by M P Singh, 14–1–14–22.

Woodmansey, David. 2018. “Ten New Cases of Alabama Rot Confirmed.” *Vet Times*, January 17.

World Health Organization. 2010. “International Statistical Classifcation of Diseases and Related Health Problems

10th Revision.” http://www.who.int/classifications/icd/icdonlineversions/en/.

World Health Organization. Dept. of Epidemic and Pandemic Alert and Response. 1999. “WHO Recommended Surveillance Standards.” Geneva: Geneva : World Health Organization. http://www.who.int/iris/handle/10665/65517.

World Organisation For Animal Health. 2018. “World Organisation for Animal Health (OIE) Protecting Animals, Preserving Our Future.” http://www.oie.int/.

World Organisation of Family Doctors. 2018. “ICPC-2e.” February 26. https://ehelse.no/icpc-2e-english-version.

Wu, T S, F Y Shih, M Y Yen, J S Wu, S W Lu, K C Chang, C Hsiung, et al. 2008. “Establishing a Nationwide Emergency Department-Based Syndromic Surveillance System for Better Public Health Responses in Taiwan.” *BMC Public Health* 8 (1): 18. doi:10.1186/1471-2458-8-18.

Yli-Hietanen, J, S Niiranen, M Aswell, and L Nathanson. 2009. “Domain-Specific Analytical Language Modeling--the Chief Complaint as a Case Study.” *Int J Med Inform* 78 (12): e27–e30. doi:10.1016/j.ijmedinf.2009.02.002.

Zhou, L, G B Melton, S Parsons, and G Hripcsak. 2006. “A Temporal Constraint Structure for Extracting Temporal Information From Clinical Narrative.” *Journal of Biomedical Informatics* 39 (4): 424–39. doi:10.1016/j.jbi.2005.07.002.

Ziemann, A, N Rosenkötter, L Garcia-Castrillo Riesgo, S Schrell, B Kauhl, G Vergeiner, M Fischer, et al. 2014. “A Concept for Routine Emergency-Care Data-Based Syndromic Surveillance in Europe..” *Epidemiology and Infection* 142 (11). Cambridge University Press: 2433–46. doi:10.1017/S0950268813003452.