# PetEVAL: A veterinary free text electronic health records benchmark

**Sean Farrell[1], Alan Radford[2], Noura Al Moubayed[1], Peter-John Mäntylä Noble[2],**

[1] Department of Computer Science, Durham University
[2] Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool
**Correspondence:** sean.farrell2@durham.ac.uk

## Abstract

We introduce PetEVAL, the first benchmark dataset derived from real-world, free-text veterinary electronic health records (EHRs). PetEVAL comprises 17,600 professionally annotated EHRs from first-opinion veterinary practices across the UK, partitioned into training (11,000), evaluation (1,600), and test (5,000) sets with distinct clinic distributions to assess model generalisability. Each record is annotated with International Classification of Disease 11 (ICD-11) syndromic chapter labels (20,408 labels), disease Named Entity Recognition (NER) tags (429 labels), and anonymisation NER tags (8,244 labels). PetEVAL enables evaluating Natural Language Processing (NLP) tools across applications, including syndrome surveillance and disease outbreak detection. We implement a multistage anonymisation protocol, replacing identifiable information with clinically relevant pseudonyms while establishing the first definition of identifiers in veterinary free text. PetEVAL introduces three core tasks: syndromic classification, disease entity recognition, and anonymisation. We provide baseline results using BERT-base, PetBERT, and LLaMA 3.1 8B generative models. Our experiments demonstrate the unique challenges of veterinary text, showcasing the importance of domain-specific approaches. By fostering advancements in veterinary informatics and epidemiology, we envision PetEVAL catalysing innovations in veterinary care, animal health, and comparative biomedical research through access to real-world, annotated veterinary clinical data.

## 1 Introduction

The growing availability of veterinary electronic health records (vEHRs) from sources such as the Small Animal Veterinary Surveillance Network (SAVSNET) (Sánchez-Vizcaíno et al., 2015), Companion Animal Veterinary Surveillance Network (CAVSNET) (Sheng et al., 2022), and VetCompass (Royal Veterinary College (RVC); McGreevy et al., 2017) presents an unprecedented opportunity to advance veterinary medicine. These datasets support disease surveillance, epidemiological research, and clinical decision-making (Farrell et al., 2023b; Bode et al., 2022; Radford et al., 2011; Sánchez-Vizcaíno et al., 2017; Singleton et al., 2020). However, vEHRs differ from human biomedical records in syntax, lexicon, and clinical expression (Davies et al., 2024b), requiring adaptation of existing computational tools. Additionally, first-opinion vEHRs often contain diagnostic uncertainty due to limited specialist access, resource constraints, and financial considerations (Robinson et al., 2016).


```
[
    {
        "savsnet_id": 1111025,
        "text": "Brought in with Coco who has
           conjunctivitis; fluo neg. no blepharospasm or
           rubbing.otherwise nad. Adv monitor, ini send to Smith
           Referrals",
        "icd_11_chapter": "Diseases of the visual system",
        "disease_ner_entities":[[12,36,"Conjunctivitis"]],
        "anonymisation_ner_entities": [[16,20,"PER"],
           [123,138,"LOC"]]
    }
]
```


Figure 1: Example data for a single consult with a unique consult, the free text clinical EHR, the ICD-11 chapter multi-label classification and NER entities for both anonymisation and disease extraction task

Despite these challenges, vEHRs offer unique advantages for biomedical research. Unlike human records, which are tightly regulated under laws such as HIPAA and GDPR, vEHRs face fewer legal constraints (Sun et al., 2020), making them a viable test bed for developing analytical methods. Their relative accessibility enables researchers to explore novel computational approaches without the ethical and regulatory barriers associated with human health data (Kol et al., 2015; Starkey et al.,



*Proceedings of the 24th Workshop on Biomedical Language Processing (BioNLP 2025)*, pages 341–353

2005; Trott et al., 2004).

Advancing natural language processing (NLP) for vEHRs is critical for global health, supporting the World Health Organisation's (WHO) One Health initiatives in zoonotic disease surveillance and antimicrobial resistance (AMR) monitoring (Bidaisee and Macpherson, 2014; Radford et al., 2011). Enhanced NLP tools improve threat detection and trend analysis in animal populations, strengthening public health responses across human, animal, and environmental health domains (Kol et al., 2015; Robertson et al., 2000; Van Duijkeren et al., 2004). Beyond public health, NLP-driven solutions facilitate large-scale epidemiological studies, identifying risk factors and treatment outcomes that enhance companion animal welfare (Lund, 2015; Farrell et al., 2023b).

Traditional veterinary disease surveillance relies on manual coding or rule-based methods, which are time-intensive and prone to human error (Hsia et al., 2010; Miñarro-Giménez et al., 2018; Turchin et al., 2006). In contrast, NLP-driven approaches offer scalable, automated solutions for extracting clinical insights from free-text records. Developing these methods within veterinary medicine improves animal welfare and contributes to the refinement of computational tools that may later be adapted to human bioinformatics research. Neural network approaches to disease coding have evolved considerably over time. Pioneering work introduced DeepTag (Nie et al., 2018), establishing a foundation that subsequently refined into the more advanced VetTag framework (Zhang et al., 2019). The field has progressed significantly with recent innovations leveraging pre-trained LLMs (Farrell et al., 2023a; Boguslav et al., 2024). Complementary research has expanded our understanding of generative models for veterinary entity extraction for clinical signs (Wulcan et al., 2024) and for body condition scoring (Fins et al., 2024).

In this paper, we contribute the following:

1. PetEVAL: The first veterinary EHR benchmark – A publicly available free-text vEHR dataset, establishing a standard for veterinary NLP research.

2. Rigorous manual anonymisation – Every record underwent manual anonymisation with at least two independent reviews, including verification by a veterinary clinician, ensuring complete removal of sensitive data.

3. ICD-11 syndromic classification – Syndromic labels were assigned using the ICD-11 framework, supplemented with domain-specific annotations to ensure clinically relevant labeling.

## 2 Literature Review

The adoption of EHRs has revolutionised medical research, offering vast amounts of health data for analysis (Gunter and Terry, 2005; Cowie et al., 2017). While structured EHR data has been extensively used in epidemiological studies (Krumholz et al., 2014; Hamer et al., 2024; Hlatky et al., 2014; Williamson et al., 2020), up to 80% of EHR information exists in unstructured formats, primarily as free-text clinical notes (Kong, 2019). These unstructured notes capture clinical insights often lost in structured formats (Birman-Deych et al., 2005; Singh et al., 2004). Excluding this data from research can significantly impact the validity of findings (Ford et al., 2013; Jensen et al., 2017; Price et al., 2016; Barak-Corren et al., 2017). However, utilising unstructured data presents challenges in patient privacy protection, particularly regarding re-identification risks (Simon et al., 2019; Abouelmehdi et al., 2017; Dorr et al., 2006). Automated EHR anonymisation has become a critical focus in addressing these challenges. Benchmarks like the i2b2/UTHealth corpus and MIMIC-3 database have been established to evaluate de-identification models (Stubbs and Uzuner, 2015; Stubbs et al., 2017; Meystre et al., 2010; Aberdeen et al., 2010). Approaches range from rule-based systems (Cao et al., 2003) to neural networks (Liu et al., 2019) and pre-trained language models (Yoon et al., 2023; Chen et al., 2021). Recent advancements in learning-based methods show promise in automating de-identification (Leevy et al., 2020; Lee et al., 2022). However, these methods face challenges with performance instability when applied to heterogeneous real-world data (Abu-El-Rub et al., 2022; Yang et al., 2019). Deep learning approaches have been proposed to address these issues, but their effectiveness is limited by small training datasets and performance degradation on out-of-distribution EHRs (Syed et al., 2022; Lee et al., 2021; Jiang et al., 2017).

## 3 PetEVAL

### 3.1 The SAVSNET Dataset

We utilise data from the Small Animal Veterinary Surveillance Network (SAVSNET), a sentinel network of 253 volunteer first-opinion veterinary practices across the United Kingdom that have collected vEHRs since March 2014. This network has accumulated over 12 million EHRs, with participating practices selected based on their practice management software compatibility with the SAVSNET data exchange system. During each consultation with a clinician or nurse, comprehensive data includes species, breed, sex, neuter status, age, owner's postcode, insurance and microchipping status, and a detailed free-text clinical narrative. These narratives may contain information about symptoms, diagnoses, treatments, procedures, or other clinical matters. Owners can opt out of data collection during any consultation. The SAVSNET group operates under ethical approval from the University of Liverpool Ethics Committee (RETH001081), ensuring adherence to established ethical standards. Figure 1 provides a sample data point in JSON format.

### 3.2 Tasks

#### 3.2.1 Task 1 - Anonymisation

Ensuring the privacy and security of EHRs is crucial for safeguarding the personal information of pet owners and facilitating the easy sharing of data use in clinical and academic research. The dataset is labelled with NER entities and spans applied to pseudo-anonymised contextual placeholders. The objective is to maintain the integrity and utility of clinical information within the EHR while effectively anonymising various types of personal data. This includes names (both animal and human), location details (such as city, town, and addresses), organisation names (including attending veterinary practices, referral hospitals, kennels, and laboratories), contact details (emails, phone numbers), id-numbers (passport numbers, insurance policy numbers, MRCVS codes), and any other explicit identifiers. The anonymisation is compliant with the HIPPA Safe Harbour (Sun et al., 2020).

#### 3.2.2 Task 2 - Syndromic Disease Classification

Given the critical role of monitoring national disease outbreaks in public health, effective surveillance systems can provide invaluable insights, such as in informing clinicians of key symptoms to observe, enabling researchers to identify aetiological agents, and establishing an automated reporting mechanism for public health agencies to facilitate swift notification of changes in disease occurrence. However, the task is not straightforward, particularly when dealing with novel diseases or syndromes with unknown symptoms. Effective outbreak reduction strategies hinge on the ability to detect outbreaks with minimal cases. To address these challenges, the dataset is provided with ICD-11 chapters (World Health Organisation (WHO), 2022), which includes contextual discussions such as symptoms and diagnoses. The task is structured as a multi-label classification problem, as a consult or condition may cover a range of presenting symptoms. Performance is evaluated using multi-label classification metrics, including precision and recall, macro-average F1-Score, and weighted F1-Score.

#### 3.2.3 Task 3 - Disease Extraction

Identifying specific diseases is critical for downstream epidemiological studies, which aim to reveal novel risk factors, seasonality, and other trends. This task is particularly challenging due to the private healthcare nature of veterinary practices in the UK and much of the world. Confirmation diagnostic tests are rare, as owners often wish to avoid the inherent costs, opting instead to take the advice of clinicians or due to the lack of available resources or expertise not found in first opinion practice. Additionally, the presence of negations is common within vEHRs, especially within the first opinion setting, where it is estimated that 11% of mentioned diseases are negated (Cheng et al., 2017) which complicates the task further. In our study, the dataset is labelled with the diagnostic disease contained within it. This process is framed as NER task using the IOB2 format, wherein the entity of 'disease' and its spans are provided. Evaluation utilises SeqEval for precision, recall, and F1-score (Nakayama, 2018).

## 4 Methods

### 4.1 Dataset Construction

Our dataset comprises three subsets: a training set of 11,000 records, an evaluation set of 1,600 records, and a test set of 5,000 records. We selected only consultations recorded before 2020 and restricted the dataset to consultations involving only

Table 1: Evaluation of NER performance on veterinary clinical text data anonymised according to HIPAA Safe Harbor guidelines. The table presents entity type distribution across training, evaluation, and test splits, with comparative performance metrics (precision, recall, F1-score) between 'BERT-base-uncased', 'PetBERT', and LLaMA 3.1 8B models across identifier categories.

| HIPAA Safe Harbor | Examples | Train/ Eval | Test Count | NER Entity | BERT-base-uncased | | | PetBERT | | | LLaMA 3.1 8B | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | P | R | F1 | P | R | F1 | P | R | F1 |
| (A) Names | Pet, Owner, Vet Names | 4790 | 1370 | PER | 0.84 | 0.93 | 0.89 | 0.93 | 0.70 | 0.80 | 0.71 | 0.65 | 0.68 |
| (B) Geographic | City, Towns, Countries | 311 | 94 | LOC | 0.95 | 0.98 | 0.97 | 0.97 | 0.97 | 0.97 | 0.78 | 0.83 | 0.80 |
| subdivisions | Vet practices, hospitals, shelters | 392 | 168 | ORG | 0.97 | 0.97 | 0.98 | 0.98 | 0.96 | 0.97 | 0.82 | 0.79 | 0.81 |
| (C) Dates | Day/month dates, appointments | 425 | 162 | TIME | 0.94 | 0.96 | 0.95 | 0.93 | 0.94 | 0.93 | 0.76 | 0.81 | 0.78 |
| (D) Telephone numbers | Client/practice phone numbers | 19 | 4 | | | | | | | | | | |
| (E) Fax numbers | n/a | None | None | | | | | | | | | | |
| (F) Email addresses | Referral/client emails | 9 | 3 | | | | | | | | | | |
| (G) Social security numbers | n/a | None | None | | | | | | | | | | |
| (H) Medical record numbers | n/a | None | None | | | | | | | | | | |
| (I) Health plan numbers | Insurance policy numbers | 33 | 20 | | | | | | | | | | |
| (J) Account numbers | Microchip Numbers | 299 | 35 | | | | | | | | | | |
| (K) Certificate numbers | MRCVS clinician codes | 51 | 17 | MISC | 0.91 | 0.97 | 0.97 | 0.95 | 0.94 | 0.94 | 0.73 | 0.69 | 0.71 |
| (L) Vehicle identifiers | n/a | None | None | | | | | | | | | | |
| (M) Device identifiers | n/a | None | None | | | | | | | | | | |
| (N) URLs | Website urls | None | None | | | | | | | | | | |
| (O) IP addresses | n/a | None | None | | | | | | | | | | |
| (P) Biometric identifiers | n/a | None | None | | | | | | | | | | |
| (Q) Photographic images | n/a | None | None | | | | | | | | | | |
| (R) Other identifiers | Passport numbers | 34 | 8 | | | | | | | | | | |

cats and dogs. To enhance generalisability, dataset splits were performed based on a pre-compiled list of veterinary practices, following the methodology outlined in (Farrell et al., 2023a). Specifically, we assigned distinct practices to training and testing sets, ensuring that models trained on the training set were evaluated on records from veterinary practices that did not contribute to training. This design minimises the risk of models overfitting to stylistic or institutional biases and provides more substantial evidence of generalisability across UK veterinary practices. We excluded empty records containing fewer than ten words or exceeding 350 words. The median narrative length in the full SAVSNET dataset is 287 words, while in PetEVAL, it is 226 words.

#### 4.1.1 Annonymisation

Each record was manually reviewed twice, targeting the removal of all potential identifiers, including names (owner, animal, and veterinary staff), locations (cities, countries, vet practices, referral hospitals, rescue centres, kennels, crematoriums, labs), dates (when they included specific years), times (if overly specific), and unique identifiers such as microchip codes, passport numbers, insurance policy numbers, vet MRCVS codes, phone numbers, and email addresses. Flagged elements were pseudonymised with context-appropriate placeholders to maintain record coherence, and corresponding spans and entity tags were generated for these placeholders. Pseudonyms were derived from separate lists for train and test splits,

For the anonymisation NER task, identifiers were mapped to standard tags: 'LOC' (cities, towns, countries), 'PER' (pet/owner/vet names), 'TIME' (specific dates/times), 'ORG' (veterinary practices, rescue shelters, labs, groomers), and 'MISC' (unique identifiers like microchips, insurance codes, contact information). The counts for each can be found within table 1. Non-clinical brand names were removed but not included in anonymisation metrics. No clinically relevant information was modified.

#### 4.1.2 Syndromic Disease Classification

The dataset was curated to support syndromic disease surveillance through the assignment of ICD-11 labels. For this purpose, 20 ICD-11 chapter codes were selected to capture a broad range of clinically relevant syndromes observed in veterinary practice. The full list of selected chapter codes is provided in Table 2. To facilitate efficient and accurate annotation, we employed a semi-automated approach wherein initial fuzzy labels were generated using the PetBERT-ICD model, a previously developed tool designed for assigning ICD-related labels in veterinary contexts. This pre-annotation step helped streamline the annotation process, reduce cognitive load for annotators, and minimise potential errors. Annotators reviewed and refined these suggested labels, ensuring alignment with clinical documentation practices in first-opinion vEHRs. To maintain the integrity of the evalua-

tion, the test set was exempt from automated label matching and underwent a full manual review by two expert annotators. Records that an initial reviewer was unhappy to determine the presence of a diagnosis were passed through an additional reviewer, and a consensus vote was taken. Finally, we ensured that the disease extraction dataset aligned with the syndromic dataset, an extracted disease therefore has a linked syndromic label.

#### 4.1.3 Disease Extraction

The dataset was developed to facilitate the evaluation of disease diagnosis extraction models from first-opinion vEHRs. Given the nature of primary care veterinary records, confirmatory diagnoses are rare, with most diagnoses being clinical assessments rather than definitive results from diagnostic testing. Therefore, any named condition mentioned in a record was annotated as a diagnosis unless explicitly negated. This includes confirmed diagnoses, differential diagnoses, and syndromic descriptions. Additionally, mentions of pathogens, such as bacteria, viruses, and parasites, were annotated as they typically are discussed as diagnoses within the narratives. We extracted diseases coded within the ICD-11 and veterinary-specific conditions not represented in human medicine. Each annotated diagnosis was linked to its corresponding span within the text, with entity tags assigned to support NER tasks. Records that an initial reviewer was unhappy to determine the presence of a diagnosis were passed through an additional reviewer, and a consensus vote was taken.

#### 4.1.4 Baseline Models

For baseline results in PetEVAL, we evaluated three pre-trained language models: 'BERT-base-uncased' (Devlin et al., 2019), a general-purpose encoder; 'PetBERT' (Farrell et al., 2023a), a veterinary domain-adapted encoder; and 'LLaMA 3.1 8B' (Team and Meta, 2024), a generalist decoder model. The encoder models were fine-tuned as token classification models using the IOB2 format for the anonymisation and disease extraction tasks, with training parameters including a mini-batch size of 32, an initial learning rate of 2e-5, and the AdamW optimiser. Early stopping was applied based on evaluation loss. For syndromic classification, both encoders were adapted for multi-label classification across 20 ICD-11 chapter codes, employing a weighted binary cross-entropy loss function with sigmoid activation to address class imbal-

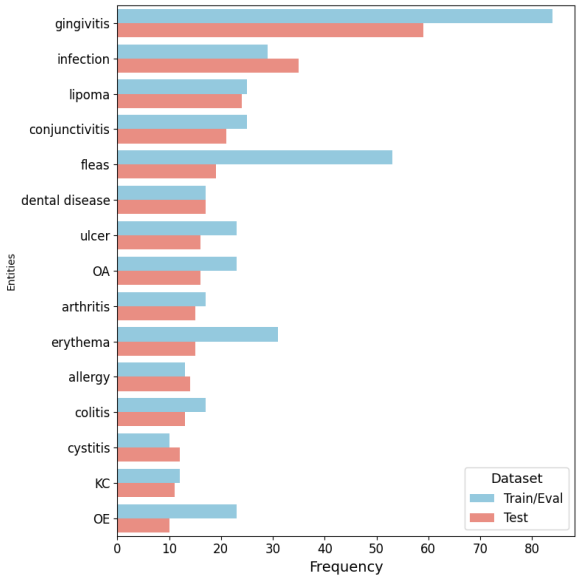


Figure 2: Distribution of the 15 most frequent disease entities extracted from veterinary electronic health records in the Train/Eval and Test sets during Task 2 (Disease Extraction).

ance. Training followed the same hyperparameter setup and typically converged beyond epoch 6. An iterative threshold analysis was conducted, varying classification thresholds between 60% and 95% in 5% increments, prioritising recall to minimise false negatives. The final classifier applied an 80% threshold and was evaluated on the test set. The decoder model was prompted with few-shot examples selected from the training set, with multiple prompt designs tested against the evaluation set before application to the full test set.

#### 4.1.5 Model Evaluation

We implemented a unified entity-level evaluation framework to ensure fair comparison between encoder (BERT) and decoder (LLaMA) architectures across anonymisation and disease extraction tasks. For encoder models, we first converted token-level IOB/BIO predictions into entity spans before applying the same entity-level F1 evaluation used for decoder models. This approach follows CoNLL methodology (Tjong et al., 2003), where all extracted entities undergo identical normalisation procedures before being exact-matched against ground truth. For both model types and tasks, we calculate precision as the ratio of correctly identified entities to total predictions, recall as the ratio of correctly identified entities to ground truth entities, and F1 as their harmonic mean. The anonymisation task evaluates the identification of privacy-sensitive entities

(LOC, PER, MISC, NAME), while disease extraction assesses the recognition of standardised disease mentions. By standardising evaluation across architectural paradigms, we enable direct performance comparison while maintaining methodological rigour in assessing clinical information extraction capabilities.

For the syndromic classification task, we assess model performance using precision, recall, and F1 scores computed against ground truth labels provided by annotators. For encoder-based models, classification uses a fine-tuned ICD-11 classifier with an optimised threshold, ensuring a balance between precision and recall for robust disease detection. For generative models, we convert outputs into a tabular format using a direct match approach on uncased text. Similarity-based methods were considered, but they yielded no performance gains, so we adopted the least computationally intensive approach. The predicted labels are transformed into a one-hot encoded vector, applying the same evaluation metrics as encoder models. Given the importance of disease surveillance, we preferentially select for recall to minimise false negatives, as missing cases could lead to undetected outbreaks. While this may increase false positives, these can be further reviewed to ensure the detection of potential health threats.

## 5 Results

### 5.1 Corpus Overview

The dataset consists of 675,935 words distributed across the training (11,000 records), evaluation (1,600 records), and test sets (5,000 records). While demographic data is not included, 68% of the records represent dogs, with a near 50-50 sex split across both species. The dataset contains information from 16,153 unique animals from various regions across the UK.

For syndromic disease classification, annotations were applied using a multi-label one-hot encoding approach aligned with ICD-11 chapter heads. Across the dataset, 9,510 annotations were made in the training set and 4,714 in the test set. The most frequent label, 'Certain infectious or parasitic diseases', was prominent due to the high occurrence of conditions like parasitic infestations. The median labels per class in the training set was 348, with an average of 0.9 labels per consultation. Notably, 8,907 consultations received at least one label, while those without a label typically represented routine checkups or non-syndromic cases.

The frequency distribution of extracted disease entities across the train/eval and test datasets is presented in Figure 2. As expected, conditions readily identifiable through visual examination, such as gingivitis, conjunctivitis, and lipoma, exhibit high representation. Furthermore, the extracted entities encompass clinical language commonly used by veterinary practitioners to indicate disease, including terms like 'infection,' 'fleas' (for flea infestation), and 'dental disease' (for unspecified dental conditions). The train/eval datasets contain 3,907 unique extracted conditions, while the test dataset comprises 2,899.

### 5.2 Inter-annotator agreement

Inter-annotator agreement was assessed on a subset of 1,000 vEHRs from the test set focused on the syndromic classification task. Two expert veterinary clinicians independently annotated the records using strictly predefined guidelines, with no communication allowed at this stage to ensure unbiased annotations. The resulting Cohen's kappa statistic was 0.722, indicating a substantial level of agreement (McHugh, 2012). This value suggests strong, though not perfect, alignment between the annotators. Disagreements were systematically reviewed, with the majority resolved through a collaborative discussion. In cases where consensus could not be reached, a third clinician provided a decisive resolution.

### 5.3 Baselines

We conducted baseline experiments with 'bert-base-uncased' and 'PetBERT' and a generative model 'LLaMA 3.1 8B' to establish reference points for evaluating more complex models. For the anonymisation task, PetBERT consistently outperformed BERT-Base across HIPAA Safe Harbor entity categories, with notable improvements in identifying names (F1: 0.80 vs. 0.89) and geographic subdivisions (F1: 0.97 vs. 0.98) (Table 1). Both models achieved high performance in structured entity types such as dates (F1: 0.93 vs. 0.95) and organisations (F1: 0.97 vs. 0.98). LlaMA 3.1, using few-shot prompting (Appendix), was behind with lower F1-scores across all categories, particularly for names (F1: 0.68) and locations (F1: 0.80).

As shown in Table 1, fine-tuned PetBERT outperformed BERT-base-uncased across most entity types, achieving a higher precision (0.93 vs. 0.84), recall (0.70 vs. 0.93), and F1-score (0.80 vs. 0.89)

Table 2: Performance Metrics for BERT-base-uncased, PetBERT, and LLaMA 3.1 8B on ICD-11 Syndromic Chapters. P = Precision, R = Recall, F1 = F1-score

| ICD-11 Syndromic Chapter | Train/ Eval | Test Count | BERT-base-uncased | | | PetBERT | | | LLaMA 3.1 8B | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | P | R | F1 | P | R | F1 | P | R | F1 |
| **Certain infectious or parasitic diseases** | 1549 | 1321 | 0.74 | 0.31 | 0.44 | 0.78 | 0.45 | 0.57 | 0.65 | 0.28 | 0.39 |
| **Neoplasms** | 774 | 499 | 0.85 | 0.77 | 0.81 | 0.90 | 0.81 | 0.85 | 0.77 | 0.65 | 0.70 |
| **Diseases of the blood or blood-forming organs** | 90 | 47 | 0.66 | 0.35 | 0.45 | 0.63 | 0.31 | 0.41 | 0.55 | 0.23 | 0.32 |
| **Diseases of the immune system** | 512 | 429 | 0.80 | 0.54 | 0.64 | 0.84 | 0.51 | 0.64 | 0.68 | 0.41 | 0.51 |
| **Endocrine, nutritional or metabolic diseases** | 572 | 305 | 0.67 | 0.60 | 0.64 | 0.69 | 0.60 | 0.64 | 0.58 | 0.45 | 0.51 |
| **Mental, behavioral or neurodevelopmental disorders** | 1121 | 469 | 0.76 | 0.34 | 0.46 | 0.79 | 0.38 | 0.51 | 0.64 | 0.27 | 0.38 |
| **Diseases of the nervous system** | 233 | 150 | 0.54 | 0.58 | 0.56 | 0.71 | 0.54 | 0.61 | 0.48 | 0.42 | 0.45 |
| **Diseases of the visual system** | 905 | 634 | 0.85 | 0.81 | 0.83 | 0.90 | 0.80 | 0.85 | 0.73 | 0.68 | 0.70 |
| **Diseases of the ear or mastoid process** | 700 | 513 | 0.83 | 0.77 | 0.80 | 0.88 | 0.78 | 0.83 | 0.71 | 0.65 | 0.68 |
| **Diseases of the circulatory system** | 276 | 181 | 0.67 | 0.33 | 0.45 | 0.71 | 0.46 | 0.55 | 0.55 | 0.29 | 0.38 |
| **Diseases of the respiratory system** | 459 | 346 | 0.80 | 0.54 | 0.64 | 0.84 | 0.57 | 0.68 | 0.68 | 0.45 | 0.54 |
| **Diseases of the digestive system** | 671 | 259 | 0.81 | 0.55 | 0.66 | 0.79 | 0.62 | 0.69 | 0.67 | 0.46 | 0.55 |
| **Diseases of the skin** | 1377 | 1018 | 0.81 | 0.62 | 0.70 | 0.88 | 0.60 | 0.71 | 0.69 | 0.51 | 0.59 |
| **Diseases of the musculoskeletal system or connective tissue** | 1171 | 722 | 0.79 | 0.73 | 0.76 | 0.83 | 0.70 | 0.76 | 0.67 | 0.61 | 0.64 |
| **Diseases of the genitourinary system** | 569 | 334 | 0.76 | 0.59 | 0.66 | 0.79 | 0.67 | 0.73 | 0.65 | 0.49 | 0.56 |
| **Pregnancy, childbirth or the puerperium** | 65 | 36 | 0.42 | 0.17 | 0.24 | 0.74 | 0.12 | 0.21 | 0.36 | 0.10 | 0.16 |
| **Certain conditions originating in the perinatal period** | 39 | 27 | 0.50 | 0.08 | 0.13 | 0.00 | 0.00 | 0.00 | 0.38 | 0.05 | 0.09 |
| **Developmental anomalies** | 191 | 95 | 0.59 | 0.19 | 0.28 | 0.70 | 0.30 | 0.42 | 0.47 | 0.15 | 0.23 |
| **Injury, poisoning or certain other consequences of external causes** | 1113 | 636 | 0.67 | 0.67 | 0.67 | 0.73 | 0.70 | 0.71 | 0.58 | 0.55 | 0.56 |
| **micro average** | | | **0.76** | **0.58** | **0.66** | **0.81** | **0.61** | **0.70** | **0.65** | **0.47** | **0.55** |
| **macro average** | | | **0.71** | **0.50** | **0.57** | **0.74** | **0.52** | **0.60** | **0.60** | **0.41** | **0.48** |
| **weighted average** | | | **0.76** | **0.58** | **0.65** | **0.81** | **0.61** | **0.69** | **0.65** | **0.47** | **0.54** |

for identifying personal names (PER) such as pet, owner, and vet names. In contrast, LLaMA 3.1 achieved lower performance across all entity types, with an F1-score of 0.68 for names. For location (LOC) and organisation (ORG) entities, PetBERT outperformed BERT-base-uncased, achieving F1-scores of 0.97 and 0.97, respectively, compared to BERT-base's 0.97 and 0.98. LLaMA 3.1 showed lower performance in both entity types, with an F1 of 0.80 for LOC and 0.81 for ORG. The comparison highlights PetBERT's superior ability to process veterinary clinical text, particularly for identifying personal and organisational entities, while Llama 3.1's performance in entity recognition remained behind.

PetBERT outperformed both BERT-Base and Llama 3.1 for the disease extraction task, achieving a precision of 0.90, recall of 0.85, and F1-score of 0.87 (Table 1). BERT-Base trailed with 0.70 precision, 0.55 recall, and an F1 of 0.60, while Llama 3.1, using a few-shot prompt (Appendix), performed worst (precision: 0.60, recall: 0.35, F1: 0.40).

## 6 Discussion

In veterinary first-opinion clinical practice, the challenge of extracting meaningful insights from vEHRs is compounded by several notable factors. Among these is the absence of standardised data conventions within free-text inputs, and inconsistencies in spelling and abbreviations used by different clinicians (Davies et al., 2024b). This is amplified by the ambiguity surrounding the interpretation of consultation events. Specifically, the lack of diagnostic details in these narratives introduces additional layers of complexity. The moderate Cohen's kappa score of 0.7, observed between two annotators—both qualified veterinary clinicians - underscores the inherent difficulties in annotating such unstructured data. Veterinary EHRs are packed with ambiguous language, clinician-specific abbreviations, and varying documentation styles, inhibiting the ability to extract information from them effectively. Even among active clinicians, the interpretation of nuanced first-opinion notes can differ, primarily due to diagnostic uncertainties, incomplete patient histories, and the lack of standardised terminology. Despite these obstacles, the intrinsic value embedded within these clinical narratives is undeniable, with applications spanning disease outbreak detection and improving public health and animal welfare standards (Davies et al., 2024a; Farrell et al., 2023a).

Generative models, such as the LLaMA 3.1 8B applied in our baseline, exhibited relatively poor performance across tasks, particularly in NER. This highlights the ongoing challenge of designing effective prompting strategies, requiring further research. Additionally, generative models present inherent difficulties in evaluation, as their flexible outputs may not align precisely with gold-standard annotations. While our strict direct match approach may penalise performance, maintaining fidelity to the intended identifier remains a priority. Over

time, we anticipate improvements in generative architectures, which may eventually surpass the limitations observed here. However, domain-adapted encoder-based models like PetBERT demonstrated superior performance across all tasks, aligning with expectations given their targeted pretraining. Beyond accuracy, their efficiency also makes them preferable for everyday deployments, especially in resource-intensive applications such as continuous disease surveillance. Given the significant environmental cost of running large LLMs (Bashir et al., 2024), there is a clear need for lightweight, domain-specific solutions that can operate effectively on consumer-level hardware, ensuring sustainability and practical usability in real-world veterinary informatics.

Strict privacy regulations in human healthcare restrict many studies to single institutions, creating discrepancies between reported performance and cross-site generalisability. PetEVAL collates from over 250 UK practices with diverse clinical approaches and provides substantial advantages for robust model evaluation. While fewer than 23% of human healthcare ML studies utilise multi-institutional data (McDermott et al., 2021), often resulting in significant biases and performance degradation when applied to external institutions (Barak-Corren et al., 2021; Burns and Kheterpal, 2020), PetEVAL's multi-institutional framework can capture practice variability and thus offers an opportunity to assess model robustness across institutions, ultimately contributing to more accurate and equitable AI-driven healthcare systems within and beyond veterinary medicine.

## 7 Conclusion

PetEVAL is the first benchmark dataset for veterinary EHRs, featuring expert-annotated resources across ICD-11 syndromic classifications, disease entity recognition, and anonymisation labels. Beyond addressing a critical gap in veterinary medicine, PetEVAL facilitates valuable comparative studies between animal and human health domains, promoting cross-disciplinary insights. As a foundational resource for veterinary informatics, this dataset promises to catalyse advancements in clinical decision support systems, enhance epidemiological surveillance capabilities, and strengthen WHO's One Health initiatives, ultimately advancing animal welfare and public health research outcomes.

## 8 Acknowledgements

We thank the data providers in veterinary practice (VetSolutions, Teleos, CVS, and other practitioners). Without their support and participation, this research would not be possible. We thank all of the annotators for their support in deriving this dataset. This work was supported by the Biotechnology and Biological Sciences Research Council [BB/T008695/1]. The carbon emissions associated with the training and evaluation of our models were entirely offset.

## 9 Limitations

Despite rigorous quality control, annotation errors are unavoidable due to the dataset's scale. Models trained on first-opinion vEHRs are inherently limited by the availability and accuracy of recorded information, often lacking confirmatory diagnostics due to financial constraints or resource limitations. Our evaluation method enforces strict token-level matching, penalising incomplete spans even when semantically close to the ground truth. While this is critical for anonymisation, it may be overly rigid for disease extraction. Similarly, our classification approach adheres strictly to predefined categories, which, while justified by the prompt, may overlook minor deviations. Future work could explore more flexible evaluation metrics and incorporate referral-level vEHRs to enhance diagnostic certainty.

## References

John Aberdeen, Samuel Bayer, Reyyan Yeniterzi, Ben Wellner, Cheryl Clark, David Hanauer, Bradley Malin, and Lynette Hirschman. 2010. The MITRE Identification Scrubber Toolkit: Design, training, and assessment. *International Journal of Medical Informatics*, 79(12):849–859.

Karim Abouelmehdi, Abderrahim Beni-Hssane, Hayat Khaloufi, and Mostafa Saadi. 2017. Big data security and privacy in healthcare: A Review. *Procedia Computer Science*, 113:73–80.

Noor Abu-El-Rub, Jay Urbain, George Kowalski, Kristen Osinski, Robert Spaniol, Mei Liu, Bradley Taylor, and Lemuel R. Waitman. 2022.

Yuval Barak-Corren, Victor M. Castro, Solomon Javitt, Alison G. Hoffnagle, Yael Dai, Roy H. Perlis, Matthew K. Nock, Jordan W. Smoller, and Ben Y. Reis. 2017. Predicting suicidal behavior from longitudinal electronic health records. *American Journal of Psychiatry*, 174(2):154–162.

Yuval Barak-Corren, Pradip Chaudhari, Jessica Perniciaro, Mark Waltzman, Andrew M. Fine, and Ben Y. Reis. 2021. Prediction across healthcare settings: a case study in predicting emergency department disposition. *npj Digital Medicine 2021 4:1*, 4(1):1–7.

Noman Bashir, Priya Donti, James Cuff, Sydney Sroka, Marija Ilic, Vivienne Sze, Christina Delimitrou, and Elsa Olivetti. 2024. The Climate and Sustainability Implications of Generative AI. *An MIT Exploration of Generative AI*. Https://mit-genai.pubpub.org/pub/8ulgrckc.

Satesh Bidaisee and Calum N.L. Macpherson. 2014. Zoonoses and one health: A review of the literature. *Journal of Parasitology Research*, 2014.

Elena Birman-Deych, Amy D. Waterman, Yan Yan, David S. Nilasena, Martha J. Radford, and Brian F. Gage. 2005. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical Care*, 43(5):480–485.

E. F. Bode, E. Mederska, H. Hodgkiss-Geere, A. D. Radford, and D. A. Singleton. 2022. Analysis of canine cardiovascular therapeutic agent prescriptions using electronic health records in primary care veterinary practices in the United Kingdom. *Journal of Veterinary Cardiology*, 39:35–45.

Mayla R. Boguslav, Adam Kiehl, David Kott, G. Joseph Strecker, Tracy Webb, Nadia Saklou, Terri Ward, and Michael Kirby. 2024. Fine-tuning foundational models to code diagnoses from veterinary health records.

Michael L. Burns and Sachin Kheterpal. 2020. Machine Learning Comes of AgeLocal Impact versus National Generalizability. *Anesthesiology*, 132(5):939–941.

Hui Cao, Peter Stetson, and George Hripcsak. 2003. Assessing explicit error reporting in the narrative electronic medical record using keyword searching. *Journal of Biomedical Informatics*, 36(1-2):99–105.

Richard J. Chen, Ming Y. Lu, Tiffany Y. Chen, Drew F.K. Williamson, and Faisal Mahmood. 2021. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering 2021 5:6*, 5(6):493–497.

Katharine Cheng, Timothy Baldwin, and Karin Verspoor. 2017. Automatic Negation and Speculation Detection in Veterinary Clinical Text. *In Proceedings of Australasian Language Technology Association Workshop,*, pages 70–78.

Martin R. Cowie, Juuso I. Blomster, Lesley H. Curtis, Sylvie Duclaux, Ian Ford, Fleur Fritz, Samantha Goldman, Salim Janmohamed, Jörg Kreuzer, Mark Leenay, Alexander Michel, Seleen Ong, Jill P. Pell, Mary Ross Southworth, Wendy Gattis Stough, Martin Thoenes, Faiez Zannad, and Andrew Zalewski. 2017. Electronic health records to facilitate clinical research. *Clinical Research in Cardiology*, 106(1):1–9.

Heather Davies, Goran Nenadic, Ghada Alfattni, Mercedes Arguello Casteleiro, Noura Al Moubayed, Sean Farrell, Alan D. Radford, and P.-J. M. Noble. 2024a. Text mining for disease surveillance in veterinary clinical data: part two, training computers to identify features in clinical text. *Frontiers in Veterinary Science*, 11:1352726.

Heather Davies, Goran Nenadic, Ghada Alfattni, Mercedes Arguello Casteleiro, Noura Al Moubayed, Sean O. Farrell, Alan D. Radford, and Peter John M. Noble. 2024b. Text mining for disease surveillance in veterinary clinical data: part one, the language of veterinary clinical records and searching for words. *Frontiers in Veterinary Science*, 11:1352239.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova Google, and A I Language. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT*, pages 4171–4186.

David A. Dorr, 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. *Methods of Information in Medicine*, 45(3):246–252.

Sean Farrell, Charlotte Appleton, Peter John Mäntylä Noble, and Noura Al Moubayed. 2023a. PetBERT: automated ICD-11 syndromic disease coding for outbreak detection in first opinion veterinary electronic health records. *Scientific Reports 2023 13:1*, 13(1):1–14.

Sean Farrell, John McGarry, Peter John Mäntylä Noble, Gina J. Pinchbeck, Sophie Cantwell, Alan D. Radford, and David A. Singleton. 2023b. Seasonality and other risk factors for fleas infestations in domestic dogs and cats. *Medical and veterinary entomology*, 37(2):359–370.

Ivo S. Fins, Heather Davies, Sean Farrell, Jose R. Torres, Gina Pinchbeck, Alan D. Radford, and Peter John Noble. 2024. Evaluating ChatGPT text mining of clinical records for companion animal obesity monitoring. *Veterinary Record*, 194(3):no.

Elizabeth Ford, Amanda Nicholson, Rob Koeling, A. Rosemary Tate, John Carroll, Lesley Axelrod, Helen E. Smith, Greta Rait, Kevin A. Davies, Irene Petersen, Tim Williams, and Jackie A. Cassell. 2013. Optimising the use of electronic health records to estimate the incidence of rheumatoid arthritis in primary care: what information is hidden in free text? *BMC medical research methodology*, 13(1).

Tracy D Gunter and Nicolas P Terry. 2005. The emergence of national electronic health record architectures in the United States and Australia: models, costs, and questions. *Journal of medical Internet research*, 7(1):e3.

Mika K. Hamer, Cathy J. Bradley, Richard Lindrooth, and Marcelo C. Perraillon. 2024. The Effect of Medicare Annual Wellness Visits on Breast Cancer Screening and Diagnosis. *Medical Care*, 62(8):530–537.

Mark A. Hlatky, Roberta M. Ray, Dale R. Burwen, Karen L. Margolis, Karen C. Johnson, Anna Kucharska-Newton, Joann E. Manson, Jennifer G. Robinson, Monika M. Safford, Matthew Allison, Themistocles L. Assimes, Anthony A. Bavry, Jeffrey Berger, Rhonda M. Cooper-DeHoff, Susan R. Heckbert, Wenjun Li, Simin Liu, Lisa W. Martin, Marco V. Perez, Hilary A. Tindle, Wolfgang C. Winkelmayer, and Marcia L. Stefanick. 2014. Use of Medicare Data to Identify Coronary Heart Disease Outcomes In the Women's Health Initiative (WHI). *Circulation. Cardiovascular quality and outcomes*, 7(1):157.

David C. Hsia, W. Mark Krushat, Ann B. Fagan, Jane A. Tebbutt, and Richard P. Kusserow. 2010. Accuracy of Diagnostic Coding for Medicare Patients under the Prospective-Payment System. *The New England journal of medicine*, 318(6):352–355.

Kasper Jensen, Cristina Soguero-Ruiz, Karl Oyvind Mikalsen, Rolv Ole Lindsetmo, Irene Kouskoumvekaki, Mark Girolami, Stein Olav Skrovseth, and Knut Magne Augestad. 2017. Analysis of free text in electronic health records for identification of cancer patient trajectories. *Scientific Reports 2017 7:1*, 7(1):1–12.

Zhipeng Jiang, Chao Zhao, Bin He, Yi Guan, and Jingchi Jiang. 2017. De-identification of medical records using conditional random fields and long short-term memory networks. *Journal of Biomedical Informatics*, 75:S43–S53.

Amir Kol, Boaz Arzi, Kyriacos A. Athanasiou, Diana L. Farmer, Jan A. Nolta, Robert B. Rebhun, Xinbin Chen, Leigh G. Griffiths, Frank J.M. Verstraete, Christopher J. Murphy, and Dori L. Borjesson. 2015. Companion animals: Translational scientist's new best friends. *Science translational medicine*, 7(308):308ps21.

Hyoun Joong Kong. 2019. Managing Unstructured Big Data in Healthcare System. *Healthcare Informatics Research*, 25(1):1.

Harlan M. Krumholz, Sharon Lise T. Normand, and Yun Wang. 2014. Trends in hospitalizations and outcomes for acute cardiovascular disease and stroke, 1999-2011. *Circulation*, 130(12):966–975.

Junhak Lee, Jinwoo Jeong, Sungji Jung, Jihoon Moon, and Seungmin Rho. 2022. Verification of De-Identification Techniques for Personal Information Using Tree-Based Methods with Shapley Values. *Journal of Personalized Medicine 2022, Vol. 12, Page 190*, 12(2):190.

Kahyun Lee, Nicholas J Dobbins, Bridget Mcinnes, Meliha Yetisgen, and Özlem Uzuner. 2021. Transferability of Neural Network Clinical De-identification Systems. *Journal of the American Medical Informatics Association*.

Joffrey L. Leevy, Taghi M. Khoshgoftaar, and Flavio Villanustre. 2020. Survey on RNN and CRF models for de-identification of medical free text. *Journal of Big Data*, 7(1):1–22.

Yi Liu, Jialiang Peng, James J. Q Yu, and Yi Wu. 2019. PPGAN: Privacy-preserving Generative Adversarial Network. *Proceedings of the International Conference on Parallel and Distributed Systems - ICPADS*, 2019-December:985–989.

E M Lund. 2015. Power of practice: using clinical data to advance veterinary medicine. *Veterinary Record*.

Matthew B.A. McDermott, Shirly Wang, Nikki Marinsek, Rajesh Ranganath, Luca Foschini, and Marzyeh Ghassemi. 2021. Reproducibility in machine learning for health research: Still a ways to go. *Science Translational Medicine*, 13(586).

Paul McGreevy, Peter Thomson, Navneet K. Dhand, David Raubenheimer, Sophie Masters, Caroline S. Mansfield, Timothy Baldwin, Ricardo J.Soares Magalhaes, Jacquie Rand, Peter Hill, Anne Peaston, James Gilkerson, Martin Combs, Shane Raidal, Peter Irwin, Peter Irons, Richard Squires, David Brodbelt, and Jeremy Hammond. 2017. VetCompass Australia: A National Big Data Collection System for Veterinary Science. *Animals 2017, Vol. 7, Page 74*, 7(10):74.

Mary L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276.

Stephane M. Meystre, 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):1–16.

Jose Antonio Miñarro-Giménez, Catalina Martínez-Costa, Daniel Karlsson, Stefan Schulz, and Kirstine Rosenbeck Gøeg. 2018. Qualitative analysis of manual annotations of clinical text with SNOMED CT. *PLOS ONE*, 13(12):e0209547.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Allen Nie, Ashley Zehnder, Rodney L Page, Yuhui Zhang, Arturo Lopez Pineda, Manuel A Rivas, Carlos D Bustamante, and James Zou. 2018. DeepTag: inferring diagnoses from veterinary clinical notes. *npj Digital Medicine 2018 1:1*, 1(1):1–8.

Sarah J. Price, Sal A. Stapley, Elizabeth Shephard, Kevin Barraclough, and William T. Hamilton. 2016. Is omission of free text records a possible source of data loss and bias in Clinical Practice Research Datalink studies? A case-control study. *BMJ open*, 6(5).

A. D. Radford, 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):310–310.

I. D. Robertson, P. J. Irwin, A. J. Lymbery, and R. C.A. Thompson. 2000. The role of companion animals in the emergence of parasitic zoonoses. *International Journal for Parasitology*, 30(12-13):1369–1377.

N. J. Robinson, R. S. Dean, M. Cobb, and M. L. Brennan. 2016. Factors influencing common diagnoses made during first-opinion small-animal consultations in the United Kingdom. *Preventive Veterinary Medicine*, 131:87–94.

Royal Veterinary College (RVC). VetCompass.

Fernando Sánchez-Vizcaíno, Philip H. Jones, Tarek Menacere, Bethaney Heayns, Maya Wardeh, Jenny Newman, Alan D. Radford, Susan Dawson, Rosalind Gaskell, Peter J.M. Noble, Sally Everitt, Michael J. Day, and Katie McConnell. 2015. Small animal disease surveillance. *Veterinary Record*, 177(23):591–594.

Fernando Sánchez-Vizcaíno, Peter John M. Noble, Phil H. Jones, Tarek Menacere, Iain Buchan, Suzanna Reynolds, Susan Dawson, Rosalind M. Gaskell, Sally Everitt, and Alan D. Radford. 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):1–13.

Zhecheng Sheng, Emma Bollig, Jennifer Granick, Rui Zhang, and Amanda Beaudoin. 2022. Canine Parvovirus Diagnosis Classification Utilizing Veterinary Free-Text Notes. *Proceedings - 2022 IEEE 10th International Conference on Healthcare Informatics, ICHI 2022*, pages 614–615.

Gregory E. Simon, Susan M. Shortreed, R. Yates Coley, Robert B. Penfold, Rebecca C. Rossom, Beth E. Waitzfelder, Katherine Sanchez, and Frances L. Lynch. 2019. Assessing and Minimizing Re-identification Risk in Research Data Derived from Health Care Records. *eGEMs*, 7(1):6.

Jasvinder A. Singh, Aaron R. Holmgren, and Siamak Noorbaloochi. 2004. Accuracy of veterans administration databases for a diagnosis of rheumatoid arthritis. *Arthritis Care & Research*, 51(6):952–957.

David A. Singleton, Gina L. Pinchbeck, Alan D. Radford, Elena Arsevska, Susan Dawson, Philip H. Jones, Peter John M. Noble, Nicola J. Williams, and Fernando Sánchez-Vizcaíno. 2020. Factors Associated with Prescription of Antimicrobial Drugs for Dogs and Cats, United Kingdom, 2014–2016. *Emerging Infectious Diseases*, 26(8):1778.

Mike P. Starkey, Timothy J. Scase, Cathryn S. Mellersh, and Sue Murphy. 2005. Dogs really are man's best friend–canine genomics has applications in veterinary and human medicine! *Briefings in functional genomics & proteomics*, 4(2):112–128.

Amber Stubbs, Michele Filannino, and Özlem Uzuner. 2017. De-identification of psychiatric intake records: Overview of 2016 CEGS N-GRID shared tasks Track 1. *Journal of Biomedical Informatics*, 75:S4–S18.

Amber Stubbs and Özlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. *Journal of biomedical informatics*, 58 Suppl(Suppl):S20–S29.

Zhaohao Sun, Kenneth David Strang, and Francisca Pambel. 2020. Privacy and security in the big data paradigm. *Journal of Computer Information Systems*, 60(2):146–155.

Mahanazuddin Syed, Kevin Sexton, Melody Greer, Shorabuddin Syed, Joseph VanScoy, Farhan Kawsar, Erica Olson, Karan Patel, Jake Erwin, Sudeepa Bhattacharyya, Meredith Zozus, and Fred Prior. 2022. DeIDNER Model: A Neural Network Named Entity Recognition Model for Use in the De-identification of Clinical Notes. *Biomedical engineering systems and technologies, international joint conference, BIOSTEC ... revised selected papers. BIOSTEC (Conference)*, 5:640.

Llama Team and Ai @ Meta. 2024. The Llama 3 Herd of Models.

Erik F Tjong, Kim Sang, and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. pages 142–147.

Darren J. Trott, Lucio J. Filippich, John C. Bensink, Mary T. Downs, Suzanne E. McKenzie, Kirsty M. Townsend, Susan M. Moss, and James J.C. Chin. 2004. Canine model for investigating the impact of oral enrofloxacin on commensal coliforms and colonization with multidrug-resistant Escherichia coli. *Journal of Medical Microbiology*, 53(5):439–443.

Alexander Turchin, Nikheel S. Kolatkar, Richard W. Grant, Eric C. Makhni, Merri L. Pendergrass, and Jonathan S. Einbinder. 2006. Using Regular Expressions to Abstract Blood Pressure and Treatment Intensification Information from the Text of Physician Notes. *Journal of the American Medical Informatics Association*, 13(6):691–695.

Engeline Van Duijkeren, Maurice J.H.M. Wolfhagen, Adrienne T.A. Box, Max E.O.C. Heck, Wim J.B. Wannet, and Ad C. Fluit. 2004. Human-to-Dog Transmission of Methicillin-Resistant Staphylococcus aureus. *Emerging Infectious Diseases*, 10(12):2235.

Elizabeth J. Williamson, Alex J. Walker, Krishnan Bhaskaran, Seb Bacon, Chris Bates, Caroline E. Morton, Helen J. Curtis, Amir Mehrkar, David Evans, Peter Inglesby, Jonathan Cockburn, Helen I. McDonald, Brian MacKenna, Laurie Tomlinson, Ian J. Douglas, Christopher T. Rentsch, Rohini Mathur, Angel Y.S. Wong, Richard Grieve, David Harrison, Harriet Forbes, Anna Schultze, Richard Croker, John Parry, Frank Hester, Sam Harper, Rafael Perera, Stephen J.W. Evans, Liam Smeeth, and Ben Goldacre. 2020. Factors associated with COVID-19-related death using OpenSAFELY. *Nature 2020 584:7821*, 584(7821):430–436.

World Health Organisation (WHO). 2022. International Classification of Diseases 11th Revision (ICD-11).

Judit M. Wulcan, Kevin L. Jacques, Mary Ann Lee, Samantha L. Kovacs, Nicole Dausend, Lauren E. Prince, Jonatan Wulcan, Sina Marsilio, and Stefan M. Keller. 2024. Classification performance and reproducibility of GPT-4 omni for information extraction from veterinary electronic health records. *Frontiers in Veterinary Science*, 11:1490030.

Xi Yang, Tianchen Lyu, Qian Li, Chih Yin Lee, Jiang Bian, William R. Hogan, and Yonghui Wu. 2019. A study of deep learning methods for de-identification of clinical notes in cross-institute settings. *BMC Medical Informatics and Decision Making*, 19(5):1–9.

Jinsung Yoon, Michel Mizrahi, Nahid Farhady Ghalaty, Thomas Jarvinen, Ashwin S. Ravi, Peter Brune, Fanyu Kong, Dave Anderson, George Lee, Arie Meir, Farhana Bandukwala, Elli Kanal, Sercan Arık, and Tomas Pfister. 2023. EHR-Safe: generating high-fidelity and privacy-preserving synthetic electronic health records. *npj Digital Medicine 2023 6:1*, 6(1):1–11.

Yuhui Zhang, Allen Nie, Ashley Zehnder, Rodney L Page, and James Zou. 2019. VetTag: improving automated veterinary diagnosis coding via large-scale language modeling. *npj Digital Medicine 2019 2:1*, 2(1):1–8.

# 10 appendices

## 10.1 Task 1: Anonymisation Prompt

Prompt: Extract Named Entities from Veterinary EHRs You are given short free-text veterinary electronic health records (EHRs). Your task is to extract named entities mentioned in the text. Focus on identifying Names (NAME) locations (LOC), organizations (ORG), temporal expressions (TIME), and miscellaneous named entities (MISC). Examples:

Input: "Raven GA castrate. Anaes: Premed ACP/Meth. Induced propofol maint iso/02. Good anaesthetic. Op: Routine open castrate. double ligated 2-0 polysorb. Skin closed intradermal." Output: RavenNAME = Raven LOC = ORG = TIME = MISC = Input: "Waffle/MG - back end irritation. Owner reports irritation round back end, rubbing bottom over last 2-3 weeks." Output: NAME = Waffle LOC = ORG = TIME = last 2-3 weeks MISC = Input: "Adv routine haem/biochem (est £603) owner will discuss with wife. - Prescription -. Date: Apr 3, 2002. Vet: Reese, Qualifications: MRCVS." Output: NAME = Reese LOC = ORG = TIME = Apr 3, 2002 MISC =

Guidelines: -Extract only named entities in the appropriate categories:

NAME: Pet Names, Owner Names, Cliniain names LOC: geographical locations, clinics, hospitals, animal shelters ORG: veterinary practices, laboratories, pharmaceutical companies TIME: dates, time periods, durations, temporal references MISC: animal names, medications, procedures, medical equipment, qualifications - List each entity under its proper category. - If multiple entities of the same type are mentioned, extract each one separately. - Maintain the exact form as mentioned in the text.

## 10.2 Task 2: Syndromic Disease Classification Prompt

You are given a free-text veterinary electronic health records (EHRs). Your task is to assign a ICD-11 chapter names based on the conditions, symptoms, and diagnoses mentioned in the text. Each assigned chapter should correspond to the primary system or disease category affected.

ICD-11 Chapters: 1. Certain infectious or parasitic diseases 2. Neoplasms 3. Diseases of the blood or blood-forming organs 4. Diseases of the immune system 5. Endocrine, nutritional, or metabolic diseases 6. Mental, behavioral, or neurodevelopmental disorders 7. Sleep-wake disorders 8. Diseases of the nervous system 9. Diseases of the eye and adnexa 10. Diseases of the ear and mastoid process 11. Diseases of the circulatory system 12. Diseases of the respiratory system 13. Diseases of the digestive system 14. Diseases of the skin 15. Diseases of the musculoskeletal system or connective tissue 16. Diseases of the genitourinary system 17. Conditions related to sexual health 18. Pregnancy, childbirth, or the puerperium 19. Certain conditions originating in the perinatal period 20. Developmental anomalies 21. Symptoms, signs, or clinical findings not elsewhere classified 22. Injury, poisoning, or certain other consequences of external causes 23. External causes of morbidity or mortality 24. Factors influencing health status or contact with health services

Examples:

1. Input: "marked signs of renal failure. not eating much. huge wt loss. not moving around much." Output: Disease of the genitourinary system

2. Input: "Bilat OE. Mild, cleaned and wax removed, no obvious sign mites. Start on ear drops, rv sooner if concerned otherwise at next vaccination on 29th." Output: Diseases of the ear and mastoid process

3. Input: "skin lesions, bloods for meds check. noticed spot like skin lesions on forehead and side of face. not rubbing/scratching. would like checked. mass on R flank, slow growing, separated masses now merged together. pulsing meloxaid for stomatogingvitis." Output: Disease of the digestive system, Disease of the skin, Neoplasms

Guidelines: - Assign at least one ICD-11 chapter name that best represents the condition(s) described. - If no condition is present then return 'None' - If multiple conditions from different systems are mentioned, include multiple ICD-11 chapter names. - Ignore non-diagnostic text (e.g., medication instructions or routine check-ups) unless relevant to a condition. - Maintain consistency in ICD-11 chapter naming as per the official classification.

### 10.3 Task 3: Disease Extraction Prompt

You are given a free-text veterinary electronic health records (EHRs). Your task is to **extract the disease names** mentioned in the text. Focus on identifying diseases or conditions specifically mentioned, ignoring general symptoms, treatments, or non-diagnostic text.

Examples:

1. Input: "marked signs of renal failure. not eating much. huge wt loss. not moving around much." Output: renal failure

2. Input: "Bilat OE. Mild, cleaned and wax removed, no obvious sign mites. Start on ear drops, rv sooner if concerned otherwise at next vaccination on 29th." Output: OE

3. Input: "skin lesions, bloods for meds check. noticed spot like skin lesions on forehead and side of face. not rubbing/scratching. would like checked. mass on R flank, slow growing, separated masses now merged together. pulsing meloxaid for stomatogingvitis." Output: Skinskin lesions, stomatogingvitis, mass on R flank

Guidelines: - Extract only disease names (e.g., "Renal failure", "Otitis externa", "Neoplasm"). - Do not include symptoms, treatment plans, or general findings (e.g., "not eating much", "Start on ear drops"). - If multiple diseases are mentioned, extract each disease separately. - Maintain consistency in naming diseases and conditions as per medical terminology.