PetBERT: Automated ICD-11 Syndromic Disease Coding for Outbreak Detection in First Opinion Veterinary Electronic Health Records

Sean Farrell [1], Charlotte Appleton [2], Peter-John Mäntylä Noble [3], Noura Al Moubayed [1]

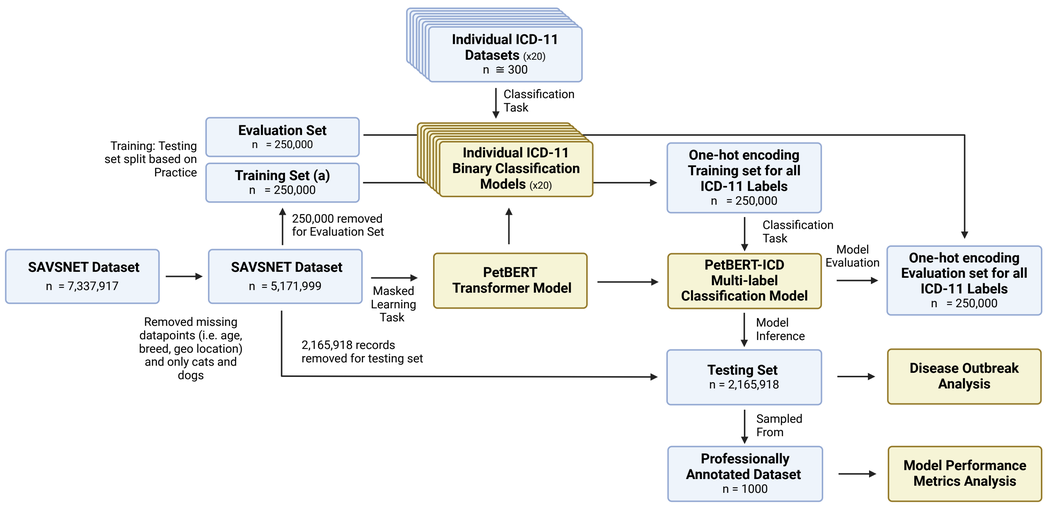
[1] Department of Computer Science, Durham University, Durham, UK
[2] Centre for Health Informatics, Computing, and Statistics, Lancaster Medical School, Lancaster University, Lancaster, UK
[3] Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool, Liverpool, UK

ABSTRACT

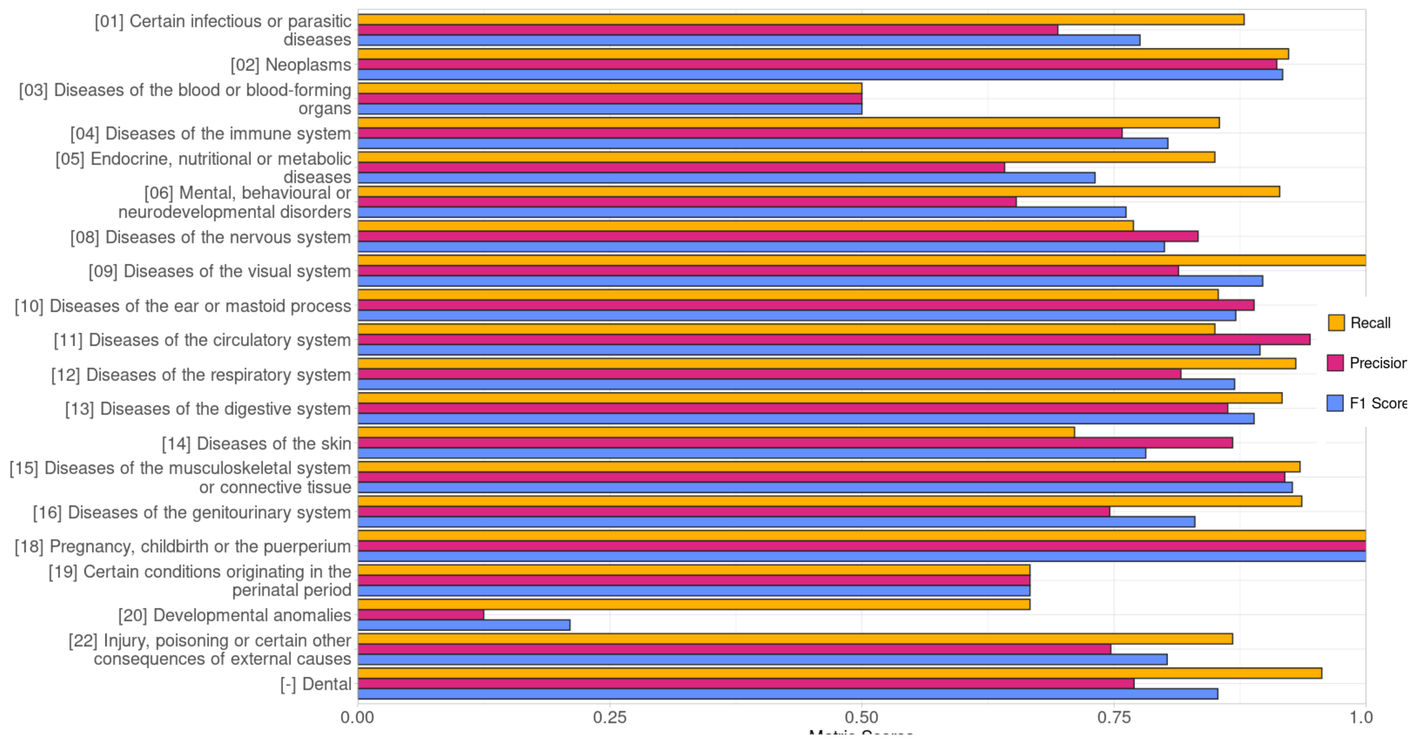
Effective public health surveillance requires consistent monitoring of disease signals such that researchers and decision-makers can react dynamically to changes in disease occurrence. However, whilst surveillance initiatives exist in production animal veterinary medicine, comparable frameworks for companion animals are lacking. First-opinion veterinary electronic health records (EHRs) have the potential to reveal disease signals and often represent the initial reporting of clinical syndromes in animals presenting for medical attention, highlighting their possible significance in early disease detection. Yet despite their availability, there are limitations surrounding their free text-based nature, inhibiting the ability for national-level mortality and morbidity statistics to occur. This paper presents PetBERT, a large language model trained on over 500 million words from 5.1 million EHRs across the UK. PetBERT-ICD is the additional training of PetBERT as a multi-label classifier for the automated coding of veterinary clinical EHRs with the International Classification of Disease 11 framework, achieving F1 scores exceeding 83% across 20 disease codings with minimal annotations. PetBERT-ICD effectively identifies disease outbreaks, outperforming current clinician-assigned point-of-care labelling strategies up to three weeks earlier. The potential for PetBERT-ICD to enhance disease surveillance in veterinary medicine represents a promising avenue for advancing animal health and improving public health outcomes.

METHOD

1. SAVSNET Dataset was split into 5.1 million training set and a 2.1 million testing set based on attending practice
2. Pre-trained BERT-base model was fine-tuned on the 5.1 million containing 500 million token dataset of clinical free-text narratives (1).
3. Individual Small PetBERT-based binary sequence classification models created for 20 target ICD-11 labels.
4. 20 models applied against a 250k training and evaluation sets
5. A multi label PetBERT-ICD model was trained and evaluated on these derived dataset
6. PetBERT-ICD applied against 2.1 million testing set for downstream analysis



RESULTS

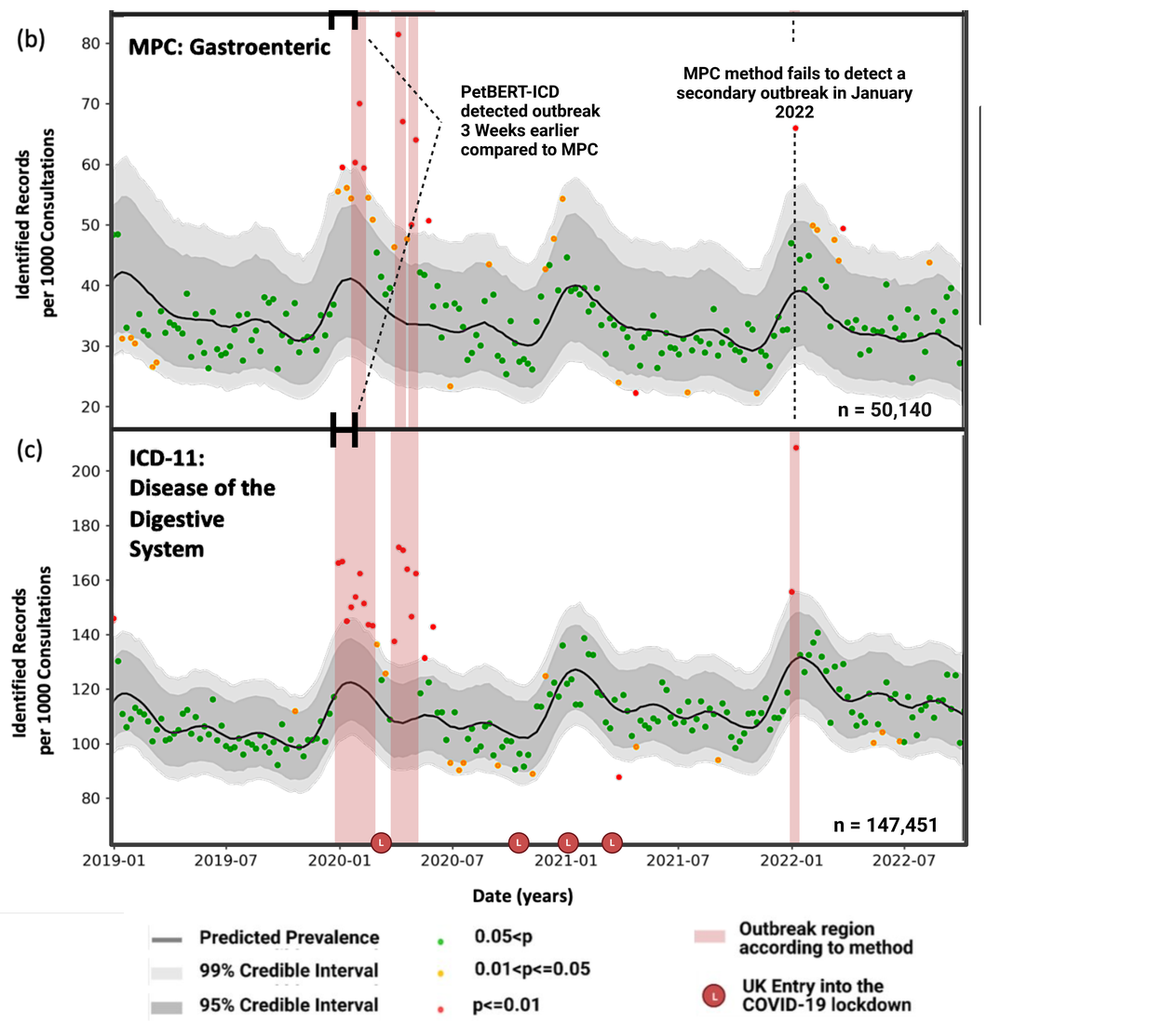


PetBERT-ICD is the additional training of PetBERT as a multi-label classifier for the automated coding of veterinary clinical EHRs with the International Classification of Disease 11 framework, achieving F1 scores exceeding 83\% across 20 disease codings with minimal annotations.

CASE STUDY: Disease Outbreak Detection

PetBERT successfully captured the 2019 canine enteric outbreak (2), as shown by the points representing each week that fell within the 95% credible intervals, indicating normal periods. Points exceeding the credible intervals for two weeks, like those beyond the 99% intervals, indicate a potential outbreak (3). More significant clustering beyond these intervals occurred during the first outbreak period, meeting the definition.

Using ICD-11 labelling, this outbreak was detected three weeks before the MPC system. A second outbreak in January 2022 was suggested by the ICD-11 data with single and paired labelling but was absent with MPC labelling.



REFERENCES

1) Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

(2) Radford, A. D., Singleton, D. A., Jewell, C., Appleton, C., Rowlingson, B., Hale, A. C., ... & Pinchbeck, G. L. (2021). Outbreak of severe vomiting in dogs associated with a canine enteric coronavirus, United Kingdom. *Emerging infectious diseases*, *27*(2), 517.

(3) Hale, A. C., Appleton, C., Noble, P. J., Pinchbeck, G. L., Rowlingson, B., Diggle, P. J., ... & Jewell, C. P. (2022). Visualising spatio-temporal health data: the importance of capturing the 4th dimension. *arXiv preprint arXiv:2211.02364*.Chicago