Surveillance of equine strangles in the United Kingdom between 2015 and 2019 based on laboratory detection of *Streptococcus equi*: Datasets and analysis code

Introduction

This repository contains datasets and code used for the manuscript as titled above. All details regarding the data source and considerations that should be noted are discussed in the manuscript; these are important to understand before embarking on any analysis of these data - in particular the differentiation between a strangles diagnosis and a positive sampling event - a single diagnosis can be made up of multiple sample events.

Referencing this dataset

Abigail McGlennon, Andrew Waller, Kristien Verheyen, Josh Slater, John Grewar, David Aanensen & Richard Newton (2020) Surveillance of equine strangles in the United Kingdom between 2015 and 2019 based on laboratory detection of *Streptococcus equi* [Dataset]. Royal Veterinary College. https://zenodo.org/badge/latestdoi/290019559

R Code

The descriptive analysis for the manuscript was performed using R. The full code is available in the *manuscriptcode.R* file. Links to the csv dataset used are included in the code as well as the required libraries. The code published here is in a format that may assist entry-level R users in navigating through the outcomes. Piping using *dplyr* was used throughout to create a step-for-step code base.

Datasets

CSV files

- ***dataset_diagnosesgen_all.csv*** is the foundational dataset where every row relates to a single Strangles diagnosis
    - `strangleslogid` denotes the unique identifier for a single strangles diagnosis
    - `eventdate` denotes the earliest date associated with each diagnosis
    - `submittingvet_prim` denotes the primary veterinary practice submitting samples for each diagnosis
    - `nuts3gid` denotes the NUTS3 gid identifier to link to the NUTS3 spatial polygon dataset *(see the **GIS** section below)*
    - `breedgen` denotes the general breed of the affected horse
    - `sexgen` denotes the sex of the affected horse
    - `premesiscategory` denotes the premises category where the affected horse was present at diagnosis
    - `age` denotes the age (in years) of the affected horse
    - `submittinglab_prim` denotes the primary laboratory submitting data relating to the diagnosis
    - `result` denotes the distinct aggregated test methods used for *S.equi* positive results for each diagnosis
- ***dataset_diagnosessampletype.csv*** represents the sample types (`sampletype`) and sample locations (horse - `samplelocation`) for each diagnosis (`strangleslogid`) where multiple lines may represent a single diagnosis - i.e. multiple sample types and from multiple locations were registered per diagnosis.
- ***dataset_clinicalsigns.csv*** represents the clinical signs noted for each strangles diagnosis. As with the *dataset_diagnosessampletype* dataset, multiple clinical signs (`cx`) are associated with a distinct diagnosis (`strangleslogid`).
- ***dataset_samplereason.csv*** represents the reason for sampling (`samplereason`) with multiple rows constituting a strangles diagnosis (`strangleslogid`).
- ***dataset_sampledecisions_type.csv*** represents the aggregated reason for sampling (`samplereasons2`) with a single row constituting a strangles diagnosis (`strangleslogid`), and in this case diagnoses are omitted where *undefined* was the only associated reason for sampling. Furthermore, the `sampletypes2` field denotes the aggregated sample types that were associated with each diagnosis.
- ***dataset_sampletestoutcome.csv*** represents the *S.equi* culture (`culture`) and real-time PCR (`qPCR`) results for each sample where at least one of these tests had a positive result. `sampledetailid` is the unique identifier for each sample tested and is not used further in this analysis.

- ***dataset_nuts3_base.csv*** is used to facilitate aggregation of diagnoses to spatial locations using table joins (rather than spatial joins) and the `gid` field correlates to the `gid` field in the GIS dataset *nuts3_regions.shp* below. `nuts_name` denotes the name of the region.

GIS (Spatial) datasets

- ***nuts3_regions.shp*** is a shapefile of the polygons making up the NUTS3 regions of the United Kingdom specifically. The European Petroleum Survey Group (EPSG) coordinate system code for this dataset is 4326 (i.e. WGS84). It contains a unique identifier `gid` which can be linked to csv datasets in this repository and the `nuts_name` associated with the area. This NUTS3 region shapefile has been amended slightly to facilitate analysis and depiction of data on the UK regional scale. In particular, smaller urban regions have been merged with their peri-urban/rural neighbours. For access to maintained NUTS datasets please go here and observe required usage rights.