Geographical distribution

STRUCTURE OF THE DATA AVAILABLE

Field name	Data type	Cleaning notes
refID	Numerical	Unique identification of a reference.
FullReference	Free-text	Full reference in the format: "all authors, YEAR, publication title,journal,issue, pages".
groupID	Numerical	This field identify UNIQUE STUDY GROUPS within each paper. If the same reference reported results for multiple groups, for instance different species, this field allows one to track which results belong to the same group, and which results form the same paper are not about the same group. refID + groupID, together, provide a unique identification of all reported animal groups in the literature review.
studyContext	Categorical	The type of study being reported, for instance outbreak investigation, prevalence estimation or a case study. Results should be considered in lights of this field.
studyContext_C	Free text	Comment field used to specify the type of study when the study context was declared to be "other".
sampStrategy	Categorical	The sampling strategy reported in the study, such as Objective (random) sampling, Selective (risk-based) sampling, etc. When the study focused on a specific population stratum, we assigned it to be "selective sampling" and used a free-text field to note which stratum (see nest field). During data cleaning we reviewed the consistency of use of the different sampling strategies, and even the consistency between study context and sampling strategy. We adopted the following practices: - When the study context was "outbreak", if all cases were described, the sampling strategy was declared as "census". Else we used the described method of selection (usually suspect sampling, convenience or unspecified) - For "case study", when we don't believe a sampling strategy is actually applicable, we selected "suspect sampling". Some contexts worth of note: - A large number of studies self-declared as "prevalence studies", but they were based on a compilation of reported cases. This explains why you will find "suspect sampling" within the study context of "prevalence study"
sampStrategy_C	Free text	Comment field used to specify the sampling strategy details, in particular when the sampling strategy declared was "Selective (risk-based) sampling". The free-text was standardized during data cleaning.
sampPoint	Categorical	Where samples were collected.
sampPoint_C	Free text	Comment field used to specify the sampling point when not in EFSAs catalogue – sampling point declared as "Other".
progType	Categorical	These categories are based on EFSA's catalogue for program type, and include terms like "monitoring active" and "monitoring passive". There are no categories to accommodate observational studies (cohort and case studies), which do not fit into any surveillance type described in this catalogue. We hesitated, however, to declare "unspecified", as we knew the type of study carried out. Therefore, we have added the category "observational study".
progType_C	Free text	Comment field used to specify the program type when none of the used categories were appropriate – progType declared as "Other".
MinAgeMonths	Numerical	(-1 if not specified)
MaxAgeMonths	Numerical	Consistency of values and ranges were checked.

sampUnit	Categorical	Sampling units. Individual animals/humans were declared as "host" (G199A), while units of vectors (ticks or sand-flies, for instance) were declared as "single" (G203A). In several studies, the samples were pooled, and results reported for the pools, particular when testing vectors. As there is no specific "pool" unit in the EFSA catalogue "SAMPUNTYP", we used "batch" in these cases.
sampUnit_C	Free text	Comment field used to specify the sampling unit when none of the used categories were appropriate – sampUnit declared as "Other".
sampUnitSize	Numerical	Sample size. Checked for consistency (for example consistency with the study type and type of sample declared) and corrected when needed.
country	Categorical	Country where the study was conducted. Checked for consistency and completeness 5 studies have no declared geographical location – these are mathematical models investigation the reproduction number of the disease, not reporting any specific occurrence.
sampArea	Free-text	No data cleaning applied as this is meant to be a free description of the area, in case too many papers refer to the same country.
startYear	Numerical	Year of start of the outbreak/study. (-1 if not specified)
StartMonth	Numerical	Month of start of the outbreak/study. (-1 if not specified)
duration	Numerical	Duration in months. (-1 if not specified)
phase	Categorical	Outbreak phase: "initial phase", "ongoing", or "resolved" and "not reported". "not investigated/not given/not relevant" was used when the phase was not reported, and also when the paper was not about a particular outbreak (meaning in reality "not relevant/not applicable"). This information should always be used in light of the "study context" information, which will determine whether the "phase" variable is relevant.
targetSpecies	Categorical	This was sometimes a host species, and sometimes a vector species.
matrix	Categorical	Animal matrix sampled, following EFSA's catalogue MTX.
matrix_C	Free text	Comment field used to specify the matrix when none of the used categories were appropriate – matrix declared as "Other".
analysisYear	Numerical	Year when the analysis were carried out (-1 if not specified).
labTarget	Categorical	As per EFSA's catalogues concept of "anMatText", the target of diagnostic tests (antibody, antigen, nucleic acid, etc).
agent	Categorical	The VBD agent.
agentSubtype	Categorical	The type of subtype reported (for instance "strain").
agentDetails	Free text	Details of the subtype. This was meant to be free-text, but during data cleaning repeated subtypes were reviewed to ensure consistency.
agentSubtypeDC1	Categorical	A form of recording the subtype used in the first DACRAH round, in which only listed subtypes could be chosen by data collectors. This was later changed to the two fields above ("agentSubtype" and "agentDetails").
labTest	Categorical	As per EFSA catalogues for the concept "anMethCode", the type of assay used for diagnosis.
labDescription	Free-text	Free description of the test details, such as cut-off values, specific procedures, etc.
nPositive	numerical	Number of positive tests.

		Please note that this field should be evaluated carefully *in light of* the type of study. For case studies, the number of positive cases will generally be the same as the number of units in the sample.
nNegative	numerical	Number of negative tests, when given.
sensitivity	Percentage	
specificity	Percentage	
prevalence	Percentage	
UCI_PR	Percentage	Prevalence upper limit of the confidence interval.
LCI_PR	Percentage	Prevalence lower limit of the confidence interval.
Incidence	Percentage	
UCI_I	Percentage	Incidence upper limit of the confidence interval.
LCI_I	Percentage	Incidence lower limit of the confidence interval.
incidUnits	text	Incidence Units. This was originally set as a free-text field by EFSA. We kept it as such, but during data cleaning all entered options were standardized.
incidUnits2	text	We are not sure why this question was duplicated in the original form, but during data cleaning it was only used if any additional information about the incidence units was available.
R0	numerical	Reproductive number.
lowerRo	numerical	Reproductive number upper limit of the confidence interval.
upperRo	numerical	Reproductive number lower limit of the confidence interval.
resInfo	Free-text	No data cleaning applied as this is meant to be a free description of the test details. Any notes relevant during the data cleaning were added here as well.
rowID	numerical	A unique number for all rows in the dataset.
ShortBibliography	Free-text	Reference in the format "First author, et al. YEAR".
Author	Free-text	List of authors
Title	Free-text	Publication title
Abstract	Free-text	Abstract
publicationYear	Free-text	Publication year.

NOTES AND WARNINGS ON DATA MEANING AND INTERPRETATION, ASSUMPTIONS AND SHORTCOMINGS

1) Data rows with the same refID are results reported from the same study
2) Individual study groups within these references receive the same groupID. These could be for instance different species reported in the same area.
3) Combinations of refiD+ groupID represent UNIQUE animal groups for which results are reported. These two fields should be used to identify multiple rows of outcomes that refer to the same animal group.
4) Data collection is performed in Distiller using "data collection forms". Each form results in one row when the data are looked in the tabular format (for instance in Excel of .CSV format). Every output can only be reported once in each form, therefore to report multiple values of the same type of outcome for the same group (say the detection window for different tests, or for different matrices), the entire form must be duplicated. Say for instance that we have a group of animals which were tested at two different phases of an outbreak. Data collectors would enter the number of positive on the first time point, with the corresponding outbreak phase. Then they would duplicate the form and change only the information about the outbreak phase, and enter the results for the new time point. This would result in two rows of data, but only one outbreak in the country in question.
5) In DACRAH1, this research objective was subjected to literature review not by the contracted CovetLab consortium, but by a team of experts within EFSA. The CoVetLab consortium followed the

systematic literature review protocol established for the references labelled with refID greater than 80000.

6) In this research objective, please pay particular attention to the study context, as not all data points refer to outbreak or case investigations.

7) The results represent the extensive literature review performed, but only peer-reviewed literature is included. The outputs do not represent a true global distribution of the pathogens, but the registered occurrence/studies in peer-reviewed literature.