Vaccines

The figure below is a scheme of the experimental design and data collection process. The boxes in the bottom are examples of individual forms of data collected (which are represented in the data as individual rows).

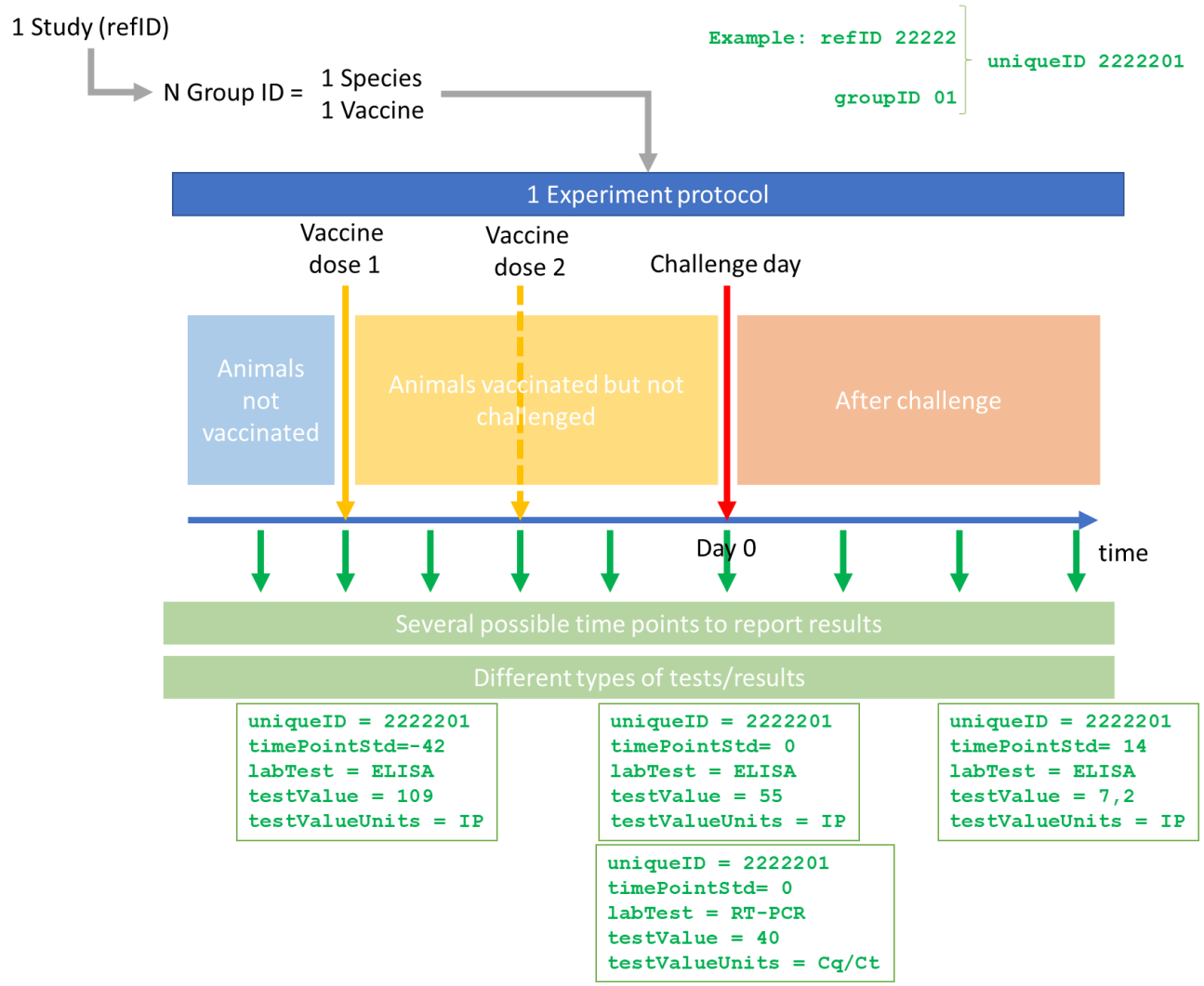

Figure 1: Experimental design of a typical vaccine study eligible for review, and examples of the generated data forms, showing how they can be identified.

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. For vaccines, these study groups could refer to different vaccine types being tested, different species, etc. Groups could be followed along multiple time points, and this field should be used to identify results belonging to the same group.
country	Categorical	Coded by EFSA. Collected in the form as a RADIO LIST.
diseaseStatus	categorical	Disease status in the country of the study.

year	numerical	Year when the study was carried out.
agent	Categorical	Vector borne disease (VBD) agent. Coded by EFSA. Collected in the form as a RADIO LIST.
agentSubtypeType	Categorical	Collected in the form as a RADIO LIST – so a fixed number of categories to choose from, but these category names are not coded following any specific catalogue.
agentDetails	Free text	During data cleaning repeated subtypes were reviewed to ensure consistency and standardization. As much as possible, this was categorized.
targetSpecies	Categorical	Coded by EFSA initially, with more items later added by the Covetlab team (using codes from the same DCF catalogue). Collected in the form as a RADIO LIST.
ageDC1	categorical	In DACRAH1 age was collected as categories (adult or young), but from DACRAH2 this was substituted by the field above.
ageMonths	numerical	Checked for consistency and completeness. All values left blank were checked by a second reviewer against the original paper to confirm that missing. In case of a specific age range, we entered the lowest age value.
sampStrategy	Categorical	Sampling strategy. Coded by EFSA. Collected in the form as a RADIO LIST.
sampUnit	Categorical	Sampling unit. Coded by EFSA. Collected in the form as a RADIO LIST.
sampUnitSize	Numerical	Checked that used consistently (corrected when needed during a data cleaning process).
route	Categorical	Route of administration. Left blank only for "unvaccinated controls". Filled in for challenge groups as well as controls which received placebo.
testSubstanceCAT	Categorical	The vaccine used. A drop-down list was presented to data collectors, containing a list of the vaccines eligible for the study – only those approved for use in the EU.
dosageFreq	numerical	Vaccine dosage frequency. Only left empty for control groups (unvaccinated or placebo)
dosageInterval	numerical	During data cleaning, checked that always filled when the dosage frequency was greater than 1.
dose	numerical	Numerical, to be interpreted with the units measure in the next field.
doseUnits	Categorical	Units for the number reported in dose.
dayDose1	numerical	In which experiment day the animals received the first vaccine dose. When papers counted the beginning of the experiment from the day of first vaccination, this was usually day 0. We recorded the day of the vaccination, the day of challenge, and later the day for each result given. Paper 50146, however, provided all results based on the day of the challenge, which was a full year after the vaccination. In this case the vaccination day is actually negative, because day 0 is the day of challenge (we used 365 days to indicate the "12 months" mentioned in the paper, and 386 days to indicate that a first vaccine dose was given 21 days prior to the final vaccine dose). This complexity/confusion is a result of the different ways results are reported, and the best we could think to cope with it was: 1) To record all results as experiment days, and to record the experiment day of the vaccination and challenge, so that all results can be compared to the vaccination and challenge day 2) To create a specific category indicating the exact point in the experiment for which results are being given (see "Experiment Status" below).
dayDose2	numerical	If animals received 2 vaccine doses, this is the experiment day in which they receive the second dose.
challengeType	Categorical	Whether animals were challenged with the virus after having received the vaccine treatment. The options are: "Challenged", "unchallenged" or "challenged with a placebo".
challengeSubstance	Free-text	Free-text specifying any details regarding the challenge. No data cleaning applied, the objective information needed is documented in other fields,

		and this was left to add any further information regarding specific inoculum preparates.
challengeDose	numerical	Numerical, to be interpreted with the units measure in the next field.
challengeDoseUnits	Categorical	Units for the number reported in dose.
challengeDay	numerical	The experiment day n which animals were challenged. Time between vaccination and challenged is interpreted by comparing this to dayDose1 and dayDose2. When directly reported, this is also recorded in the following field.
lastVaccine	numerical	Number of days since last vaccine dose AT THE TIME OF THE CHALLENGE.
timePoint	numerical	The time point for which the specific results reported in each form correspond. All the variables before, are experiment dependent. Within the same experiment, same animal group, when results were recorded at several time points, this field is used to identify at which time point the results being reported refer to.
experimentStatus	categorical	Given the different ways papers reported the experiment timeline, and even the occasional reporting of "before challenge" or "after vaccination", without a specification of the number of days elapsed, we have categorized the experimental stages as: • Not vaccinated/before vaccination • 1st vaccination day • Between vaccination doses • 2nd vaccination day • Vaccinated, not challenged • Challenge day • After challenge
labTest	Categorical	Laboratory test used to detect infection of serological response.
labDescription	Free-text	Any further details about the test.
labTarget	Categorical	The target of the detection test/indicator of infection – for instance "antibody" or "nucleic acid".
matrix	Categorical	The material collected for testing (specimen). Coded by EFSA. Collected in the form as a RADIO LIST.
nTested	numerical	Number of animals tested.
nPositive	numerical	Number of animals positive.
testValue	numerical	Used to report any specific test results (besides number of positives) which were given as numerical values.
testValueUnits	Categorical	Units for the value above.
scaleMin	numerical	When the interpretation of the test value depends on a scale (for instance clinical scores), these variables are used to report the range of the scale.
ScaleMax	numerical	
deadUnits	numerical	Number of animals dead.
mortalityTime	numerical	If provided in the paper, the time between the challenge and recorded mortality.
mortalityTimeUnits	categorical	The units of time for the value reported above.
efficacy	numerical	The efficacy, in percentage, of the vaccine when compared to a control group.
UCI_efficacy	numerical	Upper control interval for the efficacy reported.
LCI_efficacy	numerical	Lower control interval for the efficacy reported.
interventionType	categorical	Variable created to be able to separate control groups from intervention easily. Declared to be "control" if the testSubstanceCAT is declared to be "Placebo" or "Unvaccinated control"; and "vaccine" otherwise.
uniqueID	numerical	Unique identifier of groups for the whole dataset – it is a combination of refID and groupID
rowID	numerical	Unique identifier of 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.
challengeDayStd	numerical	Set to ZERO
dayDose1Std	numerical	As different studies reported different timelines, we kept the original timeline reported in the paper, but standardized all timelines to allow comparison by setting the challenge day to day ZERO, and calculating all other days in relation to the time until or since challenge
dayDose2Std	numerical	
timePointStd	numerical	
experimentStatus2	categorical	A simplified version of the experiment status, with only 4 categories: before vaccination; vaccinated by not challenged, challenge day, and after challenge.

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 a control and various treatment groups, groups of different species or age, or subjected to different experimental designs
3) Combinations of refiD+ studyGroupID represent UNIQUE animal groups for which results are reported (uniqueID). 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. In vaccine studies, the common practice is to have study groups followed over several time points, and for each time point, multiple tests could be applied and reported: for instance PCR to confirm infection, serological tests to measure immune response, observation of clinical signs, etc. There will be multiple rows of data for each group, and all the results referred to the same group can be tracked by the uniqueID. The timeline can then be reconstructed using the "timepoint" (please note we have created a "timePointStd" were the challenge day was set to zero in all studies).