Using SQL and R to simplify the analysis and reporting of a multi-target, individual animal African horse sickness surveillance program in South Africa

JD Grewar[1*], CT Weyer[2], LS van Helden[3]
[1]Equine Health Fund, Wits Health Consortium, Johannesburg, South Africa
[2]Equine Research Centre, University of Pretoria, Pretoria, South Africa
[3]Epidemiology Section, Veterinary Services, Western Cape Department of Agriculture, Elsenburg, South Africa
*johng@witshealth.co.za
Keywords: SQL, R, surveillance, African horse sickness

Abstract
The sentinel surveillance program implemented in the African horse sickness (AHS) surveillance zone of South Africa exists to confirm freedom of AHS in the zone but is undertaken in challenging conditions. There is the month to month individual animal sampling with testing using both serology and real-time PCR from different subsets of the sampling frame. The routine blood tests performed do not have DIVA (differentiating infected from vaccinated animals) capability. Subsequently, both controlled vaccination and previous outbreaks in the zone create an environment where recruitment of sentinels becomes difficult. Furthermore, the sentinel program forms only a small part of the larger AHS management system, which incorporates testing, census, vector, climate, movement and outbreak investigation components. The sentinel program, therefore, needs to conform to the vertical, cascading nature of the management database while at the same allowing for simple and repeatable analysis. In this report we detail how the querying and analysis of the AHS sentinel program in South Africa are performed from a centralised PostgreSQL environment using structured query language and R, allowing the automated reporting of an ongoing, dynamic surveillance program.

Introduction
African horse sickness (AHS) is a disease which has a significant impact on the export of live horses globally (1). This is particularly true for South Africa, where AHS is endemic. Historically South Africa's primary, non-African trading partner of choice has been the European Union. This trade is facilitated by a protocol that is specifically focused on control and testing of AHS (2). One of its requirements is a sero-sentinel surveillance program which is to be carried out monthly in South Africa's legislated AHS surveillance zone. South Africa's government and equine industry have, in an attempt to improve the sensitivity and relevance of this surveillance, included PCR-based surveillance to the serological surveillance, striving for an overall 2% minimum expected prevalence detection of potentially circulating AHS.

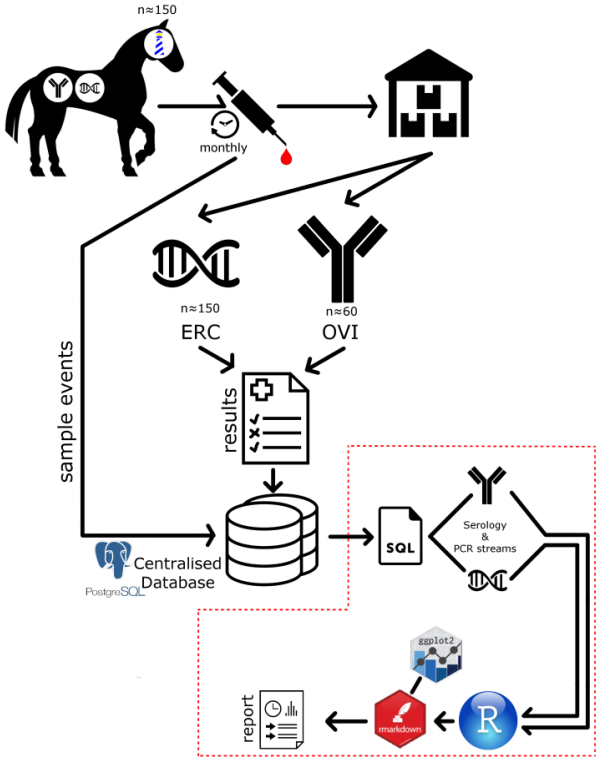
There are a number of challenges to overcome concerning this surveillance program. Neither the currently available serological test (indirect ELISA – i-ELISA) nor the RNA detecting real-time PCR used in South Africa for AHS have DIVA capabilities. Furthermore, for serological analysis, with monthly intervals between paired samples, the i-ELISA is for all intents and purposes non-quantitative and sentinels are evaluated by the permutations changing between positive, negative and suspect results for individual recruits. The recruitment of unvaccinated, previously unexposed horses for sero-surveillance is challenging. Vaccination in the surveillance zone is prohibited unless permission is given by the competent authority; however, a large proportion of the horses in the zone move to other control zones for competition purposes, and return movement to the surveillance zone require vaccination. Over and above this there have been previous outbreaks of AHS within the surveillance zone, further diminishing potential sero-sentinels. Recruitment of the added RNA sentinels is easier, although prior knowledge of vaccination is essential. A further less technical, but important hurdle is the fact that not all recruits are available every month which complicates the automated analysis of results.

Surveillance management systems ideally should be integrated into existing data capture systems. This is particularly true for a sample type such as whole blood or serum since sampling events of this nature will invariably form part of a larger system. In South Africa, the AHS sentinel surveillance makes up only one component of a greater AHS management system that incorporates the individual testing of horses for pre- and post-movement, outbreak investigation and passive surveillance testing. Horses are also registered on this system to manage census and vaccination data for holdings and horses respectively. Surveillance systems in their most simplistic form need to be capable of the efficient capture of sampling events, sample results and facilitate the automated reporting of surveillance outcomes. Figure 1 shows the basic schema for the South African sentinel surveillance system with the pertinent aspects of the analysis system described in this abstract highlighted in red stipples.

Structured Query Language (SQL) is used to communicate with databases and is the standard language for relational database management

systems. It is relatively generic across multiple database platforms. The South African AHS management system is based on a centralised, relational, object-attribute-value (OAV) PostgreSQL (3) database structure. R (4) is a free software environment used worldwide for statistical computing and graphics. It consists of base packages and collaborative packages written by various authors. In the AHS sentinel surveillance system, the RStudio (5) interface for R is utilised, improving the ease of use of the R functionality.

Figure 1: A schematic representation of the African horse sickness sentinel surveillance system. The section highlighted in red stipples relates to the methods used to automate the querying and reporting of the outcomes of the system where the serological and PCR-based testing is split and re-merged to accommodate the vertical structure of the centralised database. Using a combination of ggplot2 and rmarkdown the reporting is outputted into HyperText Markup Language (HTML). ERC – Equine Research Centre (PCR testing); OVI – Onderstepoort Veterinary Institute (serological testing)



The goal for the outcome of the analysis of the AHS sentinel surveillance program is: to capture surveillance events with minimal data structure changes and reference to the fact that said samples are for the sentinel program, and in a format that allows other sampling events in the management plan to be captured in the same table structure; to automate the querying of the data in the two streams of output i.e. the serological based versus PCR-based surveillance; to merge these two streams of output so that analysis can incorporate both surveillance methods given that holdings and horses can be recruited for both types of surveillance event; finally to graphically show this output automatically with minimal input by the end user.

Materials and Methods

The technical challenges faced in the system are primarily two-fold. The serological surveillance is analysed on a month to month basis (i.e. the change in serological status from one month to the next). However, should a horse not have been sampled in the month prior to the month under evaluation, then it is necessary to step back two months from that month under evaluation to determine the primary result in a paired series. Should a 2-month prior sample not exist then finally a 3-month backwards step is required. This highlights the second challenge. This surveillance is captured in an OAV format in relational sample event, detail and results tables. A sample event (horse and date specific) is linked to a sample detail table where a serum and/or EDTA sample is inputted. Each sample type result is then captured in a linked table referencing the relevant i-ELISA or PCR test. The vertical nature of the database is essential since the addition of each month as a new column for each horse is very bulky and creates a sparse matrix type result set with many NULL data entries, particularly since it's a long-term surveillance program with horses entering and exiting throughout.

SQL process: For the sake of brevity only the serological stream is considered here (the PCR stream consists of relatively simple SQL) and only those functions which make it possible to convert and analyse the data are included. A [cast(date_trunc(month, sampledate) as date] function allows each sample date to be pushed to the first of the month it was sampled in since the resolution of analysis is monthly. A [to_char(samplemonth::date,'J')] function assists in creating a unique Julian date character based identifier for each horse's sample event, which is used as a group-by variable later in the sequence of queries. A [cast(date_trunc('month', sampmonth - interval '1 month') as date)] function allows a separate column to be generated for each sample date indicating within which month a valid prior result would need to exist for a paired sample to be available for each horse month evaluated.

Nested queries: The ability of SQL to allow nested queries is pivotal for this analysis to work. It allows the same query (in this case results per horse per evaluation month with additional calculated prior month fields that would form part of a paired series) to be evaluated against itself. The product of this query establishes, for each unique horse sample month, whether there is a prior result in the previous month in each series. If there is then those data can

be isolated. If not the nested query can be looped to check for data going back two months and then three months from the month under evaluation respectively. The isolation of data is achieved using a [group by] query, where if a paired result is then these data are no longer available in the 2 and 3 month prior loops. Once all three loops are complete then a [union] based select query is used to merge all data into a single data series. These data are still vertical. To create a horizontal result set (i.e. for each horse the primary month date and result and the evaluated month date and result to create a paired series), two [group by] queries using [min(samplemonth)] and [max(samplemonth)] functions per horse are used to retrieve the result and date for the series data. In the final query a [case()] function is used (essentially a standard "if…else" function), that works through the permutations of the result series and allocates an outcome, as shown in Table 1.

Table 1: Permutations for each paired serology sample set outcome per horse per month of evaluation using i-ELISA testing.

Secondary Result / Primary result	Positive (+)	Negative (-)	Suspect (?)
Positive (+)	Stable +	+ → -	+ → ?
Negative (-)	- → +*	Stable -	- → ?*
Suspect (?)	? → +*	? → -	Stable ?

*permutations that require a response since these are potential indications of circulation

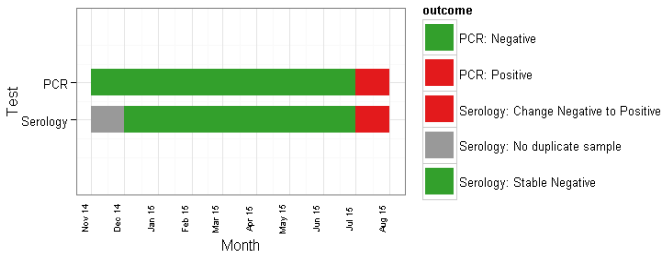
R and rmarkdown: PostgreSQL databases have the ability to establish views, which are stored queries. Views can be based on already created views, simplifying the initial setup of the system, while allowing one final view to facilitate the transfer of data to R. This final SQL string is used to pull the data to evaluate into the R environment from the web-based centralised database using a package called RPostgreSQL (6).

Using ggplot2 (7), the [geom_segment] option is used to handle the month to month timeline structure of the data, creating automated graphs for individual animals using a [for] loop for every unique horse that has any results which need following up. rmarkdown (8) is used to output the standard report in an HTML format. This package also allows the input of user parameters – in the AHS analysis script a date series is used, allowing the automated filtering of reports for a specific period of time with minimal programming knowledge required.

Results

While there are a number of descriptive statistics performed on the data, the essential HTML output focusses on individual horses that have had a suspect/positive change in serological status or a positive PCR result. The [geom_segment] output indicates both the PCR and serology status as time continues per horse requiring follow-up (Figure 2). In the illustrated case the sentinel had been vaccinated between the July and August 2015 sampling events and returned both a positive PCR and a change in serological status from negative to positive.

Figure 2: The base individual horse segment ggplot graph generated in the rmarkdown script showing the PCR and serological status for a single horse over time.



Discussion

The surveillance system discussed here lends itself to a relational database of an OAV model setup (i.e. vertical database model) since the attributes used for the analysis are defined in the surveillance plan and are not dynamic (i.e. only EDTA and serum samples are routinely tested using specific PCR and i-ELISA tests). The major advantage of using an OAV scheme is that the dynamic nature of the surveillance system (new horses being added, recruits falling out of the system and ongoing monthly samples being collected) allows the addition of horses and omission of monthly sampling without creating a sparse matrix of NULL data. At the same time database administration is minimised by not having to continually add monthly columns for each surveillance period. We show that the OAV data setup, with the help of certain SQL functions, allows the data to be converted into a horizontal result set which is more useful in an analysis environment like R.

In R the data is easily accessed from an online environment and processed using rmarkdown. The reports are standardised and can be easily compared, and include allowing basic user parameter input. While it has not yet been included it in the monthly routine analysis, the spatial component of the survey could also be evaluated within R to report on proportional sampling per area targets for the surveillance.

References

1. Zientara *et al.* (2015) OIE Rev Sci Tech 34(2): 315–327.
2. Anon. (2008) Off J Eur Union 2.9.2008(235): 16–25.
3. The PostgreSQL Global Development Group https://www.postgresql.org

4. R Core Team (2016) R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org
5. RStudio Team (2015) RStudio, Inc., Boston, MA http://www.rstudio.com
6. Conway *et al.* (2016) https://CRAN.R-project.org/package=RPostgreSQL
7. Wickham H (2009) Springer-Verlag New York
8. Allaire *et al.* (2016) https://CRAN.R-project.org/package=rmarkdown

Acknowledgements

We are grateful to the AusVet consultancy, in particular, Angus Cameron, who has over a period of time influenced our thinking regarding practical surveillance in challenging environments, promoting the use of open-source data management systems and the capturing of atomic based data. The innumerable contributors to the development of R have collectively created an exciting, dynamic environment which makes practical epidemiology in resource challenged Africa possible. We are grateful to the horse owners of the described surveillance system who ensure a sentinel-based surveillance system is possible. We are also grateful to Dr Phillippa Burger and Mrs Esthea Russouw who are responsible for the field sampling and data capture in this surveillance system. Funding of the overarching system is jointly provided by the Equine Health Fund (Wits Health Consortium) and the Government of South Africa.