How to design a second source for an effective capture-recapture analysis? The example of foot and mouth disease in Cambodia

T. Vergne[1, 5 *], V. Grosbois[1], S. San[2], T. Sothyra[3], F. Goutard[1], B. Bonté[1], A. Bouchot[4], F. Roger[1], B. Dufour[5]
[1] CIRAD, AGIRs unit, Montpellier, France
[2] National Veterinary Research Institute, Phnom Penh, Cambodia
[3] Department of Animal Health and Production, Phnom Penh, Cambodia
[4] OIE, World Organization of Animal Health, South East Asia sub-Regional Representation, Bangkok, Thailand
[5] ENVA-Afssa, USC EpiMAI, Ecole Nationale vétérinaire d'Alfort, Maisons-Alfort, France
* Corresponding author

Keywords: capture-recapture, simulation, foot-and-mouth disease, Cambodia

Introduction

Eradication of foot-and-mouth disease (FMD) in Southeast Asia by 2020 is the main objective of the "South East Asia Foot-and-Mouth Disease" (SEAFMD) campaign lead by the OIE sub-Regional Representation in South East Asia (1) who has set and now animates a passive FMD surveillance network in Southeast Asia. In Cambodia, as well as in most of the other infected countries participating to the SEAFMD campaign, under-reporting of FMD cases is an admitted fact all along the reporting chain from field to central and then regional (SEAFMD) level. A quantitative evaluation of the countries' reporting performances would thus be most useful. Capture-recapture (CR) methods could be a low cost and efficient tool for assessing these performances within a relatively short time frame.

One of these CR techniques consists in gathering, comparing and matching the cases collected by several independent surveillance sources for a focal population and time period. The analysis of such multi-source data allows one to obtain estimates of the total number of cases in the focal population and period (2,3). Usually, these analysis are performed with existing data, which can introduce uncontrolled bias if the sources fail to fulfil the conditions of application of CR methods (4). Here, we propose to combine an existing case reporting source with a new source designed by ourselves specifically for the CR analysis of FMD outbreaks reporting system in Cambodia during the year 2009. The existing source of information is the official database from the SEAFMD campaign. The second source, which we plan to generate, will be a retrospective survey using a participatory approach in villages selected independently from the official notifications. This kind of approach has already been applied in Thailand, but a very low overlap between the two sources (1 case) hampered the computation of robust estimates of the true infected population size (5). In this paper, we want to investigate two different sampling strategies for developing a second source: a random sampling and a targeted sampling. For the latter strategy, more effort is invested for sampling epidemiological units presenting relatively high probabilities of infection. With a same number of epidemiological units sampled, we expect that such a targeted strategy results in higher case detection probabilities and greater overlap with the pre-existing source than a random sampling strategy. We simulated both strategies on a virtual epizootic in order to evaluate. accuracy and precision of the official source sensibility estimate

Materials and methods

Capture-recapture with two sources: Suppose that a diseased population is screened by two imperfect surveillance systems. The system 1 detects *x* cases and the system 2 detects *y* cases. Among these *x+y* cases, there are *a* cases that are detected by both systems, *b* cases that are detected only by system 1 and *c* cases that are detected only by system 2 (*a+b=x; a+c=y*). The objective is to determine the number of cases which are not detected. Under the assumption of independence of the two sources, the probability of being detected by both sources equals the product of the probabilities of being detected by each source. This leads to the Chapman's estimate $\widehat{N_c}$ (6) of the true infected population size:

$$\widehat{N_c} = a + b + c + \frac{bc}{a+1}$$

This estimate can lead to an estimation of the sensitivity of one system, dividing the number of cases detected by this system by the Chapman's estimate. One major assumption must be met for getting unbiased estimates of the sensitivity: the two surveillance systems must be independent, *i.e.* being detected by one system doesn't change the probability of being detected by the other one (2,3).

Simulation on a virtual epizootic: A virtual epizootic was obtained by running one simulation of a simple epizootic model of an infectious disease through a 1,000 nodes scale-free graph (7). Nodes represent the epidemiological unit of interest (villages), and links represent the epidemiological connections between the nodes (like spatial proximity or cattle movement) which allow the disease to spread from one node to another one. At the end of the virtual epidemic, we get a final pattern with 202 infected nodes. On this distribution of infection, we simulated a first imperfect surveillance system (called source 1). To construct the source 1, we considered that all the cases in the simulated population have the same probability of being detected by the surveillance system. In order to evaluate the sensibility of this simulated source 1, we developed a second surveillance system (called source 2) independently of the first one. The source 2, specifically designed for the CR experiment will be simulated considering two possible sampling strategies. The first strategy consists in randomly sampling epidemiological units (same strategy as in source 1). The second strategy, which we plan to apply on the field consists, for a first phase, in focusing the sampling on the nodes which are more likely to be infected (*i.e.* nodes connected to 4 other nodes or more), and, in a second phase on all the nodes linked to an infected node detected during the first phase. This second approach is less time and money demanding and will lead to higher detection probabilities. For each simulation, we estimated the whole number of infected nodes thanks to the Chapman's estimator which permitted to calculate the sensitivity of the first source. For each strategy, 1000 simulations were run with the free software R (8).

Results

The true value of the first source sensitivity was fixed at 0.3. For each simulation, the repartition of the cases detected by the sources led to the Chapman's estimate and to an estimate of the first source sensitivity. Then we calculated the mean and the 95% confidence interval of the sensitivity for the two different strategies (random and targeted) for the 1000 simulations. On average, 180 nodes were sampled for the random approach versus 182 for the targeted one what can be considered as close enough to compare the results. The results are presented in Figure 1.

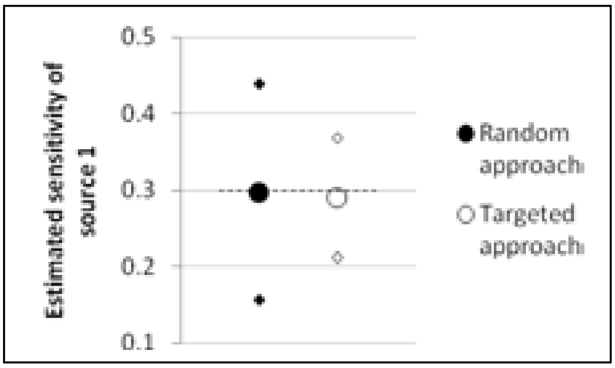

Figure 1: Estimated sensitivity of source 1 through the two approaches (the circles represent the mean of the estimates, the diamonds represent the bounds of the confidence interval, and 0.3 is the true value of the sensitivity of source 1)The two strategies led to two different results. Both allow one to calculate quite an unbiased estimate (relative bias equals -0.01 and -0.03 for the random and the targeted strategy respectively). However, even if the two estimates are unbiased, their precision differed. The random strategy resulted in a large 95% confidence interval (coefficient of variation = 0.07), whereas the targeted strategy led to a more precise estimate (coefficient of variation = 0.04). A variance analysis of the two distributions thanks to a Fisher test revealed that the difference of the two variances was significant ($p < 10^{-3}$).

Discussion

In a two-source capture-recapture experiment, if one of the sources detects randomly the epidemiological units of interest, then the Chapman's estimate will be unbiased, even if the other source experiences strong heterogeneity in the probabilities of detection (2). That's why, both approaches led to very accurate estimates. The problem with the random strategy is the large variance linked to the low number of detected units. On the contrary, the targeted strategy is much more precise because more units are detected producing a higher overlapping fraction (*i.e.* higher number of units detected by both sources). However in field studies, all the sources present heterogeneity of detection and this can lead to biased estimates if the heterogeneity of detection by the source 1 is linked to the heterogeneity of detection by source 2. So the challenge in the study we want to undertake will be to implement a second source, with a potential heterogeneity of detection, but whose heterogeneous subgroups are as independent as possible of the ones from the official notifications.

Conclusion

Regarding these results, it appears that if one wants to avoid a very low level of overlapping and thus a very large confidence interval of the estimate, as Cameron *et al.* (5) were confronted with in Thailand, one should adopt a targeted strategy for the second source. That is what we will try to do in Kampong Speu province, Cambodia, in order to evaluate the sensitivity of the reporting system to the OIE. We consider that the epidemiological unit of interest is a village with at least one herd possessing at least one clinically affected cow, which is consistent with the SEAFMD campaign outbreak definition. Many factors are known to increase the risk of FMD occurrence so in order to target our second source we will select some of the main risk factors and we will visit villages presenting these risk factors. Participatory Disease Searching, which is an inductive process of disease investigation, will then be applied to provide evidence that the disease was or wasn't present in the past. In case of a high suspicion, serum samples of one year old calves will be analyzed by an ELISA test for confirmation. If the former circulation of the virus (in 2009) can be highlighted, then the village is considered as a case for the capture-recapture analysis, and the second sampling phase is set up: a sample of the villages related to the detected one by cattle movement, are then visited. By exploiting the information available, like risk factors and animal trade between villages, we can generate a second source for the investigation of FMD outbreaks which is similar to the targeted source in the simulation study. As was shown in the results described above, this approach will allow calculation of a precise estimate. Nevertheless, the bias resulting from the dependence of the sources cannot be evaluating by a simulation mean, so the field work activity will be the most helpful tool to discuss these potential bias.

References

(1) Seafmd, 2007. *http://www.seafmd-rcu.oie.int/documents/SEAFMD%202020%20WEB%20Version.pdf*.
(2) Hook and al., 1995. *Epidemiol Rev*. **17**: 243-64.
(3) IWGDMF, 1995. *Am J Epidemiol*. **142**: 1047-58.
(4) Cormack, 1999. *J Clin Epidemiol*. **52**: 909-14.
(5) Cameron, 1999. *Survey tool box*. **8**: 183-187.
(6) Chapman, 1951. *U California Publ Stat*. **1**: 131-160
(7) Pautasso, 2008. *Ecological complexity*. **5**:1-8
(8) R Development Core team, 2008. *http://www.r-project.org*.