A Review of Current Methods of Analysis of Surveillance Data: A Way Through The Jungle

Hesterberg, U., Kosmider, R., Stewart, I., Ortiz-Pelaez, A., Del Rio-Vilas V. and Cook, A. J.C.
Veterinary Laboratories Agency, Weybridge, United Kingdom

Abstract:

Surveillance activities generate a large amount of data that are collated, analysed and reported in many ways. In the last twenty years, there has been a rapid development in health related surveillance producing a vast literature of peer-reviewed publications, conference proceedings, reports and other "grey" literature sources on the analysis of surveillance data. As part of a larger project, a literature review was conducted to identify analytical methods that are currently in use within surveillance systems, to describe the context in which they are used and what the requirements of the different methods are. To accomplish this the Medline, CAB Abstracts, Global Health and Embase, databases as well as books, conference proceedings and other grey literature were searched. In total 228 papers were reviewed and information on objectives, data and software requirements, advantages and disadvantages of the methods were recorded. The methods identified were categorized into the areas of prospective surveillance (temporal and spatio-temporal) for early disease detection, datamining, analysis of surveys (prevalence, test performance and sample size determination), prevalence estimation and evaluation of surveillance systems applicable to non-survey surveillance, estimating reporting delays and spatial analysis (disease mapping and spatial retrospective cluster detection).

Introduction:

Veterinary surveillance collects, processes and analyses animal health information through continuous scanning surveillance and structured surveys to allow decision makers to monitor and promote animal health, protect public health from disease of animal origin and protect international markets. New opportunities have arisen for the enhanced analysis of such data through improved technologies for animal identification combined with efforts of improving data quality and standardization, facilitating linkage of different datasets and developing the central collation of existing data. In Great Britain, the development of the RADAR data warehouse as a part of the Defra's Animal Health and Welfare strategy (Defra 2003) promises easier access to a greater range of animal health data in the future. We conducted a literature review of analytical methods for surveillance data to identify methods that are applicable for the enhanced use of these animal health data in order to enhance the speed and accuracy with which endemic and new diseases can be identified.

Materials and Methods:

A number of pilot searches were conducted in the PubMed bibliographic database in order to define appropriate search terms. Then we searched Medline, CAB Abstracts, Global Health and Embase, databases by means of three searches which included the following terms: surveillance, disease, population, statistic, method, application, model, approach, or analysis. (Details available from the authors on request). In addition, literature was sourced from books, conference proceedings and grey literature reports. The dates of publication for the literature database search were restricted to 1990 – 2005 inclusive.
More than 11,000 citations were considered for inclusion in the review. One panel member excluded obviously irrelevant papers; less obvious exclusions were checked by a second panel

member. References were assessed by at least two reviewers before inclusion. We concentrated on methodological references rather than applications of the methods, new developments, and references that we considered important for the improvement of surveillance. In total, 228 references were reviewed and information on objectives, data and software requirements, advantages and disadvantages of the methods in each paper were collated in an Access database. (Microsoft Corporation, 2000).

Results and Discussion:

Prospective surveillance:

Prospective surveillance aims to alert to significant increases in disease reporting which may be indicative of an outbreak situation. Upon raising the warning signal, an epidemiological investigation may be undertaken to ascertain whether it is an actual outbreak. There is a wealth of different methods described for purely temporal prospective surveillance which include statistical process control methods, regression analysis, and time series. This analysis is more frequently undertaken on temporal data but it is increasingly being applied to spatio-temporal data.

Integration of the spatial aspect into real-time surveillance allows for identification of local outbreaks where case numbers might not be sufficient to raise the overall level of cases to a degree where an alarm is triggered and allows for rapid identification of the geographical area the elevated rates originated from and consequently where further investigation and control may be required.

Methods identified in this section originate from the area of cluster detection, temporal surveillance and spatial modelling, and were adapted to the setting of prospective spatio-temporal surveillance. Based on the literature review findings, no method appears to be uniformly better than another method and it is probably advantageous to apply more than one method simultaneously. These methods were primarily applied to alternative 'syndromic' data sources in the public health sector, such as the initial complaint at arrival at an emergency department, school absenteeism, drug sales, or telephone calls to a national help line. Most methods require several years of historical data to integrate seasonal and other influencing effects into the expected counts to which observed counts are compared. Besides this, spatial case and baseline population data, that are recorded at regular intervals, are required. The efficacy of these methods depends to a large degree on how well the expected counts can be modelled.

Data mining:

Data mining techniques are applicable to high dimensional databases, where data are missing or where only a small number of records are matched with a large number of variables and is all about searching patterns in data. We identified applications of data mining to: detection of abnormalities in the prospective surveillance setting; reducing the dimensionality of a large database; searching for associations between various drugs and various side effects; predicting whether a subject is likely to be a case or not based on available attributes; and finding the level of a variable from a hierarchical tree of a variable family that is most likely to be responsible for rejecting the null hypothesis of no association. It is, however, felt that, with the applied search strategy focusing on health surveillance, only a few papers in this field have been identified, as most papers will probably be found within the information technology and computation fields.

Analysis of surveys (prevalence, test performance and sample size determination):

Stochastic simulation, using Markov Chain Monte Carlo Simulation (MCMC) and Bayesian modelling, has increased the feasibility of incorporating high levels of complexity into the output parameters of surveys: sample size calculation, prevalence estimation, substantiation of disease freedom, etc. Multi-level sampling, multiple testing regimes with imperfect tests, very low prevalence events under surveillance, etc., are no longer constraints, as they were for the traditional probabilistic methods. The iterative approach to incorporate variability and uncertainty, and the risk-based approach to the estimation of the outcome, will definitely help policy makers make

informed decisions on the implementation of surveys, or justify statements concerning disease status at regional or national level.

Prevalence estimation and evaluation of surveillance systems applicable to non-survey surveillance:
In public health, capture-recapture (CRC) methods have been applied for the estimation of the prevalence and incidence of a particular event. This approach has been applied successfully in a veterinary surveillance setting to estimate the prevalence of Scrapie in Great Britain. CRC methods require either repeated entries in one list (count models) or several sources that target the same event (multi-list methods). As a by-product of their main objective, the estimation of parameters in unobserved populations with a characteristic of interest, CRC models can produce estimates of the sensitivity of a surveillance network. Prevalence and incidence estimates can also be obtained from the application of evidence synthesis methods. Of particular interest is the estimation of frequency from multiple and heterogeneous sources of evidence. These methods also aid understanding of the causes of heterogeneity within surveillance sources. Evidence synthesis advocates the use of all sources of evidence to inform decisions, and is applicable to any surveillance setting with multiple data sources informing the parameter of interest, which is generally a frequency estimate. Evidence synthesis methods have been applied to assess the consistency of individual surveillance data sources, making any biases explicit.

Estimating reporting delays:
Techniques to model reporting delay distributions, calculate an adjusted incidence and estimate its variance have been developed (Harris JE. 1990). These also provide reliable estimates for back-calculations and, through analysis of the effect of covariates, assessment of the surveillance system and of the reliability of comparisons between systems or countries. Methods vary, and include log-linear models or Poisson regression as well as those based on adapted survival analysis techniques with particular application for covariates. The vast majority of techniques have been developed for and applied to AIDS data. These techniques could be applied to veterinary surveillance data, although adaptation may be required for some methods.

Spatial analysis:
Disease Mapping: Almost all methods identified through this review focus on mapping relative risk of disease by means of Bayesian hierarchical models. These models apply smoothing, and allow for the inclusion of covariates, spatial dependencies and in some cases temporal components. Thus they give a display of relative risk, after the removal of so-called ‘noise’(Lawson 2005), which can be mapped for further hypothesis generation. Most of the methods reviewed were applied to the mapping of human cancer cases. These models though require expert statistical knowledge, careful evaluation of their fit and availability of spatially referenced covariates and ecological variables.
Spatial Retrospective Cluster Detection: Spatial cluster tests detect spatially abnormally distributed rates and test for their significance. They can either test for these rates throughout the study region (global cluster tests), localize clusters (general cluster tests) or test if there are elevated rates in the vicinity of a hypothesized risk source (focused cluster tests) (Besag and Newell 1991). To test for abnormal rates, cluster tests need to account for the spatial distribution of the population at risk and require either spatially referenced data on the population at risk or suitable control data. These tests were initially developed for non-infectious chronic diseases and limitations related to this must be kept in mind when applying and interpreting these tests in the infectious disease surveillance setting. Overall, cluster tests are considered to be valuable screening methods, especially in the setting of less transmissible infectious diseases, or conditions such as toxicities or antibiotic resistance. Ideally they are combined with a disease map.

As the list of references would be very long, we decided not to include them here, but a full list of references is available from the authors on request. A detailed report on the findings of this review is expected to become available in future from the DEFRA website.

References:

Besag J, Newell J. (1991). The detection of clusters in rare diseases. *Journal of the Royal Statistical Society* - Series A: Statistics in Society, 154, (1), 143-155

Defra (Department for Environment, Food and Rural Affairs) (2003). A strategy for enhancing veterinary surveillance in the UK, Defra Publications, London, UK

Harris JE. (1990). Reporting delays and the incidence of AIDS. *Journal of the American Statistical Association*, 85, (412), 915-923. (7582).

Lawson AB. (2005). Spatial and spatio-temporal disease analysis. *Spatial and Syndromic Surveillance for Public Health*, John Wiley & Sons Ltd, Chichester, England, 55-75

Microsoft Corporation, 2000, Redmond , USA