Could multivariate statistical analysis contribute to the detection of emerging diseases?

Hoinville, L.J.[1], Done S.[2], Gaudie C.[2], Snow L.C.[1], Miller A.J.[1], Cook A.J.C.[1]
[1]Centre for epidemiology & risk analysis, Veterinary Laboratories Agency , Weybridge, Addlestone, Surrey KT15 3NB [2]Veterinary Laboratories Agency Regional Laboratory, West House, Station Road, Thirsk, North Yorkshire Y07 1PZ

Abstract

Rapid identification of emerging diseases, particularly those with non-specific clinical presentation, is essential to facilitate the development of effective control policies.

Post weaning multisystemic wasting syndrome (PMWS) is a disease with non-specific clinical presentation that was first identified in the United Kingdom (UK) in 1999. Porcine circovirus 2 (PCV-2) is thought to be associated with the occurrence of disease but the virus is widespread in the pig population and cannot be used to confirm the presence of PMWS.

As part of a cross-sectional study to investigate risk factors for this condition four poor doing weaned pigs were selected from each of 85 randomly selected pig farms. These animals were subjected to a full post-mortem examination including histological examination of three lymph nodes with immunohistochemical detection of PCV-2 antigen. Multiple correspondence analysis and hierarchical cluster analysis were used to identify groups of similar animals within this population on the basis of the pathological lesions present. Changes in the groups identified within a population of PMWS negative animals when additional animals with or without PMWS were added to the population were examined to determine whether these methods may be useful in detecting new diseases.

When PMWS positive animals were added to the population there was an increased likelihood that groups containing more than 50% of newly added animals would be detected suggesting that these methods may be useful in identifying emerging diseases. Further testing of these methods using data collected as part of a routine surveillance system is required.

Introduction

Detection Of Emerging Diseases

The need for timely detection of new and emerging diseases to allow the implementation of effective control strategies has been highlighted by the recent occurrence of rapidly spreading diseases like Foot and Mouth disease or those with long incubation periods like Bovine Spongiform Encephalopathy in the animal populations of the United Kingdom. (Doherr & Audige, 2001, Thurmond, 2003, Mortimer, 2003, Kuzma, 2001). A symptom based surveillance system has been suggested as a means to increase the timeliness of disease detection in the human population (Okaka et al, 2002) and development of data analysis methods which exploit all available sources of information have been recommended for the detection of emerging diseases (Wagner et al, 2001).

Multivariate Statistical Methods

Multivariate statistical methods are used for the simultaneous analysis of several related random variables collected from a number of different subjects (animals or farms) (Manly, 2000). In our study these related random variables indicated whether a number of clinical and pathological findings were present. Information about the presence of each of these signs was collected from a large number of individual animals.

PMWS

PMWS emerged as a new disease in Canada in the 1990's (Harding & Clark, 1997; Harding et al, 1998) and was first reported in the pig population of the UK in 1999 (Gresham et al, 2003).

Study Objectives

The objective of this study was to make an initial assessment of whether multivariate analysis of clinico-pathological observations could contribute to the detection of emerging diseases.

Materials and methods

Data Used

The data used in this investigation were taken from a cross-sectional study of 85 farms which was designed to investigate risk factors for PMWS occurrence (Cook et al, 2004). During this study four unthrifty pigs between 4 and 18 weeks of age were selected for post-mortem examination from each farm by a veterinary surgeon. Each of these pigs was subjected to a thorough gross post-mortem examination of all organ systems with standardised recording of all findings on a specially designed questionnaire. All data were entered into an MS access database and transferred to STATA 8 (StataCorp, 2003) for statistical analysis.

Discussions within the project team identified the conditions most likely to be present in unthrifty pigs of this age and the clinico-pathological variables most likely to distinguish between animals affected with these conditions were selected from the data available. These variables included clinical and pathological indicators of PMWS presence and the other disease conditions thought likely to occur in this group of animals and some demographic variables. Information on the presence of specific aetiological agents (e.g. PCV-2) was not included as this information is not likely to be available for all animals included in routine surveillance data sets.

Analysis Methods

Previous multivariate analyses of these data used to develop a new case definition for PMWS (Hoinville et al, 2006) identified correspondence analysis and hierarchical cluster analysis as the most appropriate methods for identifying groups of similar animals using these clinico-pathological data. The 'epipath' case definition developed in the previous analyses of the data (Hoinville et al, 2006) was used to divide the population of animals examined into those with and without PMWS. 131 of the 327 animals in the data set with complete data available were classified as PMWS positive and 196 as PMWS negative using this definition.

Initially a population of 146 PMWS negative animals were randomly selected from the 196 PMWS negative animals and multivariate analyses conducted to identify groups of animals with similar clinical presentation. Multiple correspondence analysis of the selected variables was conducted first to simplify the data and create new variables that could be used as inputs for the cluster analysis. Hierarchical cluster analysis was then conducted to identify groups of similar animals using the principal dimensions from the correspondence analysis that accounted for 99% of the variation in the data as input variables. The Euclidean distance between individuals was used to assess similarity and the distance between groups was assessed using the two furthest apart members of each group. The number of groups present was assessed by visual inspection of the resulting cluster analysis dendrogram to identify the point on the graph at which the length of vertical lines dividing groups changed length. Clinico-pathological features present in at least 80% of the animals in a group were used to describe the main clinico-pathological features of each group.

Two further populations were examined using the same analysis methods. Firstly, the remaining 50 PMWS negative animals were added to the population and the correspondence and cluster analysis repeated to identify groups present in this larger population of 196 PMWS negative animals. Secondly, the correspondence and cluster analysis was repeated using the 146 PMWS negative animals originally selected and a random selection of 50 of the PMWS positive animals. In each of these analyses the population was divided into fourteen groups, the same number of groups created in the initial analysis of 146 PMWS negative animals. The re-allocation of the original 146 PMWS negative animals to new groups and the proportion of each new group that came from the 50 animals that had been added to the population were examined after the addition of PMWS positive and PMWS negative animals to the population. These changes in group structure helped to assess the stability of the cluster analysis and the ability of these methods to detect new diseases.

Results

The frequency with which the variables selected for use in our analyses occurred in the populations of PMWS positive and PMWS negative animals is shown in Table 1.

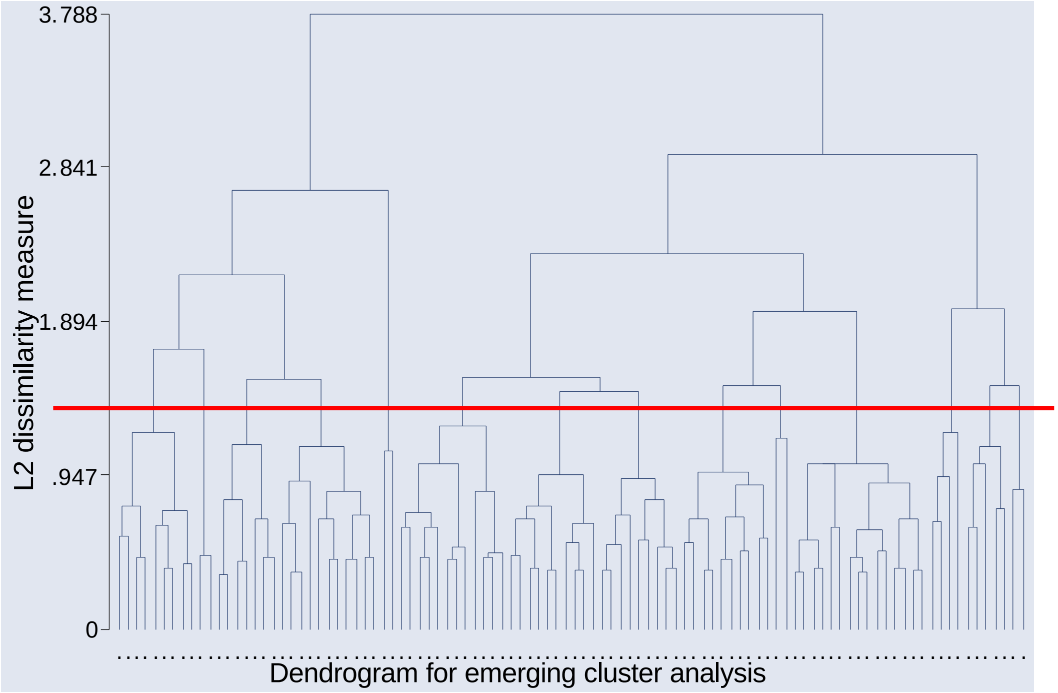
Table 1: Frequency with which selected signs occurred in populations of PMWS negative and PMWS positive animals

	PMWS Negative animals		PMWS positive animals	
	All animals (196)	Animals selected (146)	All animals (131)	Animals selected (50)
Conjunctivitis	6.1	5.5	14.4	18.0
Greasy pig	2.6	2.1	3.8	2.0
Cranial lung consolidation	52.0	54.8	66.4	62.0
Caudal lung consolidation	32.1	32.9	35.1	28.0
Pleurisy	32.1	34.3	30.5	30.0
Caudal lung haemorrhage or necrosis	7.7	7.5	3.8	6.0
Endocarditis	1.5	1.4	0.8	2.0
Pericarditis	16.8	18.5	21.4	26.0
Ulceration of caecum, colon or ileocaecal junction	14.3	13.0	16.0	20.0
Thickening of distal ileum	5.1	5.5	15.3	14.0
Peritonitis	31.6	34.9	24.4	26.0
Haemorrhage in lymph nodes	11.7	12.3	5.3	6.0
Infarct in spleen	3.6	3.4	4.6	10.0
Pale muscles	4.6	4.1	12.2	10.0
Multiple swollen joints +turbid synovial fluid	5.1	5.5	6.9	8.0
Superficial inguinal lymph node enlargement	26.5	27.4	77.1	74.0
Hairy skin	46.9	44.5	61.8	76.0
Pale skin	50.0	48.0	61.8	66.0
Pars oesophagus lesions	23.5	27.4	34.4	40.0
Flaccid heart	6.6	7.5	10.7	8.0
Large kidney	15.8	15.1	25.2	26.0
Fluid faecal consistency	39.8	37.7	43.5	46.0
Enlarged tracheobroncial LN	43.4	43.8	58.8	64.0
Enlarged ICC LN	25.5	24.0	43.5	46.0
Low weight for age	74.5	74.0	100	100
Region – other	41.3	41.7	17.6	18.0
Region – Scotland	5.6	4.8	6.1	4.0
Region – York & Humber	14.3	13.7	36.6	38.0
Region – East	38.8	39.7	39.7	40.0
Herd type - Breeder	66.8	67.1	55.0	58.0

Correspondence analysis of the 146 randomly selected PMWS negative animals revealed that the variation in the original variables could be explained using 11 principal dimensions. The first two principle dimensions in this data set explained 66% of the variation in the original data and 99% of the variation could be explained using the first 8 principal dimensions.

The dendrogram resulting from the hierarchical cluster analysis using the first 8 principal dimensions as input variables is shown in Figure 1. The red line shows the position at which the vertical lines on the dendrogram become longer, division of the population into groups at this level produced 14 groups of animals.

Figure 1: Dendrogram resulting from cluster analysis of 146 PMWS negative animals using first 8 principal dimensions created in correspondence analysis



The number of animals and main clinico-pathological features present in the animals in each of the 14 groups produced in the three separate cluster analyses are summarised in Table 2. When 50 additional PMWS negative animals were added to the population to produce a population of 196 PMWS negative animals the largest proportion of new animals in a group was 50% in one group with two animals one of which came from the additional 50 animals. When 50 PMWS positive animals were added to the same population of 146 PMWS negative animals there were three groups in which more than 50% of the animals present were derived from the population of 50 additional animals. One of these groups contained only 4 animals all of which were from the population of 50 PMWS positive animals.

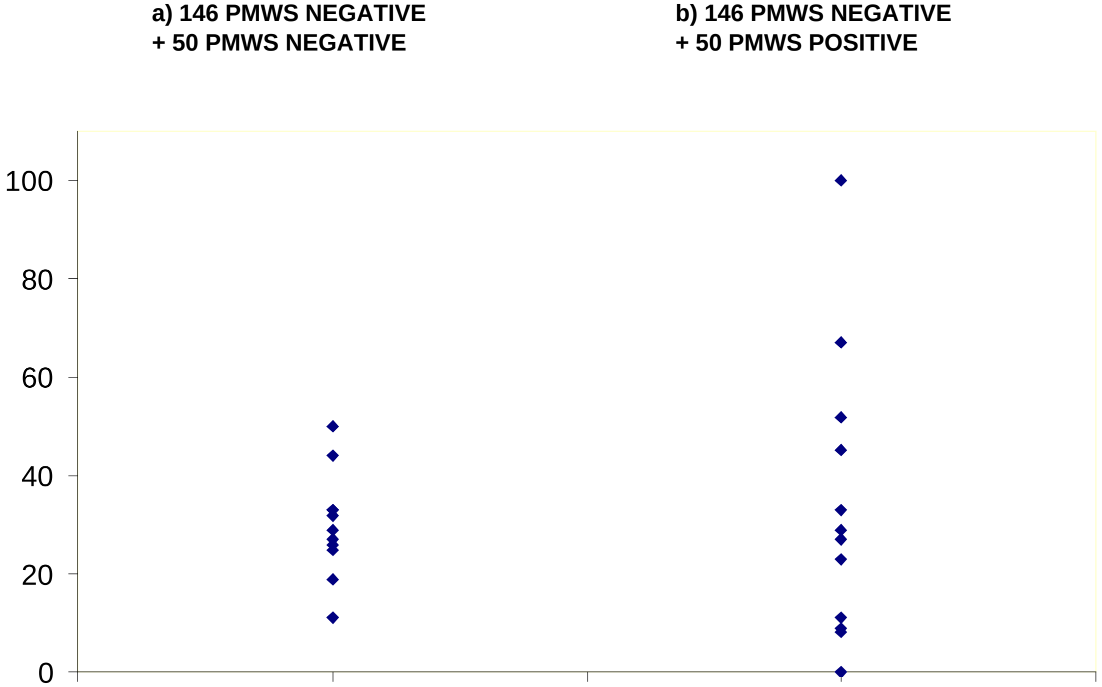
Table 2: Main clinico-pathological features, number of animals (N) and percent of group derived from additional 50 animals (% new) in groups formed in cluster analysis of three different populations

146 PMWS negative animals with additional 50 PMWS negative animals added			Original 146 PMWS negative animals		146 PMWS negative animals with additional 50 PMWS positive animals added		
Main clinico-pathological features	**N**	**% new**	**Main clinico-pathological features**	**N**	**Main clinico-pathological features**	**N**	**% new**
Pale skin, fluid faeces, wasting	34	32%	Pale skin, wasting	25	Pale skin, hairy skin, fluid faeces	33	27%
Wasting	42	26%	Wasting	27	Wasting	45	9%
Wasting, hairy skin	9	44%	Hairy skin	10	Skin hairy, SI lymph node large	14	50%
TB lymph node large	36	11%	TB lymph node large	20	TB and IC lymph nodes large	22	23%
TB and IC lymph node large	22	27%	TB and IC lymph node large, cranial lung consolidation	10	IC and SI lymph nodes large, pale skin, wasting, pars oesophagus lesions	4	100%
Cranial lung consolidation, pericarditis, pleurisy, TB lymph node large	11	9%	Cranial lung consolidation, pleurisy	17	Pleurisy. TB lymph node large	18	11%
Cranial and caudal lung consolidation, pleurisy, peritonitis, pale skin	8	13%	Cranial lung consolidation, pleurisy, pericarditis, TB lymph node large, pars oesophagus lesions	7	Cranial and caudal lung consolidation, TB lymph node large, pleurisy	12	33%
Cranial and caudal lung consolidation, pleurisy	9	11%	Pleurisy, TB, SI and IC lymph node large, skin hairy	3	Cranial lung consolidation, TB lymph node large, pleurisy	9	56%
Cranial and caudal lung consolidation, pleurisy, hairy skin, wasting	6	11%	Cranial and caudal lung consolidation, pleurisy, peritonitis, pale skin	12	Cranial lung consolidation, pleurisy	13	8%
Cranial lung consolidation, pale and hairy skin, wasting	7	29%	Cranial and caudal lung consolidation, pleurisy, hairy skin, wasting, fluid faeces, greasy pig disease	2	Cranial lung consolidation, pleurisy, pericarditis, peritonitis, fluid faeces, pale skin	11	45%
Caudal lung haemorrhage, spleen infarction, SI lymph node large	4	25%	Hairy skin, wasting	5	Spleen infarction, hairy skin	7	29%
Caudal lung haemorrhage, hairy skin, wasting	2	50%	Caudal lung haemorrhage, spleen infarction	4	Caudal lung haemorrhage, cranial and caudal lung consolidation, pleurisy, pericarditis, TB lymph node large, pars oesophagus lesions	3	0%
Cranial lung consolidation, SI and TB lymph nodes large, pleurisy, hairy skin, large kidney	3	33%	Cranial lung consolidation, pleurisy, TB and SI lymph nodes large, skin hairy, flaccid heart, pars oesophagus lesions	2	Caudal lung haemorrhage, cranial lung consolidation, pleurisy, pericarditis, peritonitis, TB and IC lymph node large, hairy skin	3	66%
Endocarditis, pericarditis, peritonitis, TB lymph node large, lymph node haemorrhage,	3	33%	Endocarditis, pericarditis, peritonitis, TB lymph node large, lymph node haemorrhage, pars oesophagus lesions	2	Endocarditis, pericarditis, peritonitis, TB lymph node large, lymph node haemorrhage, pars oesophagus lesions	2	0%

TB = tracheobronchial, IC = ileocaecocolic, SI = superficial inguinal

The distribution of the percent of new animals in each group when PMWS negative and PMWS positive animals were added to the population is shown in Figure 2.

Figure 2– Distribution of the percent of new animals in each group when a) 50 PMWS negative or b) 50 PMWS positive animals were added to the population of 146 PMWS negative animals



Discussion

The use of cluster analysis to identify groups of similar individuals is controversial due to the variation between the results obtained using different methods on the same data set and the difficulty in deciding how many groups are present (Everitt, 1979). It has been suggested that the results of cluster analysis should be judged largely by whether the classification produced is useful or not (Everitt, 1979; Anderberg, 1973). In these analyses it was possible to divide the population into distinguishable groups with different clinico-pathological features using these methods although there was considerable overlap between the features of animals in different groups.

There were some groups which appeared to be relatively stable when additional PMWS negative animals were added to the population but others groups from which animals were re-allocated to several different groups. This was consistent with the failure to identify very discrete groups and suggests a number of overlapping groups of animals with different combinations of respiratory and reticulo-endothelial lesions.

However, the main objective of these analyses was not to identify distinct groups of animals but to determine whether it is likely that a change in the pattern of groups present could be used to indicate the presence of a new disease. These analyses were able to identify a difference between the proportions of new animals present in groups when PMWS positive animals were added to a PMWS negative population compared to that when PMWS negative animals were added. When PMWS positive animals were added to the population there were several groups produced in which more than 50% of animals were from the additional animals added to the original population. This did not occur when PMWS negative animals were added. When PMWS negative animals were added to the population no groups were identified in which more than 50% of the animals were from the population of new animals added.

These results suggest that it may be possible to develop criteria which could be used to examine the results of cluster analysis of clinico-pathological data collected in different time periods that could be used to aid in the detection of new diseases.

Whether these methods would be sensitive enough to detect new diseases would depend on the number of cases of any emerging disease and how different the clinico-pathological presentation of the disease was from the presentation of those conditions previously present in the population. The degree of disruption of the groups of animals identified using clinico-pathological data when a new disease occurs will depend on the clinico-pathological features of the new condition and the proportion of the population examined that are affected with the new condition. In this study all of the animals added to the population were thought to be affected with a new disease and this is unlikely to be the case in a surveillance data set in which animals with a new disease will occur at the same time as animals with conditions already present in the population.

The ability to identify new diseases using these methods will also depend on the nature of the information available, information from a representative sample of animals over a sufficiently long time period with details of a variety of pathological lesions present would be required. The information collected would have to include a wide range of lesions in order to be sure of including information about lesions that may occur in diseases not yet identified in the population. The data used in this investigation were collected as part of a study to investigate the occurrence of PMWS and were not designed to discriminate between the major conditions likely to be present in the pig population. It is possible that if surveillance data were collected specifically to identify groups of animals with different diseases it would be easier to identify distinct groups of animals and to determine whether new conditions were occurring in the population.

In practice these analytical methods would be only part of a surveillance system designed to detect emerging diseases with other aspects of routine surveillance data collection, including field visits by practicing veterinarians and VLA veterinary investigation officers, being informed by and providing information to guide these analyses.

Although the inclusion of demographic variables had little influence on the results of the investigations conducted in this study it is possible that the use of additional information may help to identify new diseases occurring in certain types of animal or restricted to certain areas of country.

The sensitivity, specificity and timeliness of these methods in detecting new diseases could be tested if a suitable surveillance data set including sufficient clinico-pathological information was available. Ideally this data set would include cases of a disease that emerged in the population during the course of data collection.

Conclusions

Cluster analysis of clinico-pathological data may be useful in identifying the presence of emerging diseases although the sensitivity of this approach needs to be tested using a detailed surveillance data set which includes data from animals affected with a new disease. These investigations highlight the need for the collection of standardized clinico-pathological surveillance data which could be used for further evaluation of the methods presented here and for the detection of emerging diseases using these or other methods. One possible source of such data would be the syndromic surveillance data collected in VLA's 'farmfile' project

Acknowledgements

We are very grateful to the farmers for providing the information used in these analyses, to the veterinary surgeon's and VLA colleagues who contributed to the collection and management of the data and to DEFRA for funding this work.

References

Anderberg MR (1973). Cluster analysis for applications Academic Press New York.

Cook A, Gaudie C, Miller A, O'Connor J & Done S. (2004) Surveillance for PMWS in UK – results from a cross-sectional study. Pig Veterinary Society Spring Conference May 2004.

Doherr MG & Audige L (2001) Monitoring and surveillance for rare health-related events: a review from the veterinary perspective. *Philosophical Transactions of Royal Society London B* 356 1097-1106

Everitt BS (1979). Cluster Analysis. Heineman, London.

Gresham A, Cook AJC, Thomson JR & Kennedy S (2003). Survey of veterinary practitioners on PMWS and PDNS in the UK. *The Veterinary Record* 153 400-403.

Harding JCS & Clark EG (1997) Recognizing and diagnosing postweaning multisystemic wasting syndrome (PMWS). *Swine Health and Production* 5 201-203.

Harding JCS, Clark EG, Stokappe JH, Willson P & Ellis JA (1998). Postweaning multisystemic wasting syndrome: epidemiology and clinical presentation. *Swine Health and Production* 6 249-254

Hoinville LJH, Done S, Gaudie C, Snow L, Miller AJ & Cook AJC (2006) Defining PMWS: the use of multivariate statistical analysis to develop case definitions for emerging diseases. *Journal of infectious diseases* (submitted)

Kuzma CD (2001) Working together to fight emerging diseases *JAVMA* 218 1688-9

Manly BFJ, Multivariate statistical methods A primer (1986). Chapman & Hall, London.

Mortimer PP (2003) Five postulates for resolving outbreaks of infectious disease *Journal of Medical Microbiology* 52 447-51

Osaka K, Takahashi H & Ohyama T (2002). Testing a symptom-based surveillance system at high-profile gatherings as preparatory measure for bioterrorism *Epidemiology and Infection* 129 429-34

StataCorp (2003) Stata Statistical Software: Release 8.0 College Station, TX; Stata Corporation

Thurmond MC (2003) Conceptual foundations for infectious disease surveillance. *Journal of Veterinary Diagnostic Investigation* 15 501-514

Wagner MM, Tsui FC, Espino JU, Dato VM, Sittig DF, Caruana RA, McGinnis LF, Deerfield DW, Drudzel MJ& Fridsma DB (2001) The emerging science of very early detection of disease outbreaks *Journal of Public Health Management Practice* 7 51-9.