General enquiries on this form should be made to:
Defra, Science Directorate, Management Support and Finance Team,
Telephone No. 020 7238 1612
E-mail: research.competitions@defra.gsi.gov.uk

SID 5 Research Project Final Report

- **Note**
  In line with the Freedom of Information Act 2000, Defra aims to place the results of its completed research projects in the public domain wherever possible. The SID 5 (Research Project Final Report) is designed to capture the information on the results and outputs of Defra-funded research in a format that is easily publishable through the Defra website. A SID 5 must be completed for all projects.
- This form is in Word format and the boxes may be expanded or reduced, as appropriate.
- **ACCESS TO INFORMATION**
  The information collected on this form will be stored electronically and may be sent to any part of Defra, or to individual researchers or organisations outside Defra for the purposes of reviewing the project. Defra may also disclose the information to any outside organisation acting as an agent authorised by Defra to process final research reports on its behalf. Defra intends to publish this form on its website, unless there are strong reasons not to, which fully comply with exemptions under the Environmental Information Regulations or the Freedom of Information Act 2000.
  Defra may be required to release information, including personal data and commercial information, on request under the Environmental Information Regulations or the Freedom of Information Act 2000. However, Defra will not permit any unwarranted breach of confidentiality or act in contravention of its obligations under the Data Protection Act 1998. Defra or its appointed agents may use the name, address or other details on your form to contact you in connection with occasional customer research aimed at improving the processes through which Defra works with its contractors.

Project identification

1. Defra Project code: SE4002

2. Project title:
   Development of farm-specific biosecurity risk management strategies for cattle herds and sheep flocks

3. Contractor organisation(s):
   The Royal Veterinary College
   Hawkshead Lane
   North Mymms, Hatfield, Hertfordshire
   AL97TA
   United Kingdom

4. Total Defra project costs (agreed fixed price): £ 299,830.66

5. Project: start date ..... 01 April 2004

   end date 31/3/2006

6. It is Defra's intention to publish this form.
Please confirm your agreement to do so..... **YES** ☒ **NO** ☐

(a) When preparing SID 5s contractors should bear in mind that Defra intends that they be made public. They should be written in a clear and concise manner and represent a full account of the research project which someone not closely associated with the project can follow.

Defra recognises that in a small minority of cases there may be information, such as intellectual property or commercially confidential data, used in or generated by the research project, which should not be disclosed. In these cases, such information should be detailed in a separate annex (not to be published) so that the SID 5 can be placed in the public domain. Where it is impossible to complete the Final Report without including references to any sensitive or confidential data, the information should be included and section (b) completed. NB: only in exceptional circumstances will Defra expect contractors to give a "No" answer.

In all cases, reasons for withholding information must be fully in line with exemptions under the Environmental Information Regulations or the Freedom of Information Act 2000.

(b) If you have answered NO, please explain why the Final report should not be released into public domain

Executive Summary

7. The executive summary must not exceed 2 sides in total of A4 and should be understandable to the intelligent non-scientist. It should cover the main objectives, methods and findings of the research, together with any other significant events and options for new work.

The reduction of the risk of disease introduction into livestock farms is one of the objectives of DEFRA's Animal Health & Welfare Strategy. The spread of an infectious disease depends strongly on the characteristics of the causal agent. However, farm management practices can contribute significantly to the likelihood of disease spread between herds and flocks. Biosecurity is defined in this report as the measures taken on a farm to prevent the introduction of disease agents into a herd or flock.

In this project, a knowledge-driven approach was used to develop a risk scoring algorithm using risk factors affecting farm-level biosecurity which would require farm visits, based on a combination of information in the published scientific literature and expert opinion. Then, a data-driven approach was used to develop a biosecurity risk scoring algorithm for cattle and sheep farms based on externally measurable factors and therefore not requiring farm visits.

The approach chosen for the first objective of the project, the development of the knowledge-driven biosecurity risk scoring algorithm, included a systematic review of the published literature and an expert opinion workshop (EOW). The systematic review was used to identify risk factors that were significantly associated with introduction of disease. The EOW was conducted to check whether these risk factors were applicable to the UK and whether there were any additional risk factors. Based on the findings of the systematic review and the EOW, farm-level biosecurity risk scoring algorithms were developed. Both, the systematic review and the EOW were used to summarize knowledge regarding biosecurity management at the farm-level.

This systematic review of the literature focused on factors influencing farm-level biosecurity risk on cattle and sheep farms in the UK, using 10 diseases that were considered to be most important for the UK industry. From an initial list of 3,320 cattle- and 1,294 sheep-related publications, 29 cattle- and 8 sheep-related were considered suitable for inclusion in the review. Overall, direct contact between the same species was identified as the most important risk factor for disease introduction, with indirect contact having a less significant role. Information on best practice of farm-level biosecurity was also examined.

The EOW was effective for eliciting knowledge on risk factors and on the impact of preventive measures in a semi-quantitative fashion from a group of experts. The methodology is particularly useful in areas where limited published knowledge is available. But it needs to be recognised that the conclusions were based on data strongly influenced by subjective experiences. It is recommended that this approach is used in combination with systematic reviews, so that the consistency of the findings can be evaluated. In the current case, the findings of the systematic review and the EOW indicate a good agreement in the conclusions, and both suggest that purchase of livestock is the most important risk factor for potential introduction of infectious diseases to a herd/flock.

Based on the findings from the systematic review and the EOW, a farm-level biosecurity risk scoring algorithm was generated for both sheep and cattle farms.
The second objective of the project was aimed at developing a data-driven farm-level biosecurity risk scoring algorithm not requiring farm visits. Several analytical methods were applied to identify the most important variables from a range of externally measurable parameters extracted from various databases. Analyses were conducted separately for cattle and sheep farms. Subsequent to a small-scale pilot study of cattle holdings in Norfolk County, the main study was conducted using population data on cattle and sheep holdings in Wales in 2004, and aimed at identifying risk factors associated with the occurrence of endemic diseases. The database compiled from 12 data sources contained information about 15,845 cattle holdings and more than fifty variables. A subset of 33 variables was included in the final dataset for analysis. The two disease outcome variables were presence of bovine tuberculosis and a pool of 10 infectious and parasitic diseases. A second dataset was generated with 18,937 sheep holdings with 24 variables for analysis. The outcome was a pool of 9 infectious and parasitic diseases defined under Objective 1 of this project. The variables were divided into four groupings: Holding demographics, animal movements, densities and environmental variables. The statistical analysis was conducted using regression and data mining techniques, considering differential weightings of importance of false-positive and –negative farm classifications. For the outcome variable 'TB status' on cattle farms, the regression model indicated an increase of risk with increasing farmed area, cattle density, herd size and 'no. of days during 2004 a holding moved cattle off farm'. A decreased TB risk was associated with increasing total rainfall, average temperature, 'no. of Welsh holdings animals were moved to/from in 2004', 'being located within 5km distance from a nature reserve', number of linkages to other holdings in Wales and land cover. The classification tree analysis identified the following most important variables for TB risk: Cattle density, 'no. of days during 2004 a holding moved cattle off farm' and herd size. Using the outcome 'TB status or presence of other diseases' for cattle farms, the regression model indicated an increasing disease risk with increasing cattle density, herd size, 'no. of days during 2004 a holding moved cattle off farm' and 'no. of Welsh holdings animals were moved to in 2004', and if soil type was peat soils and marshes. Disease risk was reduced with increasing total rainfall, farm location being within 5km distance from a nature reserve and increasing 'no. of Welsh holdings animals were received from in 2004'. The results from the classification tree analysis for this outcome variable were the same as for the one for TB risk. Exploratory analysis of the movement data revealed a pattern where holdings characterised by only having movements within Wales having a different pattern compared with those also linking with holdings outside Wales. The latter group of cattle farms tended to have a higher risk of disease. The outcome variable for the analysis for sheep flocks was the presence of at least one from a group of sheep diseases. The significant risk factors included in the regression model for sheep farms were: farmed area, flock size, 'herd size if mixed-species holding', 'no of other holdings sheep were moved to in 2004', 'no. of sheep in 2004 moved on to the holding', total rainfall, average temperature and being within or at less than 5km of a nature reserve. Movement variables were the most important variables when identifying positive holdings using classification tree analysis. The results from the statistical analyses were used to develop semi-quantitative biosecurity risk scoring algorithms for cattle and sheep holdings. The criteria for inclusion of the variables in the algorithms were derived from their importance as a risk/protective factor in the logistic regression and the classification tree analyses. The algorithms were defined as a sequence of mutually exclusive criteria which were added to form an overall risk score. The criteria for cattle holdings included whether any cattle movements occurred, whether cattle movements occurred to/from other holdings in Wales, herd size, cattle density, farmed area, proximity to a nature reserve, amount of total rain, soil type and land cover. The criteria for sheep holdings included flock size, 'herd size if mixed-species holding', farmed area, average number of sheep movements per year on/off the premises and being close to a nature reserve.
Both biosecurity risk assessment algorithms can be used to conduct farm-specific risk assessments, which could then inform the development of tailored risk management strategies. Different approaches could be applied, for example the use of the externally measurable risk score could identify high biosecurity-risk farms and then be followed up with on-farm risk assessments that result in farm-specific risk management plans. Alternatively, both algorithms could be applied on farms participating in animal health plans. It should be noted that the performance characteristics of the algorithms (predictive values) in a field situation were not

evaluated in this study. As a next step, the approach needs to be refined and tested in epidemiological and economic terms on farms using intervention studies.

Project Report to Defra

8. As a guide this report should be no longer than 20 sides of A4. This report is to provide Defra with details of the outputs of the research project for internal purposes; to meet the terms of the contract; and to allow Defra to publish details of the outputs to meet Environmental Information Regulation or Freedom of Information obligations. This short report to Defra does not preclude contractors from also seeking to publish a full, formal scientific report/paper in an appropriate scientific or other journal/publication. Indeed, Defra actively encourages such publications as part of the contract terms. The report to Defra should include:
   - the scientific objectives as set out in the contract;
   - the extent to which the objectives set out in the contract have been met;
   - details of methods used and the results obtained, including statistical analysis (if appropriate);
   - a discussion of the results and their reliability;
   - the main implications of the findings;
   - possible future work; and
   - any action resulting from the research (e.g. IP, Knowledge Transfer).

Project Objectives

1) Conduct a systematic review of existing knowledge relating to methods for:
   - classification of cattle herds and sheep flocks according to biosecurity risk, and
   - enhancement of bio-security levels on cattle and sheep farms

   Use this information to develop knowledge-driven models for farm-level biosecurity risk scores and biosecurity management strategies

2) Identify risk factors associated with disease introduction to cattle herds and sheep flocks in GB based on retrospective data analysis of:
   - environmental factors, and
   - description of contact networks between herds/flocks

   Combine models from Objective 1 with risk factor information to develop data-driven models for farm-level biosecurity risk scoring systems

The general objective of this project was to develop a methodology for risk assessment and management of biosecurity on cattle and sheep farms in Great Britain. The combination of knowledge- and data-driven models was aimed at providing decision-support for defining management strategies aimed at reducing the risk of introduction of infectious diseases to cattle herds and sheep flocks. The delivery of practical evidence-based approaches aimed at enhancing farm-level biosecurity is one of the priorities of the DEFRA's new Animal Health and Welfare Strategy (AHWS) (Scudamore, 2004).

Objective 1: Systematic review and development of knowledge-driven models for farm-level biosecurity risk scores and biosecurity management strategies

One of the major objectives of DEFRA's Animal Health & Welfare Strategy is the reduction of the risk of disease introduction into livestock farms. The spread of an infectious disease depends strongly on the characteristics of the causal agent. However, farm management practices can contribute significantly to the likelihood of disease spread between herds and flocks. Biosecurity is defined as any practice or system that prevents the spread of infectious agents from infected to susceptible animals, or prevents the introduction of infected animals into a herd, region, or country in which the infection has not yet occurred (Radostits 2001). Another, more strict definition has been proposed by Dargatz et al (2002), who state that biosecurity is the outcome of all activities undertaken by an entity to preclude the introduction of disease agents into an area that one is trying to protect. Biocontainment is used to prevent spread of a disease within a herd or flock when the disease is already present. In this report the more narrow definition, i.e. prevention of the introduction of disease agents into a herd or flock, will be used as the definition of biosecurity.
The knowledge-driven part of this project consisted of a systematic review of the published literature and an expert opinion workshop (EOW). The systematic review was aimed at defining the risk factors

that were significantly associated with introduction of disease. The EOW was conducted to assess whether these risk factors were applicable to the UK livestock production system and whether there were any additional risk factors. Based on the findings of the systematic review and the EOW, generic templates for farm-level biosecurity scores were developed. Both, systematic review and the EOW were used to summarize knowledge related to biosecurity risk assessment and management at the farm-level.

Systematic Review

Introduction

The approach described in the Cochrane Handbook for Systematic Reviews of Interventions published by the Cochrane Collaboration was used as the basis for this systematic review (Egger and Smith 2001). A review protocol was produced in consultation with a group of experts from around the UK. In total, ten diseases for both cattle and sheep were selected to evaluate biosecurity. The following criteria were considered when selecting diseases for inclusion in the systematic review: Estimated herd prevalence in the UK (largely based on VLA surveillance reports), the economic impact of introduction into a herd/ flock, the zoonotic impact of the disease and the estimated availability of quantitative data on risk of disease introduction. For cattle, bovine tuberculosis, Johne's disease, salmonellosis, BVD, IBR, *Neospora caninum*, *Campylobacter fetus* ssp *venerealis*, leptospirosis, FMD and liver fluke were considered to be most relevant. In the case of sheep, multiple resistance to helminths, psoroptic mange, *Chlamydophyla* abortion, footrot, Jaagsiekte, scrapie, louping ill, caseous lymphadenitis, maedi visna and toxoplasmosis were the diseases focussed on in the systematic review.

Materials and methods

In order to select relevant peer-reviewed publications, the following internet search engines were used: PubMed, CABI and ISI Web of Knowledge (ISI WoK). Also, proceedings from World Buiatrics Congresses, conferences of the Society for Veterinary Epidemiology and Preventive Medicine and the British Cattle Veterinary Association were searched to obtain as much published material as possible. In the search, publications written in the English language were used primarily. Articles written in other languages were also included, when the reviewers were able to understand their content on the basis of the English language summary. Search phrases included the words: "introduction", "risk", "herd", "flock", "cattle", "cow", "bovine", "sheep", "ovine" and the specific name of the disease of interest. In general, the search formula was as follows: ('bovine' OR 'cow' OR 'cattle') AND ('herd' OR 'introduction' OR 'risk') AND ('disease name') for cattle diseases and ('ovine' OR 'sheep') AND ('flock' OR 'introduction' OR 'risk') AND ('disease name') for sheep diseases. The reference lists in the publications identified by the search were examined to identify further publications for possible inclusion in the review.

Publications which provided estimates of the contribution of individual risk factors to the likelihood of introduction of infectious diseases to cattle or sheep farms were included in this review. Differences in study design and populations were not considered in this review. The risk factors for each disease were extracted from each publication and presented in a table for further examination. Similarly defined risk factors from different publications were included as a single generic risk factor, in order to allow producing a generic template for the risk of disease introduction at the herd level. The variation in effect sizes (ie. odds ratios) across diseases was expressed by presenting their range.

Results

Searches of the PubMed, CABI and ISI Web of Knowledge (WOK) databases yielded 3,320 cattle-'hits', and 1,294 sheep-related 'hits'. The abstract or summary of each article was scrutinised and an initial selection of publications was performed. A total of 29 publications for the 10 selected cattle and 8 for the selected sheep diseases met the review selection criteria as summarised in Table 1 and Table 2.

Table 1: Number of publications identified by the three literature search engines for each cattle disease, and the number of publications selected for the review

Disease/ Agent	ISI WoK	PubMed	CABI	Used
Bovine herpes virus 1	20	103	316	5
Bovine viral diarrhoea virus	74	218	400	1
Foot-and-mouth disease	43	45	170	2
Campylobacter fetus	25	52	128	0
Salmonellosis	60	91	185	5
Leptospirosis	10	108	173	2
Bovine tuberculosis	19	100	216	7
Johne's disease	30	132	253	4
Neospora caninum	38	81	123	3
Fasciola hepatica	13	21	73	0
Total	319	930	1964	29

Table 2: Number of publications identified by the three literature search engines for sheep disease, and the number of publications selected for the review

Disease/ Agent	ISI WoK	PubMed	CABI	Used
Scrapie	19	207	200	4
Maedi Visna	2	49	120	0
Jaagsiekte	2	8	24	0
Chlamydophyla abortion	1	30	86	0
Louping ill/ Tick borne fever	2	10	6	0
Footrot	0	26	78	1
Caseous lymphadenitis	1	18	45	1
Toxoplasmosis	1	26	60	1
Psoroptic mange	6	19	50	0
(Multi)resistant helminths	2	67	129	1
Total	36	460	798	8

The generic risk factor list relating to the risk of disease introduction to farms is presented in Table 3 for cattle and Table 4 for sheep. A visual assessment of the ranges of the disease-specific odds ratio estimates reveals that the mixing of different herds was associated with the highest risk of disease introduction. Both, larger herds and those that introduced at least one head of cattle per year are more at risk of disease introduction. Visits by animal health professionals and the use of protective clothing, boots, and disinfectant footbaths tend to have a protective effect.

Table 3: Summary table of odds ratio values by risk factor and disease for cattle farms (ranges are presented for multiple studies as well as their number)

	Disease										
Risk Factor	**BHV1**	**BVDV**	**FMD**	**Campylobacter**	**Salmonella**	***Leptospira hardjo***	**Bovine TB**	***Mycobacterium paratuberculosis***	***Neospora caninum***	***Fasciola hepatica***	**Min-Max**	**N**
Mixing of different herds *	1.3	5.1-28.6 (2)	1.6		0.1-12.6 (4)		13.4-48.8 (3)	2.1-2.5 (2)	0.1-8.8 (2)		0.1-48.8	15
More than 100 cow herd	5.4-9.0 (2)				1.5-19.2 (3)	1.9-7.2 (2)	4.2-15.8 (2)	1.7	2.1		1.5-19.2	11
Introduction >= 1 head of cattle	0.2-6.9 (5)	1.8-5.4 (2)	1.0-2.2 (2)		2.7-4.3 (2)		5.8	1.6-4.0 (2)			0.2-6.9	14
Contact with other ruminant spp.					0.56	6.7	2.4	3.5-16 (2)			0.6-16	5
Mixed herd					4.1		4.9				4.1-4.9	2
Surface water			1.6		0.5-2.1 (2)	7.9					0.5-7.9	4
Wildlife contact							1.4-3.4 (4)	3.4-4.1 (2)			1.4-4.1	6
Contiguous to infected herd							2.4-5.0 (4)				2.4-5.0	4
Visitors >= once a week	2.6-4.1 (2)										2.6-4.1	2
Over the fence contact		2.5									2.5	1
Feeding only grass					0.3-13.2 (4)						0.3-13.2	4
Farm > 2 miles away	0.2-1.2 (3)				4.3						0.2-4.3	4
Dog present									1.6-11.5 (3)		1.6-11.5	3
Shared equipment					0.6						0.6	1
Overall, boots and disinf. Dip	0.43				0.2-6.2 (2)						0.2-6.2	3
Professional visitors	1.7	0.24			0.52						0.2-1.7	3

*: Mixing of different herds: mixing more than 25% of cattle from the herd with other cattle, which do not originate from the original herd. In mixing there is an intense contact, i.e. using a shared pasture.

Table 4: Summary table of odds ratio values by risk factor and disease for sheep farms (ranges are presented for multiple studies as well as their number)

	Disease											
Risk Factor	**Scrapie**	**Maedi Visna**	**Jaagsiekte**	**Chalmydophyla**	**Louping ill /TBE**	**Footrot**	**Caseous lymphadenitis**	**Toxoplasma**	**Psoroptic mange**	**Resistant helminths**	**Min-Max**	**N**
Mixing of flocks*	4.2-5.3 (3)					7.9				4.0	4.0-7.9	5
Purchase of animals	2.8-8.5 (2)						5.8				2.8-8.5	3
Farm <100 m sea level	0.9					9.2		0.83			0.8-9.2	3
Use contract shearers							5.8				5.8	1
Flock size >100 sheep	4.1-10.2 (2)										4.1-10.2	2
Vermin control								2.3-4.1 (2)			2.3-4.1	2
Flock <25 years old										2.3	2.3	1
Pure breed vs. commercial	1.6										1.6	1
# of anthelmintic treatments										1.04	1.0	1
Dip sheep for ectoparasites							0.4				0.4	1
Feeding lambs concentrates	0.17										0.2	1
No sheep compost used	0.04										0.04	1

* Mixing of flocks: mixing more that 25% off sheep from the flock with other sheep, which do not originate from the original flock. In mixing there is an intense contact, i.e. using a shared pasture.

Discussion

A significant finding from the review was the small number of publications assesssing risk factors for introduction of infectious diseases to cattle and sheep farms. Although a relatively large number of candidate publications were initially identified, less than 1% met the pre-set review criteria. However, despite the use of a comprehensive search strategy it is very likely that a significant number of reports were not included, particularly unpublished reports. Another relevant finding was that there was a trend towards an increase in the number of relevant publications over the past 10 years, suggesting that more research effort is being directed towards defining risk factors for disease introduction.

It is acknowledged that in summarising the data for the risk factors reported in the selected publications by disease, there is a significant risk of generating biased odds ratios. Therefore, we recommend that the odds ratio calculated for each risk factor per disease is to be interpreted cautiously, but we still believe that the general direction of effects provides meaningful information.

In the cattle, and particularly, the sheep risk factor comparison, the limited availability of quantitative data is striking. One approach to addressing this lack of knowledge in the short-term is to use an expert opinion workshop (EOW) to identify the major risk factors for disease introduction and to estimate the contribution of each factor to the risk of introduction of specific infectious disease of cattle and sheep based on expert opinion.

Very few publications were identified that provided quantitative data on the impact of the implementation of biosecurity measures on risk of disease introduction. Instead, the majority of articles provided generic recommendations based primarily on knowledge of the mode of transmission and epidemiology of specific infectious disease agents. Therefore, in the absence of appropriately designed field studies which have quantified the efficacy of implementation of various preventive measures, the only recommendation that can be made is the implementation of measures which are effectively the reverse of the risk factors identified in Table 3 for cattle and Table 4 for sheep. This is the approach that Van Schaik et al (2001) adopted when modelling the economic impact of the implementation of preventive measures on the incidence of several cattle diseases (BHV1, BVDV, *L. hardjo* and salmonellosis).

Expert Opinion Workshop

Introduction

The results from the systematic review presented above indicate that the published literature provides only limited information about the factors that are related to the risk of disease introduction. The preventive measures that are described in the literature are presented in a generic fashion, so no quantitative judgement about their efficacy can be made. As part of the current project, an expert opinion workshop (EOW) was conducted, aimed at quantifying risk factors and the effect of different preventive measures. In addition to the risk factors and preventive measures, the likelihood of introduction of specific diseases and their impact when introduced were assessed. Also, the usefulness of information in currently available external databases that could be accessed to obtain information about the presence of risk factors on a farm was evaluated.

Materials and Methods

The experts were selected based on their veterinary expertise and employment background. A range of expertise was required covering all aspects of detection, control and prevention of infectious diseases of cattle and sheep. The veterinary experts came from geographic regions and professional activity backgrounds in the UK, and were complemented by two experts from abroad. The group of 19 experts consisted of nine experts from veterinary practice, six from government animal health services, and four university academics.

The EOW questionnaire consisted of five parts and the same ten diseases were addressed as in the systematic review. In Part I of the questionnaire, the experts were asked to estimate the likelihood of disease introduction and the impact of introduction to both, a disease-free and an endemic farm in the UK. The experts could express the likelihood and the impact using a semi-quantitative scale from 1 to 5. Score 1 represented very low, 2 low, 3 moderate, 4 high, and 5 very high likelihood or importance. Part II of the questionnaire dealt with the likelihood of disease introduction in a quantitative fashion. The experts were asked to assess the likelihood of disease introduction (10 diseases in beef/dairy cattle and sheep) on an annual basis. In Part III, the experts were asked to provide an opinion on the relative contribution of ten different risk factors (identified from the systematic review) to the risk of disease introduction to cattle and sheep farms using three different buying-in policies: a closed herd/ flock, a herd/ flock that buys-in less than 5% of the herd on an annual basis and a herd/flock that buys-in more than 5%. In this part, the experts were asked to allocate a total of 100% across the ten risk factors plus a non-specific 'factor' representing 'baseline risk'. The experts were told that each factor could be considered either as increasing (positive) or decreasing (negative) the risk of disease introduction. For the evaluation of the value of using data from external databases for estimating the risk of disease introduction, 16 different external data recording systems were examined in Part IV. In Part V, 19 preventive measures were evaluated with respect to their efficacy for reducing the risk of disease introduction. Again, the experts were asked to estimate this for farms operating the same three different buying-in policies, mentioned above. The experts were asked to identify those measures that would cumulatively reduce the risk of disease introduction by at least 50%. As in Part III, the experts were asked to attribute a relative weighting to the different measures in order to quantify their individual impact.

For Parts I to III of the EOW a semi-Delphi approach (Gallagher 2004; Horst 1998) was used. In this case, this implied that these questionnaire parts were sent out to the experts approximately 4 weeks in advance of the EOW for them to complete. The returned responses were summarised and presented on the day of the EOW. This approach was chosen for two reasons: Firstly, we wanted to use the summaries as a starting point for reaching a degree of consensus between the experts, and secondly we wanted to familiarise the experts with the data capture method we had chosen to use throughout the EOW.

The responses from each questionnaire were entered into a spreadsheet database and checked for completeness, and typing errors. The responses for beef and dairy herds in Parts I, II, III and V were compared statistically using non-parametric (Wilcoxon signed rank test) or paired T-tests, depending on the distributional characteristics of the data. To test whether there was a difference in the likelihood of introduction between diseases (Part II), a linear mixed effects (LME) model, including expert as a random effect, was used to analyse the beef-, dairy-cattle and sheep data. The same approach was applied to test whether the responses in relation to the likelihood of introduction based on Part I (semi-quantitative) and Part II (quantitative) were consistent.

Results

All experts completed the questionnaires required for this workshop. Most experts commented that it was quite demanding to evaluate this large number of risk factor/disease/farm management systems in a single questionnaire. There was a query regarding the definition of a beef farm, and it was agreed

that for the purpose of the workshop this was defined as being a beef suckler herd. Another issue was the definition of the factor 'baseline risk' in Part III. It was explained that this term was needed to complete the sum of 100%, so everything that was not explained by the specified risk factors was included in 'baseline risk'.

Part I

Overall, the ranking of likelihood of introduction of individual diseases to beef and dairy farms was very similar, with BVDV, BHV-1 and *L. hardjo* being estimated to have moderate to high likelihood of being introduced. For sheep farms, footrot, chlamydophyla abortion and psoroptic mange were considered by the experts to be the diseases with the highest likelihood of being introduced. Comparing beef and dairy farms, the likelihood of introduction of *M. avium* spp*. paratuberculosis (MAP)* ($p=0.020$), liver fluke $p=0.007$) and campylobacter ($p<0.001$) were different. There was a difference between beef and dairy in relation to the impact of salmonella ($p=0.025$), campylobacter ($p=0.005$) and *Fasciola hepatica* ($p=0.034$) in a naive herd, and, the impact of the introduction of IBR into an endemic herd ($p=0.083$).

Part II

Comparing between beef and dairy cattle farms, the estimates for the likelihood of introduction were lower in dairy farms for both liver fluke and campylobacter ($p<0.05$). The linear mixed effects model revealed that the likelihood of introduction was significantly different between the selected diseases. For cattle, the estimated annual likelihood of introduction into a farm of BVDV (0.33) and *L. hardjo* (0.27) was significantly higher than for the other diseases whereas FMD (0.01), bovine tuberculosis (0.09) and *Fasciola hepatica* (0.10) were less likely to be introduced ($p<0.0001$). For sheep farms, the introduction of scrapie (0.09), Maedi Visna (0.10), jaagsiekte (0.13) and louping ill (0.08) was considered to be less likely, whereas caseous lymphadenitis (0.15), toxoplasmosis (0.22) and psoroptic mange (0.24) were more likely to be introduced than other diseases ($p<0.0001$). There was a significant correlation between expert's responses to the semi-quantitative (Part I) and the quantitative (Part II) section of the questionnaire ($p<0.0001$).

Part III

The summarised findings indicate that there are several risk factors where the experts believed that they influence the risk of disease introduction. Ranking the percentage contribution estimates of each risk factor as part of the total risk of disease introduction was used to identify the most important risk factors. Introduction of a cow/sheep to a farm, mixing of herds, over-the-fence contact and usage of pasture that is contiguous to infected farms are the most important risk factors. For many diseases, the use of protective clothing on the farm is considered to be a risk-reducing factor. It is interesting to note that generally the experts felt that most of the risk of disease introduction to 'open' cattle farms could be explained by the listed risk factors, but for diseases such as campylobacter and *Fasciola hepatica* in closed herds a significant proportion of the risk could not be explained. The experts did not believe that there were significant differences between beef and dairy herds ($p=0.99$) in relation to the percentage contribution to total risk of all diseases and risk factors evaluated. However, there were statistically significant differences ($p<0.0001$) between the estimates of the percentage contribution of individual risk factors to the likelihood of introduction of diseases (see Table 5 and Table 6). When comparing the different risk factors with each other, 'over-the-fence contact' ($p=0.098$ for open herds and $p<0.0001$ for closed herds, and 'being contiguous to an infected herd' ($p<0.0001$) were considered to be significant risk factors for all 3 types of buying-in policies included in this analysis. For open herds, 'introduction of a calf/heifer/cow/bull' and 'mixing of herds' were factors that differed significantly from the other factors ($p<0.0001$ and $p=0.015$ respectively). For sheep farms, 'over-the-fence contact' had a significantly higher ($p<0.0001$) contribution towards total risk (across all buying-in policies), and in open herds, 'purchase' and 'mixing of flocks' were also significantly higher, compared to the other factors ($p<0.0001$). There was no significant relationship between percentage contribution to total risk and the different breeds, different buying-in policies and the different diseases.

Table 5: Proportion of total risk (across all diseases considered and for both production systems) due to specific risk factors for introduction of disease into cattle farms (mean and 95% confidence interval)

	% risk of introduction		
Risk factor	**Closed herd**	**Buying <5%/y**	**Buying >5%/y**
Baseline risk	23.3 (21.3-25.2)*	14.0 (12.7-15.4)*	13.7 (12.2-15.1)*
Introduction >1 calf/heifer/cow/bull	NA	26.3 (24.8-27.8)*	35.7 (33.9-37.6)*
Mixing of different herds	NA	15.5 (14.4-16.6)*	13.4 (12.4-14.4)*
Over-the-fence contact	23.5 (21.6-25.4)*	11.6 (10.8-12.4)*	9.7 (9.0-10.4)*
Contiguous to infected herd	24.3 (23.1-25.6)*	14.4 (13.6-15.3)*	12.1 (11.3-12.9)*
Non-professional visitors (1/week)	4.0 (3.5-4.4)	2.9 92.6-3.2)	2.5 (2.3-2.8)
Professional visitors (1/week)	5.0 (4.3-5.7)	2.7 (2.1-3.2)	1.5 (0.8-2.1)
Overall, boots and disinfectant dip	-6.2 (-6.9- -5.4)	-4.7 (-5.3- -4.2)	-4.3 (-4.7- -3.8)
Contact other ruminant species	10.0 (8.7-11.3)	6.5 (5.7-7.3)	5.7 (5.0-6.3)
Contact with non-ruminant species	10.8 (9.3-12.4)	7.4 (6.3-8.6)	6.4 (5.4-7.4)
Shared equipment	5.7 (5.2-6.3)	4.0 (3.6-4.4)	3.5 (3.2-3.8)

* estimate is significantly higher, compared to other risk factors (p<0.05)

Table 6: Proportion of total risk (across all diseases considered) due to specific risk factors for introduction of disease into sheep farms (mean and 95% confidence interval)

	% risk of introduction		
Risk factor	**Closed herd**	**Buying <5%/y**	**Buying >5%/y**
Baseline risk	52.8 (48.4-57.1)*	25.9 (22.6-29.1)*	20.8 (17.4-24.1)*
Pure breed	3.7 (2.6-4.8)	3.4 (2.4-4.4)	3.5 (2.4-4.5)
Farm <100 m above sea level	3.1 (1.9-4.3)	1.7 (0.7-2.7)	1.5 (0.6-2.3)
Purchase of animal(s)	NA	32.3 (30.1-34.5)*	42.5 (40.0-45.2)*
Mixing of flocks	NA	23.0 (21.3-24.8)*	20.9 (19.2-22.5)*
Over-the-fence contact	25.4 (21.6-29.2)*	8.7 (7.5-10.0)*	7.3 (6.3-8.2)*
Non-professional visitors	2.8 (1.6-4.0)	1.4 (1.0-1.9)	1.3 (0.9-1.8)
Professional visitors/shearers	8.7 (6.2-11.3)	4.0 (2.9-5.1)	3.2 (2.2-4.2)
Contact other ruminant species	7.7 (5.9-9.5)	4.3 (3.2-5.4)	3.6 (2.7-4.4)
Sheep dipping applied	-3.8 (-5.6- -2.1)	-4.0 (-5.6- -2.38)	-3.7 (-5.4- -1.9)
Vermin control (cats/poison)	0.6 (-1.2-2.4)	0.3 (-1.5-2.0)	0.3 (-1.4-1.9)

* estimate is significantly higher, compared to other risk factors (p<0.05)

Part IV

The experts differed in their opinion about the accuracy of the estimates of the contribution that external databases could make towards describing selected specific risk factors of disease introduction. None of the listed databases were considered capable of providing extremely reliable data on the contribution of specific risk factors to the introduction of disease to sheep farms. For cattle farms, it was felt that the risk factors 'introduction of a cow' and 'being contiguous to an infected herd' can be defined adequately using external databases.

Part V

As in Part III, there is no difference between beef and dairy farms in relation to the importance of preventive measures. In only one scenario (protective clothing in case of preventing salmonella introduction), 'buying-in less than 5%' differed significantly from those 'purchasing more than 5% of cattle per year'. The results of the LME model indicate that there was no significant difference between specific diseases with respect to the estimated percentage impact of the specified preventive measures. There was a difference in the magnitude of the percentage effect between preventive measures ($p<0.0001$). Evaluating the importance of preventive measures (as a percentage), several factors were reported by the experts. 'Maintaining a closed herd' (56%), 'double fencing' (8%), 'cattle-proof perimeter fence' (8%), and 'no grazing on perimeter pastures' (10%) were more important factors for reducing the risk of disease introduction into closed herds than the remaining measures ($p<0.05$). For both categories of open herds, 'becoming a closed herd' (44%) and 'purchase of certified disease-free animals' (15%) were considered to be more important risk mitigation measures for disease introduction than the other measures included in this analysis ($p<0.05$).

Discussion

The EOW was an intensive experience, both for the experts and the researchers. The Delphi approach was essential for enabling the experts to complete the questionnaire survey on the day of the workshop. Considering the current findings, we can conclude that direct contact with cattle and sheep from outside the farm is the most important risk factor for the introduction of disease to cattle and sheep farms, respectively. The magnitude of the estimated contribution of this risk factor varies between the

different diseases evaluated, reflecting differences in the modes of transmission and biology of the causative infectious agents. Overall, similar rankings of the different risk factors were generated by the EOW and the systematic review. For prevention of disease introduction, becoming and remaining a closed herd are the most important preventive measures. When animals are to be introduced, it is important to consider their origin and the source herd's disease status, to treat, vaccinate and quarantine them. Prevention of indirect contact becomes more important for closed herds.

Proposal for Biosecurity Risk Assessment Algorithm

The information generated by the EOW and the systematic review was used to develop a farm-level risk assessment biosecurity algorithm which is described here.
For each disease, a relative weighting was calculated based on the products of the likelihood of introduction and the economic impact. These weightings were multiplied by the average score (0-100) assigned to each risk factor for each disease and a weighted mean for each risk factor was calculated. It was not possible to directly compare the weighted mean risk factor impact scores between the three farming styles, because the closed herds did not have data for the risk factors 'introduction of livestock' and 'mixing of herds/flocks'. Therefore, in order to allow for such a comparison between closed herds and herds buying-in over 5% of their livestock annually, we adjusted the weighted mean risk factor impact scores by removing these two risk factors, which contributed 51% in dairy, 52% in beef and 61% in sheep farms to the total score. Then, we adjusted the remaining common risk factors to equal 49 (dairy), 48 (beef) and 39 (sheep). Similarly, to be able to compare farms buying-in <5% with farms buying-in >5% of their livestock we rescaled the weighted mean risk factor impact scores using the difference in impact of buying-in cattle. Based on the finding that there was no difference between the risk factor impact scores for beef and dairy cattle, the scores for both were averaged to represent overall scores for all cattle herds.
The resulting biosecurity risk assessment algorithms are presented in Table 7 for cattle and in Table 8 for sheep. These can be used to calculate a biosecurity risk score for individual farms, by considering the presence/absence of the various factors and summing up the applicable scores. The resulting risk score could assist in the decision process on the need for a biosecurity risk management plan (if the score is above a set value), and then if one is required the individual components of the total risk score could be help prioritising different risk management options involving modification of different sets of these key risk factors.

Table 7: Biosecurity risk assessment algorithm for disease introduction risk scoring of cattle herds

Risk factor for disease introduction to a cattle herd	Score
Baseline risk	11
Contact with non-ruminant species	5
Contact other ruminant species	5
Contiguous to infected herd	12
Non-professional visitors (1/week)	2
Over the fence contact	11
Overalls, boots and disinfectant dip	-4
Professional visitors (1/week)	2
Shared equipment	4
Mixing of different herds	15
Introduction of a calf/heifer/cow/bull	24
Introduction of more than 5% of herd on annual basis	13

Table 8: Biosecurity risk assessment algorithm for disease introduction risk scoring of sheep flocks

Risk factor for disease introduction to a sheep flock	Score
Baseline risk	20
Pure breed	2
Farm <100 m above sea level	2
Over the fence contact	8
Non-professional visitors	1
Professional visitors/shearers	4
Contact other ruminant species	4
Sheep dipping applied	-3
Vermin control (cats/poison)	0
Mixing of flocks	21
Purchase of animal(s)	28
Introduction of more than 5% of flock on annual basis	13

Conclusion

The systematic review combined with the EOW was very useful for identifying generic key risk factors for farm biosecurity. It also showed that a lot of information is based on anecdotal evidence, which emphasizes the necessity for using an EOW in addition to review of published data. It was also necessary to use 'important' diseases as the basis of the risk factor identification, since generic biosecurity has not been studied extensively. It was surprising though how much consistency there was in terms of the types of risk factors and their relative importance across diseases. The most important risk factors for introduction of hazards were related to introduction of livestock, as would be expected. It needs to be emphasized that while the identified factors had been known, as far as we know it had not been attempted before to produce a relative ranking across diseases and combine the information into a semi-quantitative score. The outcome of this component of the project has been a farm-level biosecurity risk assessment algorithm that will allow producing a semi-quantitative biosecurity risk score for cattle and sheep farms. The result will inform the development of tailored risk management strategies. The algorithms need to be tested on farms and it is likely that they will have to be revised in the light of field experience. The sensitivity/specificity of these algorithms which require farm visits can be improved by combining it with information generated under the second objective of the current project, relating to biosecurity risk assessment based on externally (not requiring farm visit) measurable factors.

Objective 2: Identify risk factors associated with disease introduction to cattle herds and sheep flocks in GB based on retrospective data analysis

Introduction

The second research objective of the project is based on using a data-driven approach for producing a risk scoring algorithm for farm-level biosecurity using externally-measurable factors (EMF). For the purposes of this study, these are defined as all attributes potentially associated with disease presence that can be accessed for the unit of interest (animal, farm, etc.) without requiring additional specific data collection activities, since the information has already been collected for some other purpose. EMFs at individual animal level are more difficult to collect than herd-level data, and can be retrieved only in animal populations with accurate identification systems. The EMFs used here can be grouped into four categories: Environmental, animal movements, demographic/farming-system patterns and densities.

In the first year of the project, a pilot study was conducted as a preliminary step in order to assess the availability and accessibility of data sources and the suitability of the analytical approach where statistical methods, spatial analysis and data mining techniques were applied to predict disease presence in a sub-population of the cattle holdings in Norfolk County in 2002.

A second study has been conducted in the final phase of the project where the presence of bovine tuberculosis and of a group of other common diseases was analysed for cattle and sheep holdings in Wales in 2004. The results of data analyses were then used to develop a semi-quantitative biosecurity risk-score system. The results from this work are presented here.

Many of the assumptions and theoretical approaches used in this study are similar to the ones described in the report for the pilot study submitted in May 2005. For further details on spatial

resolution and data aggregation see milestone report 1 (S2a-3) of this project. Equally, a detailed description and discussion of the characteristics of the databases used for the analysis, the suitability of the disease data, its biases and completeness and a detailed explanation of the analytical methods can be obtained from that particular report.

Materials and Methods

Data sources

The datasets used in this study have been collated using the following sources and methods:

- Higher Education Institutions have restricted free access to selected databases for research and teaching purposes. Previous registration of the institution is required (UKBORDERS-EDINA) and acknowledgment of the conditions of use in other cases (UK Postcode Directory)
- Purchase: Agcensus (EDINA)
- Free access to the public through the web: Protected sites' datasets
- Official requests: Animal Movement Licensing System (AMLS), Cattle Tracing System (CTS), SOILSCAPE, LANDMAP, climate data, geo-references of agricultural holdings, TB database etc.
- Non-official requests: Farmfile, CTS data, VETNET, SND

Outcome data

- Farmfile-Veterinary Investigation Diagnosis Analysis (VIDA) Veterinary Laboratories Agency (VLA)
- Scrapie Notification Database (SND) and Tb database. Veterinary Laboratories Agency (VLA)
- TB database: Centre for Epidemiology and Risk Analysis (CERA). Veterinary laboratories Agency (VLA)

Externally measurable factors

- UKBORDERS-EDINA
- Geo-referenced database of agricultural holdings in Great Britain (DEFRA)
- UK Postcode Directory-National Statistics (©Crown Copyright 2004. Source: National Statistics / Ordnance Survey)
- Agricultural Census 2004 (Assembly of Wales)
- VETNET: Animal Health Information System of the State Veterinary Service (SVS)
- Cattle Tracing System (CTS). British Cattle Movement System (BCMS). Veterinary Laboratories Agency
    - Two main datasets were obtained containing movement data:
    - All cattle movements on and off holdings in Wales during 2004: This data was used to calculate the number of days and number of animals moved on and off a Welsh cattle holding.
    - All cattle movements on/off holdings in Wales during 2004 to/from holdings in Wales.
- Animal Movement Licensing System (AMLS). British Cattle Movement System (BCMS)- Rural Payment Agency (RPA)
- Protected sites datasets (© Crown Copyright. www.ccw.gov.uk)
- Digital Boundary Data for Designated Wildlife Sites and related information. The following datasets containing specific designated areas of environmental importance were selected for the study:
- National Nature Reserve (NNR)
- Local Nature Reserve (LNR)
- Areas of Outstanding Natural Beauty (AONB)
- UK Census 2001-National Statistics. 2001 Census Aggregate Outputs. Economic & Social research Council (ESRC). Accessed via Manchester Information & Associated Services (MIMA).
- Agcensus 2004- EDINA
- Land Cover Map 2000: Centre for Ecology & Hydrology
    - 1 km grid raster data with aggregate classes and sub-classes
- 5Km monthly mean rainfall and temperature data (199-2004). Geographic Information Unit- DEFRA-authorised by the Met Office
- Land-Form PROFILE Contours & Digital Terrain Model for Wales, with a 10Km buffer into England. Geographic Information Unit- DEFRA- authorised by Ordnance Survey
- NATMAP soilscapes for Wales. National Soil Resources Institute (NSRI) - Cranfield University. Silsoe. Bedford

Cattle Study Population

According to the aggregate Agricultural Census 2004 data published by the Assembly of Wales, the total number of cattle holdings in Wales was 13,966. The unique identifier of the study units is the County-Parish-Holding number (CPH). From the initial list of 23,319 CPHs in Wales, the identification of the cattle holdings of the study population was based on a combination of three data sources:

- CPHs that submitted bovine samples or specimens to any VLA-Regional Laboratory during 2004
- CPHs that were tested for TB in Wales during 2004
- CPHs that registered cattle movements into CTS during 2004

In this study, it was assumed that holdings in Wales are those with county numbers in their CPHs between 52 and 60. Following a cross-check with the geo-references and postcodes, some of these

CPHs appear to be located in neighbouring English counties but generally close to the Welsh border, and these were therefore included in the final study population. All non-farming holdings that could be identified through holding type in various data sources were removed from the final study population. This was the case for markets, abattoirs, show grounds, artificial insemination centres, collection centres, etc. Thus it is expected that the number of included holdings which do not correspond to the typical farm structure has been reduced to a minimum. The final study population contained 15,845 CPHs which represents 110% of the holdings reported by the 2004 Census.
The unit of interest was the holding defined by the CPH (country-parish-holding). The only identifier of the cattle and sheep holdings available was the CPH number for which a geo-reference was obtained. The location of each holding was determined using the Easting and Northing coordinates obtained from the Agricultural Census database. The point locations of all cattle holdings in the study population were mapped using ArcGIS 9 (© ESRI).
A dataset was generated from these various data sources containing information about 15,845 cattle holdings and more than fifty variables for each of them. A subset of 33 variables was included in the final dataset to be used for analysis. Raw variables were available at different numerical scales, continuous to binary (multi-species, presence of sheep, any disease, closed/open) and categorical (herd size, etc).
The following two outcome variables were analysed:

- Outcome 1: Bovine tuberculosis. A holding was declared positive if there was a confirmed breakdown of TB in the holding during 2004.
- Outcome 2: At least one of 10 selected diseases (see Table 9) was recorded in Farmfile as having been diagnosed on the holding in 2004 and/or a breakdown of bovine TB was confirmed on the holding during 2004.

Table 9. Disease included in the Outcome 2 (Any disease in cattle)

BVD: BVD, persistently infected, congenital disease due to BVD	Salmonellosis (*S.typhimurium*)
Mucosal disease	Rotavirus infection
IBR: IBR/IPV, foetopathy due to IBR/IPV	Mastitis (10 causative agents)
Neospora: foetopathy due to Neospora	Pneumonia (10 causative agents)
Mycobacterium paratuberculosis: Demonstration of AFB or positive to ELISA	*Fasciola hepatica*

The number of holdings that could not be geo-referenced due to the absence of geographical coordinates in the database was 828 (5.2%). Attributes requiring geo-referencing for linking them to the holding were then not available in the final dataset for these holdings. There were 637 (4%) positive observations for Outcome 1 in the dataset and 1,372 (7.5%) for Outcome 2.The name and description of each variable used in the analysis are described below.

Demographics:

- **TVETNET**: Holding type using VETNET data source. Reclassified in three categories: beef, dairy, other.
- **CATSHE**: Mixed holding (cattle and sheep) or not.
- **TCTS**: Holding type based on CTS data source. Only two types of holdings appear in the final study population: agricultural holdings with land and landless keeper. Other holding types as in CTS database do not appear because markets, abattoirs, Artificial Insemination Centres, show grounds, etc. have not been included in the study population.
- **FARMA**: Total area farmed based on the Agricultural Census 2004.
- **HERSIZE:** Herd size was estimated by combining the information available in two different data sources: VETNET and TB database.

Densities:

- **CATDENCEN:** Number of cattle in the 5 km^2 grid cell area where the holding is located. This value is taken directly for each 5 km grid cell from the Agcensus 2004 database.
- **CATDENSITY:** Smoothed number of cattle in the 5 km^2 grid cell area within which a holding is located. This is calculated by interpolating the surface of the study area for each 5 km^2 grid cell from point location data of the cattle population as provided by the Agcensus 2004 database. The output is a "smoothed" value of the no. of cattle per 5km^2 grid cell calculated by using ordinary kriging based on a spherical semivariogram and a variable search radius of 12 points, using Spatial Analyst in ArcGIS 9 (©ESRI).
- **SHEDENCEN:** Number of sheep in the 5 km^2 grid cell area where the holding is located. This value is taken directly for each 5 km grid cell from the Agcensus 2004 database.
- **CATBUF5:** Number of cattle holdings within a 5 km radius buffer area surrounding a holding's point location.
- **ADDPOST:** Number of addresses in the post code area where the holding is located, using UK Postcode Directory-National Statistics.
- **POPCAS:** Number of total human population in the Census Area Statistics (CAS) where the holding is located.

Animal movements:

Two datasets were produced by retrieving cattle movement data from CTS in order to characterize cattle movements and assign attributes to the study units.

Dataset 1

It contained almost 4 million of paired movements that had as either origin or destination CPHs in Wales during 2004. So, movements on- and off-holdings not located in Wales were allowed as long as one of the CPHs of the paired movement was in Wales. All births and deaths were deleted from the dataset so that only movements of live animals were considered. A total of 13,515 CPHs in Wales moved a total of 583,845 cattle off their premises during 2004. During the same period 269,180 cattle were moved on to holdings in Wales. This dataset resulted in the following variables for analysis:

- **DAYMOVON:** Total number of days on which the holding moved animals on to the premises during 2004.
- **CATMOVON:** Total number of animals moved on to the holding during 2004.
- **DAYMOVOFF:** Total number of days on which the holding moved animals off the premises during 2004.
- **CATMOVOFF:** Total number of animals moved off the holding during 2004.
- **OPENCLOSEDTOTAL:** Did the holding have cattle movements on to and off the farm registered during 2004?
- **OPENCLOSEDON:** Did the holding have cattle movements on to the premises registered during 2004?
- **OPENCLOSEDOFF:** Did the holding have cattle movements off the premises registered during 2004?

Dataset 2

This dataset contained all cattle movements that had as origin and destination agricultural holdings in Wales during 2004. It reflects the internal movement structure and contacts between cattle holdings in Wales. Movements to abattoirs were removed and only markets were left as non-farming holding in the dataset. With these paired movements of cattle within Wales a directed unvalued (multiple movements between same holdings are not counted) network has been built. The network contained 13,250 nodes (CPHs) and 47,085 links. The main centrality measures of the nodes of the network were calculated and included as attributes in the dataset for analysis:

- **INDEGWAL:** The indegree of a holding is the number of other holdings (CPHs) from which animals were moved to the holding in 2004.
- **OUTDEGWAL:** The outdegree of a holding is the number of other holdings (CPHs) to which animals were moved in 2004.
- **BETWAL:** Relative betweenness of each holding is defined as the proportion of the maximum betweenness a holding can have in the network in 2004. Betweenness quantifies the number of times a holding has been involved in movements connecting any pair of holdings in the holding network (Wasserman and Faust, 1994).
- **OUTCLOSWAL:** Output closeness of each holding in the network in 2004. Closeness reflects how close a holding is to other holdings in the network (Wasserman and Faust, 1994). This is not geographical distance, but rather whether animals move directly from holding A to holding B or whether there have been intermediate movements to other holdings before an animal has moved from holding A to B. Output closeness of a holding only takes into account the movement links from the holding of concern.
- **INCLOSWAL:** Input closeness of each holding in the network in 2004. Input closeness of a holding takes into account only the incoming movements from other holdings in the network to the holding of concern.

Environmental factors

- **SOILECAT:** 24 soil classes according to NATMAP soilscapes have been aggregated into four main soil types: Combinations of loamy soils, combination of acid soils, combination of freely draining soils and other types (peat soils, salt marsh, sand dune, etc.).
- **TEXTCAT**: Three classes of soil texture according to NATMAP soilscapes: loamy, peaty and sandy.
- **DRAINCAT:** Six classes of soil drainage according to NATMAP soilscapes and re-classified into four types: freely draining, impeded and slightly impeded drainage, surface wetness/naturally wet and variable.
- **FERTCAT:** Ten classes of soil fertility according to NATMAP soilscapes re-classified into four types: high, moderate, low and lime-rich.
- **LANDNATMAP:** Sixteen classes according to LANDNATMAP re-classified into three types: combinations of arable land, combinations of grassland and other (moorland, forestry, etc.)
- **HABICAT:** Twenty classes of habitats according to NATMAP soilscapes and re-classified into four types: combinations of pasture and woodlands, combination of grassland and grass moors, combination of wet areas and other (coastal salt marsh, sand dune vegetation, etc.).
- **TOTRAIN:** Total rainfall in 2003**.**
- **AVETEMP:** Annual average temperature in 2003.
- **5KMAONB:** Within or at less than 5km to an Area of Outstanding Natural Beauty.
- **5KMNNR:** Within or at less than 5km to a National Nature Reserve (NNR).
- **5KMLNR:** Within or at less than 5km to a Local Nature Reserve (LNR).
- **WITHISSSI:** Within a “Site of Special Scientific interest” (SSSI).

Sheep Study Population

According to the aggregate Agricultural Census 2004 data published by the Assembly of Wales, the total number of sheep holdings in Wales was 15,483. From the initial list of 23,319 CPHs in Wales, the identification of the sheep holdings in the study population was performed using a combination of three sources:

- CPHs that submitted ovine samples or specimens to any VLA-Regional Laboratory during 2004 (2,260)
- CPHs that appear to have sheep in the Agricultural Census 2004 (15,483)
- CPHs that registered sheep movements in the Animal Movement Licensing System (AMLS) in 2004 (12,420).

The final study population contained 18,937 CPHs which represented 122% of the holdings reported by the Census in 2004. A dataset was generated from these various data sources containing information about 18,937 sheep holdings and more than fifty variables. A subset of 24 variables was included in the final dataset to be used for analysis. The outcome variable "Any disease in sheep" was defined on the basis of at least one of 9 selected diseases (see Table 10) being recorded in Farmfile as having been diagnosed on the holding during 2004

Table 10: Diseases included in the Outcome 3 (Any disease in sheep)

Caseous lymphadenitis	Maedi-Visna
Toxoplasmosis	Pulmonary carcinomatosis (Jaagsiekte)
Fusobacterium necrophorum (Footrot)	Sheep scab
Louping ill	Parasites (Haemonchosis, Nematodiriasis)
Tickborne fever	

The number of holdings that could not be geo-referenced due to the absence of geographical coordinates in the database as 6,170 (32.6%). There were 425 (2.2%) positive observations in the dataset. The name and description of each variable used in the analysis is described below.

Demographics

- **MIXED**: Mixed holding (cattle and sheep) or not.
- **HOLTYPE**: Holding type based on AMLS holding type classification. Only two types of holdings appear in the final study population: agricultural holdings, domestic premises and unknown
- **FARMA**: Total area farmed based on the Agricultural Census 2004.
- **TOTCAT**: Herd size was estimated by combining the information available in three different data sources: Agricultural Census 2004, VETNET and TB database. Only calculated for the mixed holdings (cattle and sheep).
- **TOTSHE:** Flock size was extracted from the Agricultural Census 2004.

Densities

- **SHEDENCEN:** Number of sheep in the 5 km^2 grid cell area where the holding is located. This value is taken directly for each 5 km grid cell from the Agcensus 2004 database.
- **SHEBUF5:** Number of sheep holdings within a 5 km radius buffer area surrounding a holding's point location.

Animal movements

All sheep movements on and off holdings in Wales in 2004 were extracted from the Animal Movements Licensing System (AMLS). It contained 124,016 pairs of movements that had as either origin or destination CPHs in Wales during 2004. The following movement variables were extracted from this dataset for analysis:

- **DAYMOVON:** Total number of days on which the holding moved sheep on to the premises during 2004.
- **SHEMOVON:** Total number of sheep moved on to the holding during 2004.
- **DAYMOVOFF:** Total number of days on which the holding moved sheep off the premises during 2004.
- **SHEMOVOFF:** Total number of sheep moved off the holding during 2004.
- **INDEGREE:** Number of holdings from which animals were moved to the holding during 2004.
- **OUTDEGREE:** Number of other holdings to which animals were moved during 2004.

Environmental factors (same description as in the cattle dataset)

- **SOILECAT, TEXTCAT, DRAINCAT, FERTCAT, LANDCOVER, TOTRAIN, AVETEMP, 5KMAONB, 5KMNNR, 5KMLNR** and **WITHISSSI**.

Analytical Methods

The data were analysed using two different multivariable analysis methods: Logistic regression and classification tree analysis. Logistic regression produces an equation linking different risk factors for calculating risk estimates given possible patterns of risk factor values. The effect of the individual variables is expressed using odds ratios. Classification tree analysis divides the data into subgroups on basis of risk factor categories aiming to maximise the proportion of observations within a single outcome category within each subgroup. The algorithm tests all possible cut-off points for numeric

variables and all classes in categorical variables until best discrimination between outcome categories has been achieved. Subgroups can be further subdivided using the same or other risk factors until either only one outcome category is represented or the minimum number of observations within the subgroup has been achieved. The result of this process is a tree that can then be used to classify observations in groupings associated with a certain probability of being in one of the outcome categories. The hierarchical structure of the classification trees also indicates the relative importance of the variables for influencing the outcome of interest and the interaction among the variables (Zhang et al. 2006).

Factors associated with the occurrence of the outcomes of interest at holding level during 2004 were identified by fitting a random effect multivariate logistic regression model using STATA 9.0 (Stata® Corporation 2005). Parish was included as a random effect as a relatively crude method of taking account of spatial dependence. Variables with less than 250 observations were not included in the final model and the criterion for inclusion was forward selection based on the likelihood ratio test using a cut-off p-value of 0.1 for entry of variables into the model. Environmental attributes were extracted using ArcGIS 9 (© ESRI) from the corresponding GIS layers, and attributed to the corresponding farm locations using ArcView Spatial Analyst (© ESRI). Numeric variables were tested using both continuous and categorical scales. Aggregation of the numeric variables in three categories was done using the 33^{th} and 66^{th} percentiles for some variables or applying the cut-off points derived from the classification trees for others. Multi-category environmental variables were aggregated on the basis of biological considerations in relation to the most important features of the variable. Tests for linearity of effects were conducted using categorical variables as well as tests for interaction. A comparison of nested fitted models was conducted using likelihood ratio tests and the Akaike Information Criterion (AIC).

Several classification trees were generated using the software WEKA 3.4.4 (©1999-2005 University of Waikato, New Zealand). A metaclassifier for cost-sensitive learning was initially served to a set of 33 and 24 variables for each dataset in an attempt to predict the classes of interest with the least expected misclassification cost rather than the most likely one. The C4.5 algorithm was used to build classification trees within the metaclassifier, as described by Quinlan (1993).

We developed three classification trees with increasing cost-ratios for false negatives relative to false positives (5:1, 10:1 and 20:1) according to the following specifications: 10-fold cross-validation, reweighing of the training data according to the different costs assigned to each class and allowing only leaves with 200 or more instances in each. The reduced-error pruning procedure uses three folds: One fold for pruning and the remainder for building the tree following the steps of the C 4.5 algorithm.

The third method applied is the extraction of different sets of classification rules from partial decision trees. A classification rule is a prediction rule of the form: IF <conditions> THEN <prediction (class)> (Carvalho et al. 2002). The PART algorithm was used as implemented in WEKA 3.4.4 (©1999-2005 University of Waikato. New Zealand). Since our objective was to identify holdings at increased risk of disease, we specified the classification algorithm such, that it maximized the sensitivity through increasing the relative weight of false negatives. This means we applied the same cost-sensitive metaclassifier with the same three penalties for misclassification of false negatives: 5:1, 10:1 and 20:1. The minimum number of observations in each rule was set to 50 whereas only terminal nodes with 200 or more observations were allowed in the classification trees. This less stringent criterion results in the inclusion of more variables, more combinations of variables and more variety of cut-off points for the same variables in the rules, which make them more difficult to interpret. The analysis was conducted for each of the two outcome variables. Descriptive analyses of the classification rules and the quantitative assessment of the different types of variables in the rules that predict positive outcomes were performed.

As part of the development of the biosecurity risk algorithm, the correlation and the dimensionality of the movement variables was analysed using factor analysis in STATA 9.0 (Stata® Corporation 2005).

Results

Outcome 1: Bovine TB in cattle holdings

Random effect logistic regression model:

The final random effects logistic regression model was based on data from 6,871 holdings out of 15,845 and the variable parish was included as a random effect (n=880). High cattle density, herd size, number of days in year cattle were moved off the holding and farmed area are associated with a significant increase in the odds of the outcome of interest occurring. Total rainfall, average temperature, location within or at less than 5km of an AONB, 'indegree' in the Wales network,

'outdegree' in the Wales network, output closeness Wales network and land cover (grassland) reduce the odds for the presence of TB in the study population (Table 11). Land cover was forced into the final model because of its confounding effect on other significant environmental variables like TOTRAIN. An alternative model was built replacing the movement variables by a composite of them: open/closed. The final model using these variables includes 7,412 observations with parish as a random effect (n=889). OPENCLOSEOFF is a significant movement variable with a five-fold increase in the odds of having TB if animals are moved off the premises. Although the main effects for categorised average temperature and total rainfall are associated with a reduced odds of TB in the final model, a significant interaction (OR= 1.98; P=0.025; 95% CI: 1.08-3.61) between them produces a two-fold reduction of the protective effect for average temperature >=9 and total rainfall >=900 ml.

Table 11. Variables included in final random effect logistic regression models for Outcome 1 (TB in cattle)

	Movement variables			Open/closed		
Variable names	**OR**	**P value**	**95% CI**	**OR**	**P value**	**95% CI**
CATDENSITY (binary >2,630)	4.14	P<0.001	2.55-6.72	4.15	P<0.001	2.62-6.58
HERDSIZE (categorical)						
<30	ref			ref		
30-88	3.81	P<0.001	1.98-7.34	3.73	P<0.001	2.19-6.33
>88	4.4	P<0.001	2.2-8.8	5.24	P<0.001	3-9
DAYMOVOFF (categorical)						
<4	ref					
4-11	2.79	P=0.001	1.55-5.02			
>11	5.25	P<0.001	2.84-9.68			
FARMA (categorical)	1.27	P=0.026	1.03-1.57	1.37	P=0.002	1.12-1.68
5KMAONB	0.37	P=0.001	0.2-0.67	0.42	P=0.003	0.24-0.73
TOTRAIN (binary >=900)	0.43	P=0.001	0.26-0.7	0.38	P<0.001	0.24-0.62
AVETEMP (binary >=9)	0.55	P=0.004	0.36-0.82	0.57	P=0.005	0.39-0.84
OUTDEGWAL (categorical)	0.7	P=0.001	0.57-0.86			
OUTCLOSWAL (categorical)	0.77	P=0.021	0.62-0.96			
LANDNATMAP (grassland)	0.78	P=0.095	0.58-1.04	0.75	P=0.053	0.57-1
INDEGWAL (categorical)	0.81	P=0.008	0.7-0.94			
OPENCLOSEDOFF				5.6	P=0.044	1.05-56.26

Classification trees:

All variables included in the three trees are shown in Table 12 in hierarchical order so that variables at the top of the table have more ability to discriminate observations classified as positive and appear at the top of the classification trees. Classification performances of the trees are shown in Table 13.

Table 12. Hierarchical ranking of variables (top=highest; bottom=lowest) in the three classification trees for Outcome 1 (TB in cattle)

Tree 5:1	**Tree 10:1**	**Tree 20:1**
CATDENSITY	CATDENSITY	DAYMOVOFF
DAYMOVOFF	DAYMOVOFF	CATDENSITY
HERSIZE	TOTRAIN	HERSIZE
TOTRAIN	HERSIZE	CATMOVOFF
INDEGWAL	CATDENCEN	SHEDENCEN
	SHEDENCEN	ADDPOST
	OUTCLOSWAL	TOTRAIN
	CATSHE	OUTCLOSWAL
	LANDNATMAP	SOILCAT
	ADDPOST	CATSHE
	OUTDEGWAL	DAYMOVON
	FERTCAT	OUTDEGWAL

Table 13. Classification performance of the three trees using different values for ratio of positive versus negative misclassification cost for Outcome 1 (TB in cattle)

	Tree 5:1	**Tree 10:1**	**Tree 20:1**
Sensitivity	18.5%	36.9%	64.2%
PPV	25.7%	14.1%	9.6%
Specificity	97.8%	90.6%	74.8%
NPV	96.6%	97.2%	98%
Error rate	5.4%	11.5%	25.6%
Area under ROC	59.9%	74.1%	75.7%

Three of the top five variables in the three trees are common: CATDENSITY, DAYMOVOFF and HERSIZE. TOTRAIN appears in two trees (5:1 and 10:1) and four other variables appear in one of the trees: INDEGWAL, CATDENCEN, CATMOVOFF and SHEDENCEN.

Trees 5:1 and 10:1 have very low sensitivity, whereas Tree 20:1 achieves moderate levels of correct classification (64.2%) for positive observations. However, this rise from 36.9% to 64.2% produces an increase in the number of false positives from less than 10% to 26%, an error rate of 25.6% and a very low positive predictive value of less than 10%. The increase in the misclassification penalty for false negatives results in a greater importance of movement and densities variables and the environmental

variables become less important, so that TOTRAIN, the third most important variable in the Tree 10:1, drops to seventh in the Tree 20:1.

Classification rules:

The sets of classification rules for the three misclassification penalties contain 36, 52 and 57 rules, respectively. Of these 10, 14 and 23 predict the outcome of interest (TB). For example CATDENSITY appears in ten of the 23 positive classification rules 20:1. AVETEM appears in five of the fourteen positive classification rules 10:1. Density variables are included most frequently in the three sets of rules, with an increasing weight as sensitivity is forced up through the higher misclassification ratio (see Table 13). CATDENSITY, CATDENCEN AND SHEDENCEN are consistently present in most of the rules. DAYMOVOFF is the most important movement variable with a generally consistent pattern and weight throughout the sets. Demographic variables are least commonly represented in the rules (Table 14). However HERSIZE is in the top five of the list with a cut-off point increasing with increasing misclassification cost.

Table 14. Relative contribution of different variable groups to the positive classification rules for classification trees with different misclassification cost ratios for Outcome 1 (TB in cattle)

	Misclassification cost ratio		
Variable group	**Tree 5:1**	**Tree 10:1**	**Tree 20:1**
Demographics	12.1%	18.1%	8.6%
Densities	39.4%	31.3%	41.4%
Movements	30.3%	27.7%	24.1%
Environment	18.2%	22.9%	25.9%
	100%	100%	100%

Outcome 2: 'Any disease' or bovine TB in cattle holdings

Random effect logistic regression model:

The final random effects logistic regression model was based on data from 8,656 holdings out of 15,845, and parish was included as a random effect (n=935). High cattle density, location in peat soils and marshes, increasing herd size, number of days cattle were moved off the holding and 'outdegree' in the Wales network increase the odds of a holding having 'any disease' or TB. Increasing rainfall, location within or at less than 5Km of an AONB and increasing 'indegree' in the Wales network reduce the odds of the presence of TB or 'any disease' in the study population (Table 15).
An alternative model was developed replacing the movement variables by a composite of them as open/closed. The final model using these variables includes 10,047 observations with parish included as a random effect (n=951). The odds of having TB or 'any disease' in an open herd are six times the odds in a closed one when adjusted for SOILCAT, CATDENSITY, HERSIZE, DAYMOVOFF, OUTDEGWAL, 5KMAONB and TOTRAIN.
HERSIZE is the only demographic risk factor in the final model resulting in increased odds of disease, with two movement variables, one environmental and one density. As for Outcome 1, some movement variables appear to have an effect that increases and others have one that reduces the odds of disease in a holding. This aspect is discussed in the "Movement Data" section.

Table 15 Variables included in the random effect logistic regression models for Outcome 2 ('Any Disease' or TB in cattle)

	Movement variables			Open/closed		
Variable names	**OR**	**P value**	**95% CI**	**OR**	**P value**	**95% CI**
SOILCAT (peat soils, marshes, etc.)	3.23	P=0.05	0.99-10.46	2.6	P=0.095	0.84-8
CATDENSITY (>2630)	2.15	P<0.001	1.58-2.91	2.29	P<0.001	1.73-3
HERDSIZE (categorical)	2.09	P<0.001	1.79-2.43	2.97	P<0.001	2.63-3.36
DAYMOVOFF	2.07	P<0.001	1.76-2.43			
OUTDEGWAL	1.02	P=0.012	1-1.04			
5KMAONB	0.57	P=0.001	0.42-0.79	0.61	P=0.001	0.45-0.82
TOTRAIN (categorical)	0.84	P=0.007	0.74-0.95	0.8	P<0.001	0.71-0.9
INDEGWAL (categorical)	0.86	P=0.002	0.78-0.94			
OPENCLOSEDTOTAL				6.36	P<0.001	3.12-12.97

Classification trees:

All variables included in the three trees are shown in Table 16. Classification performance of each of the trees is shown in Table 17.

Table 16. Hierarchical ranking of variables (top=highest; bottom=lowest) in the classification trees for Outcome 2 ('Any disease' or TB in cattle)

Tree 5:1	Tree 10:1	Tree 20:1
DAYMOVOFF	DAYMOVOFF	DAYMOVOFF
CATDENSITY	HERSIZE	HERSIZE
HERSIZE	CATDENSITY	CATDENSITY
TVETNET	FARMA	FARMA
INDEGWAL	DAYMOVON	OUTDEWAL
CATMOVOFF	TOTRAIN	SHEDENCEN
	OUTDEWAL	CATBUF5
	CATBUF5	
	OUTCLOSWAL	
	SHEDENCEN	

Table 17. Classification performance of the three trees using different values for positive misclassification cost for Outcome 2

	Tree 5:1	Tree 10:1	Tree 20:1
Sensitivity	35.1%	70.3%	88.5%
PPV	26.5%	18.5%	13.4%
Specificity	92.4%	70.7%	45.6%
NPV	93.7%	96.2%	97.7%
Error rate	14.1%	29.4%	50.7%
Area under ROC	72.2%	75.8%	71.6%

Three of the top five variables in the three trees are the same as for Outcome 1 (TB): DAYMOVOFF, HERSIZE and CATDENSITY. FARMA appears in two trees (Tree 10:1 and Tree 20:1) and four other variables appear in one of the trees: TVETNET, INDEGWAL, DAYMOVON and OUTDEGWAL. Trees 10:1 and 20:1 have high sensitivity with similar specificity and predictive values. However, only Tree 10:1 has an error rate (29.4%) of less than 50%. Positive predictive values remain low in both trees with a proportion of false positives over 50% in Tree 20:1. The increase in the penalty for misclassification does not result in a change in the type of variables included in the trees. Density, movement and demographic variables are equally distributed across the three sensitivity scenarios. TOTRAIN is the only environmental variable included in the trees.

Classification rules:

The sets of classification rules for the three misclassification penalties contain 41, 53 and 36 rules, respectively. Of these, 19, 24 and 22 predict the outcome of interest. Environmental variables are the most important ones in all sets of rules, with FERTCAT, 5KMAONB and SOILCAT being most frequent amongst them. However, the weight of this variable type decreases with increasing misclassification costs (see Table 18). The top five variables in all sets are DAYMOVOFF, HERSIZE, CATDENSITY, CATBUF5 and TVETNET. The relative importance of the different types of variables for this outcome is much more evenly distributed than for Outcome 1 (TB in cattle). The four groups of factors have a similar role in the 20:1 rules which is reflected by the presence of one variable of each class in the top ranking of variables.

Table 18. Relative contribution of different variable groups to the positive classification rules for classification trees with different misclassification cost ratios for Outcome 2 ('Any disease' or TB in cattle)

	Misclassification cost ratio		
Variable group	**Tree 5:1**	**Tree 10:1**	**Tree 20:1**
Demographics	20.3%	12.5%	21.7%
Densities	12.7%	17.1%	27.8%
Movements	26.6%	35.2%	20.0%
Environment	40.5%	35.2%	30.4%
	100%	100%	100%

Outcome 3: 'Any disease' in sheep holdings

Random effect logistic regression model:

The final random effects logistic regression model for the sheep study population was based on data from 11,695 holdings out of 18,937 and parish was included as a random effect (n=892). Three significant risk factors increasing the odds of disease are demographic variables, two are movement variables and three are environmental variables.
Farmed area, flock size, herd size, 'outdegree', number of sheep moved on to the holding, total rainfall, average temperature and being within or at less than 5km of a National Nature Reserve (NNR) are significant risk factors for the occurrence of any of the specified diseases in sheep (see Table 19).

Table 19. Variables included in the final random effect logistic regression model for Outcome 3 (Any disease in sheep)

Variable names	OR	P value	95% CI
TOTSHE (categorical)	1.85	P<0.001	1.43-2.39
5KMNNR	1.65	P<0.001	1.28-2.13
OUTDEGREE (>1)	1.53	P=0.001	1.19-1.98
SHEMOVON (categorical)			
0	ref		
1-27	1.47	P=0.009	1.1-1.96
>27	1.27	P=0.124	0.93-1.72
FARMA (categorical)	1.33	P=0.03	1.02-1.73
AVETEMP (categorical)	1.29	P=0.003	1.09-1.53
TOTCAT (categorical)	1.21	P=0.02	1.03-1.44
TOTRAIN (categorical)	1.16	P=0.04	1.00-1.35

Classification trees:

All variables included in the two trees are shown in Table 20. Classification performances of each of the trees are shown in Table 21.

Table 20. Hierarchical ranking of variables (top=highest; bottom=lowest) for the two classification trees for Outcome 3 (any disease in sheep)

Tree 10:1	Tree 20:1
SHEMOVOFF	SHEMOVOFF
DAYMOVOFF	DAYMOVOFF
5KMNNR	TOTSHE
OUTDEGREE	5KMNNR
TOTCAT	OUTDEGREE
FARMA	SHEDEN
DAYMOVON	SHEMOVON
SHEDEN	TOTCAT
TOTSHE	FARMA
SHEMOVON	SHEBUF5

Table 21. Classification performance of the two classification trees for Outcome 3 (any disease in sheep)

	Tree 10:1	Tree 20:1
Sensitivity	10.1%	33.6%
PPV	9%	5.8%
Specificity	97.7%	87.5%
NPV	97.9%	98.3%
Error rate	4.3%	13.7%
Area under ROC	63.4%	70.1%

SHEMOVOFF and DAYMOVOFF are the top two variables in both trees with 5KMNNR and OUTDEGREE appearing among the top five variables in both trees as well. Movement variables are the most frequent variables in the trees. 5KMNNR is the only environmental variable. FARMA, TOTSHE and TOTCAT are the three demographic variables present in both trees.
Trees 10:1 and 20:1 have both very low sensitivities. The increase in the penalty for misclassification of positive observations results in a substantial increase in sensitivity from 10.1% to 33.6. The importance of the four groups of variables does not change substantially between the two trees only with replacement of a movement (DAYMOVON) by a density variable (5KMBUFFER).

Classification rules:

The sets of classification rules for the three misclassification penalties contain 21, 55 and 83 rules. Of these, 3, 15 and 33 rules predict the outcome of interest with overall sensitivities of 5.4%, 15.3% and 26.4%, respectively. SHEMOVON appears in 25 of the 33 positive classification rules for Tree 20:1. 5KMBUFFER appears in 11 of the 15 positive classification rules for Tree 10:1. Movement variables are the most frequent group in the three sets of rules accounting for more than 50% of variables with an increasing weight as sensitivity is forced up (see Table 22). Although all movement variables increase their importance with increasing misclassification penalty, SHEMOVON is most affected by the increasing false negative misclassification cost. It only appears once and twice in rules 5:1 and 10:1, respectively, but it is the most important variable in Tree 20:1 through its presence in 25 rules. Environmental variables are the second most important group in the Tree 5:1, but they decrease their presence in the subsequent sets with only 17% in Tree 20:1. Overall, density variables are least influential in the three sets of rules. Demographic variables increase their participation along the rules but only account for 14.8% in the Tree 20:1.

Table 22. Relative contribution of different variable groups to the positive classification rules for classification trees with different misclassification cost ratios for Outcome 3 (Any disease in sheep)

	Misclassification cost ratio		
Variable group	**Tree 5:1**	**Tree 10:1**	**Tree 20:1**
Demographics	7.7%	10.7%	14.8%
Densities	7.7%	17.9%	4.7%
Movements	53.9%	50.0%	63.3%
Environment	30.8%	21.4%	17.2%
	100%	100%	100%

Biosecurity risk scoring system

The information generated by the different modelling approaches was integrated in a qualitative fashion to develop a semi-quantitative risk-score.

Criteria used for including variables in the development of the algorithm

The cut-off points that the classification algorithms identifies for particular variables may vary from tree to tree or from rule to rule. Equally, variables from different groupings have to be considered, such as environmental or population parameters (movement, demographics, densities). However, it was possible to identify a consistent pattern across modelling approaches amongst variables that best discriminated the outcomes of interest. Whereas, variables with low discriminatory power in the classification trees were also not significant in logistic regression models. These variables also had inconsistent patterns in the cut-off points within and between classification trees. The selection of cut-off points for individual variables was conducted based on the outputs from the different analyses as explained in the previous paragraphs. In some cases, the cut-off points produced by the classification trees were used to categorise numeric variables in the logistic regression during the uni- and multivariable analysis. For example, CATDENSITY has been transformed into a binary variable (less/equal *versus* greater than 2,630 per sqkm) in the final logistic regression model, because this cut-off point consistently appears as the best in the classification tree analysis for partitioning the dataset. In the final risk scoring algorithm protocol, CATDENSITY was split into three categories using two cut-off points: the above-mentioned rounded down to 2,500 and another one at 1,000 which is the most frequent cut-off point for the positive (classifying herds with the outcome) classification rules.

We have not differentiated between TB and the pool of other diseases to produce the final protocol since the ultimate goal of this research project component is to classify holdings by their biosecurity status and not for the risk of any particular disease.

Movement data

Two types of movement variables have been analyzed in this study. One aggregates cattle movements on and off the holding regardless of their origin and destination. The second refers to the movements within Wales and the variables are centrality measures of the nodes/holdings of the network built with these paired movements. In a factor analysis of the movement variables, only factors with Eigenvalues greater than 1 were considered (Manly 2005). The analysis generated two factors that meet this condition (see Table 23). Both account for over 90% of total variance. Factor 1 explains 63.7% of the variability and has unrotated positive loadings for all variables. Factor 2 accounts for 28.8% of the variability and has unrotated positive loadings for the variables of the first type and negative loadings for the Welsh network parameters.

The orthogonal varimax solution detects a clear pattern of the loadings of each group of variables. The first group of variables has high positive rotated loadings whereas the network parameters have all low rotated loadings. The factor measures the extent to which holdings move animals to/from any place compared to movements to/from Wales. It could be labelled “movements to/from GB rather than to/from Wales”. The second factor has high rotated loadings for the most important variables of the network and very low ones for the non-network parameters (see Table 24). The most important network variables appear in the top left corner of the scatterplot (not shown here, but in Milestone report) whereas the other group of variables are concentrated at the bottom right corner. The interpretation of this factor together with Factor 1 leads us to conclude that there is a strong component in the variability of the movement data that could be due to holdings that if they move cattle tend to do so within Wales. Holdings with movements in general have different pattern compared with those only contributing to the Welsh movement network.

This analysis was conducted to assist with the interpretation of the results of the three above-described analytical methods and the role of the movement variables in the risk profiles of the study population. The result of the factor analysis allows us to better understand why normal movement variables appear to be strong risk factors for the introduction of TB and other diseases in Wales whereas network

parameters of the intra-Wales network appear to be protective. Even if a holding is open, the risk of introduction of disease will be different according to area of origin and destination. This finding has been incorporated into the risk scoring protocol by increasing the risk score if the holding has many movements and decreasing the risk score if the holding only had movements within Wales.

Table 23: Eigenvalues and proportion of variance of the first four factors

Factor	Eigenvalue	Proportion of variance explained
Factor 1	3.64	63.7%
Factor 2	1.64	28.8%
Factor 3	0.53	9.3%
Factor 4	0.29	5.1%

Table 24: Loadings of the main two factors (orthogonal varimax rotation)

Variable	Factor 1	Factor 2
DAYMOVOFF	0.75	0.15
CATMOVOFF	0.94	0.12
DAYMOVON	0.5	0.12
CATMOVON	0.91	0.12
INDEGWAL	0.41	0.8
OUTDEGWAL	0.03	0.85
BETWAL	0.05	0.88
OUTCLOSWAL	0.006	0.1
INCLOSWAL	0.21	0.1

Risk scoring protocol for cattle holdings

Individual risk profiles based on variables at holding level (demographics/farming and movements) have been combined with area-level variables (environmental and densities). The selection criterion of variables for inclusion in the final risk scoring protocol is as follows: “those included in the logistic regression models for the two outcomes, are among the top five variables in ranking of variables of the classification trees and are the most frequent in the classification rules”. The proposed cut-off points by the classification tress and the positive sets of classification rules have been selected to add incremental scores for the different categories.

The protocol is a sequence of mutually exclusive criteria for the selected variables except for criterion 1 on movements. The conclusion in relation to the assessment of each criterion results in a particular score (see Table 25).

The individual scores can be positive or negative according to the impact of each variable on the risk of the outcomes. Positive scores have three levels (+, ++, +++) and are for outputs that increase the risk of disease/s.

Negative scores have two levels (-, - -) and are for outputs that decrease the risk of disease/s. Outputs that neither increase nor decrease the risk have a null score (0).

The final score of a holding is the sum of all individual scores. Consequently, the final score can range between (-5) and (+12).

The resulting risk scoring algorithm for cattle holdings will result in the following risk groupings: Score <=4 (Low risk), Score 5-8 (Medium risk) and score 9-12 (High risk).

Table 25. Algorithm for biosecurity risk scoring of cattle holdings in Wales (with example interpretation in Results column)

Variable group	Criterion	Response and associated score			Example Result
Demographic/ movements	1.- Open (moving on-off at least one animal):	Yes	No		Yes
	If *No* go to 2	**+**	**-**		**+**
	If *Yes*:				
	1.1. Frequency of movements off to any location:	<5	5-10	>10	5-10
		o	**++**	**+++**	**++**
	1.2 Movements only within Wales:	Yes	No		No
		-	**o**		**o**
	2.- Herd size:	<30	30-80	>80	<30
		o	**++**	**+++**	**o**
	3. - Farmed area (in ha):	<50	50-100	>100	<50
		o	**+**	**++**	**o**
Environment/ densities	4.- Cattle density in the area (in heads per sqkm) :	<1,000	1,000-2,500	>2,500	<1,000
		+	**++**	**+++**	**+**
	5.- Close to/within an Area of Outstanding Natural Beauty	Yes	No		No
		- -	**o**		**o**
	6.- Total rainfall (in mm)	<1,000	>=1,000		<1,000
		o	**-**		**o**
	7.- Land cover	Grassland	Arable		Arable
		-	**o**		**o**
	8.- Soil type	Peat/marshes	Other		Other
		+	**o**		**o**
				Total risk score	**4**

We estimated the Positive Predictive Value (PPV) of the protocol of the high-risk level by calculating the number of holdings that have the outcome “TB and/or any other disease” among the holdings with top scores according to the risk scoring algorithm. These herds meet the following criteria: ‘Herd size > 80’ AND ‘farmed area > 100’ AND ‘number of cattle moved off > 10’ AND ‘in an area with more than 2,500 cattle per sqkm’ AND ‘in arable or other land cover’ AND ‘total rainfall > 1,000’ AND ‘not within 5km of an AONB’ AND ‘in soil other than peat/marshes’. The probability of being positive was 22.6%, which is three times higher than the probability of a randomly selected holding being positive.
A typical low-risk holding fits the following criteria: ‘herd size < 30’ AND ‘farmed area < 50’ AND ‘number of cattle moved off <5’ AND ‘in an area with less than 1,000 cattle per sqkm’ AND ‘within 5km of an AONB’. It has a probability of having a positive outcome (ie any of the selected diseases or TB) of 5.5%, which is lower than the 7.5% prevalence in the study population.

Risk scoring protocol for sheep holdings

A similar approach has been applied to develop the risk scoring algorithm for sheep holdings. The criterion to include variables in the protocol is that they were included in the final logistic regression model, were amongst the top five ranked variables in the classification trees and among the most frequently included in the rules of the classification trees. The cut-off points identified by the classification trees and the positive sets of classification rules were used to add incremental scores for the different response categories for the criteria.
The variables included in the algorithm are: ‘Flock size’, ‘movements on and off the premises’, ‘herd size for mixed holdings’ and ‘proximity to a National Nature Reserve (NNR)’ (see Table 26).
The score ranges between 2 and 11. It is proposed to interpret the calculated risk scores for sheep holdings using the following biosecurity risk groupings: Score <=4 (Low risk), score 5-8 (Medium risk) and score 9-11 (High risk).

Table 26. Algorithm for biosecurity risk scoring of sheep holdings in Wales (with example interpretation in Results column)

Variable group	Criterion	Response and associated score			Example Result
Demographic/ movements	1.- Flock size	<150	150-750	>750	<150
		+	**++**	**+++**	**+**
	2. Mixed holding (with cattle)	Yes	No		Yes
	If No go to 3	**o**	**-**		**o**
	If Yes:	<30	30-80	>80	30-80
		o	**+**	**++**	**+**
	3. Farmed area (in ha)	<15	15-70	>70	15-70
		+	**++**	**+++**	**++**
	4. Movement of sheep off the premises	Yes	No		Yes
		+	**-**		**+**
	5. Movements of sheep on to the premises	Yes	No		No
	If No go to 6	**o**	**-**		**-**
	If Yes:	<50	50-100	>100	
		+	**++**	**+++**	
Environment/ densities	6.- Close to/within a National Nature reserve	Yes	No		No
		++	**o**		**o**
				Total risk score	**4**

A high risk sheep holding would meet the following criteria: 'flock size > 750' AND 'farmed area > 70ha' AND 'herd size if mixed >80' AND 'outdegree > 1' AND 'number of sheep moved on > 100' AND 'within 5km from a National Nature reserve'. It has a 3.7% probability of being positive, which is 1.65 times higher than the probability of being positive for a randomly selected holding (2.2%).
A typical low-risk sheep holding has 'flock size <150' AND 'herd size if mixed <30' AND 'farmed area < 15' AND 'no movements off the premises' AND 'number of sheep moved on < 50' AND 'not within 5km of a National Nature reserve', and a 0.2% probability of being positive, which is substantially lower than the probability of a randomly selected holding of being positive (2.2%).

Discussion and Conclusions

We have developed data-driven models for predicting the presence of a pool of diseases in the cattle and sheep population of Wales, and summarized the results into a biosecurity risk scoring algorithms. The identification of different types of holding attributes that can be accessed without requiring on-farm visits has been a priority throughout this component of the project.
The collation of data from fourteen different data sources had the disadvantage of producing complex datasets with a significant amount of "noise", a large number of variables, non-linear dependency structures, missing values, imprecise data and errors (Dilly 2006). Despite these disadvantages, it has enriched the pool of attributes available for the study units of analysis (cattle and sheep holdings), and resulted in proposals for biosecurity risk scoring algorithms.
We did not differentiate between two cattle outcome variables, bovine TB and the combination of bovine TB and a pool of other diseases, when producing the final risk scoring protocol since the ultimate goal of this research project component was the classification of holdings according to their biosecurity status. We are unable to assess the predictive accuracy for biosecurity, as disease reporting and TB testing data were used as proxy indicators for biosecurity. The combination of the two cattle outcomes ('TB or any disease') has resulted in a dataset with a low prevalence (7.5%) but still higher compared to the dataset with 637 positive observations (4%) using bovine TB as outcome.
However, the disease data source used for the sheep study population only contains diagnoses recorded by the VLA diagnostic laboratories for selected diseases which had been identified in the systematic review under project objective 1. It is likely to be more strongly affected by reporting bias than the cattle data, since the criteria to submit ovine specimens or samples to the VLA might condition the profile of the holdings given the huge difference in the market value between cattle and sheep and the lack of systematic testing in contrast to TB in cattle. Scrapie was not included in the sheep disease data. Data of holdings with confirmed cases of scrapie during 2004 were obtained from the Scrapie Notification Database (SND). However, it was decided to not include the data in the analysis due to the epidemiological features of the disease (chronic, long incubation period, low specificity of clinical signs) and the type of cases recorded in the database, most of them being clinical suspects notified to the Animal Health Divisional Offices (AHDO). As a result, the sheep dataset only contains 425 (2.3%) positive holdings.
Data mining models such as used here are strongly affected by the prevalence of the class of interest. For this reason it was not possible to produce any classification trees based on the misclassification

ratio of 5:1. It was only possible to produce an output when the cost-sensitive metaclassifier was forced to increase the sensitivity of the tree. The performance of the trees and rules in the sheep dataset reached a maximum sensitivity of 33% for the tree with the misclassification ratio of 20:1.
Another premise of this component of the research project is that the analysis of potential determinants of the biosecurity status of a holding would not be based on prior evidence of the true status of a particular holding or group of holdings. Therefore, the biosecurity status would be quantified using the externally measurable factors (EMFs) as proxies of biosecurity. The classification of the EMFs into four distinctive groups has been useful when interpreting the results of the analysis. The integration of area-(environmental and densities) with holding-level (demographics and movements) factors in the risk scoring algorithm prevents the development of a purely spatial predictive model whose main outcome would be the classification of geographic high risk areas in Wales. Demographic and movement characteristics of a particular holding may increase its biosecurity risk score despite being located in an area of low risk. This holistic approach towards biosecurity and risk of disease introduction does not deny the importance of conducting further studies where purely spatial prediction models are to be developed.
The resulting semi-quantitative biosecurity algorithm based on externally measurable risk factors represents an interpretation of the results of the various data exploration exercises conducted under this project component. The relative weighting of the different factors needs to be further optimised within the geographical context of its application. It should then be possible to achieve predictive values that are tailored to local prevalence situations.
The biological and epidemiological interpretation of the different factors was not considered to be the primary objective of this study. As one would have expected, whether a herd was closed or not was the most important risk factor in the analyses and also in the systematic review / EOW. The use of movement factors in this analysis context was novel, and produced some interesting results. Their interpretation is complex, but as a general pattern it appears that the larger the geographical area within which animal movement occurs (only Wales versus GB) the higher the risk of disease. Several of the other risk factors such as being close to a nature reserve or area of outstanding beauty or soil type are likely to be confounded with unmeasured risk factors and there biological interpretation should be explored further. In this context it should be noted that the algorithm derived from the data-driven analysis presented here is tailored to the livestock population it was derived for, ie that of Wales.
The need to provide a new and robust evidence base is a message emphasized not only by DEFRA's new Animal Welfare and Surveillance Strategy, but is now widely acknowledged across Europe (www.euragri.org). Biosecurity is probably one of the areas where policy and advice have been traditionally driven by knowledge and experience. Although valuable, the combination of this expertise with research findings would result in a more robust decision-making process. This idea has been embedded in this project from the development of the original research proposal until the writing of these conclusions. The two components of this project represent the integration of the best research evidence through the systematic review under Project Objective 1 and the development of biosecurity risk scoring algorithms from the quantitative analysis under Objective 2 taking account of the expert knowledge derived from the expert opinion workshop under Project Objective 1.
The results from the two study objectives can be used in combination for the assessment of on-farm biosecurity risk. The selection of farms based on this approach as high or low-risk should be validated in field studies, particularly the use of the EMFs as proxy measures of the biosecurity status of cattle holdings. The value of the risk scores also has to be evaluated as part of the on-farm risk assessment which could be used for informing the development of cost-effective farm-specific biosecurity risk management strategies.

Overall Conclusions

This project has generated two different risk scoring algorithms for biosecurity risk assessment of sheep and cattle farms, one uses externally measurable variables and the other is based on factors which have to be measured on farm. Both algorithms have been based on data of varying quality. In the case of the externally measurable factors it was based on quantitative analysis of a database resulting from compiling relevant animal movement indicators, environmental data etc. into a single database (ie. data-driven approach). The on-farm risk scoring algorithm was based on a combination of systematic review data and expert opinion (ie. knowledge-driven approach). Neither of the two algorithms has been assessed in terms of its predictive value with field data, although for the one based on externally measurable factors it was possible to calculate predictive sensitivity and specificity for the contributing factors in relation to the dataset used in this analysis.

The information generated by the two algorithms could be interpreted jointly in a number of different ways. The risk score derived from externally measurable factors could be used for screening of low biosecurity farms, possibly accepting a significant number of false positives, since these could be filtered out by then applying the on-farm risk scoring algorithm to screening all positives. It would also be possible to consider series or parallel interpretations of the two methods. Following this risk assessment, the data produced by both methods will inform the development of a biosecurity risk management strategy tailored to individual farms. As a next step, the quantitative characteristics of the approaches should be determined in a field study, and the impact of the use of such a biosecurity risk assessment/management approach on farm-level disease patterns and economic performance should be assessed.

References

Assembly of Wales. Summary Statistics. Agricultural Census 2004. http://www.wales.gov.uk/keypubstatisticsforwales/content/publication/agriculture/2005/sb51-2005/sb51-2005-tables.xls (Accessed 16-5-2005)

Breiman L., Friedman J.H., Olshen R., Stone R. (1984). Classification and regression trees. Wadsworth & Brooks. Monterey, California.

Carvalho D.R. and Freitas. A.A. New results for a hybrid decision tree/genetic algorithm for data mining. Proc. 4th Int. Conf. on Recent Advances in Soft Computing (RASC-2002), pp. 260-265. Published in CD-ROM (ISBN: 1-84233-0764), Nottingham Trent University, Notthingham, UK. Dec. 2002.http://www.cs.kent.ac.uk/people/staff/aaf/pub_papers.dir/RASC-2002-Deborah.pdf

Dargatz D.A., Garry F.B., Traub-Dargatz J.L. (2002). An introduction to biosecurity of cattle operations. Vet. Clin. North Am. Food Anim. Pract. 18:1-5.

Dilly R. (1996). Data Mining. An Introduction. Based on material supplied by Sarabjot S. Anand and the DMIG, University of Ulster Jordanstown. Version 2.0, Feb 1996. The Queen's University of Belfast. (http://www.pcc.qub.ac.uk/tec/courses) (Accessed 27-4-2006).

Egger M., Smith G.D. (2001). Principles and procedures for systematic reviews. In M. Egger, G.D. Smith, D.G. Altman (eds) Systematic reviews in health care: Meta analysis in context. BMJ Publishing Group, London, UK. 23-42.

European Agricultural Research Initiative (EURAGRI) (2005). Aims and objectives of the XIX EURAGRI Members Conference 21-23 September 2005. (http://euragri.csl.gov.uk/aims.cfm) (Accessed 25-8-2005).

Gallagher E. (2004). Studies in risk analysis, with emphasis on risk assessment and risk communication. PhD Thesis, University of London, London, UK.

Gilbert M., Mitchell A., Bourn D., Mawdsley J., Clifton-Hadley R., Wint W. (2005). Cattle movements and bovine tuberculosis in Great Britain. Nature 435, 491-496.

Golovnya M. (2005). Data mining with decision trees: an introduction to CART®. Data mining and data quality training course. Madrid 3-5 May 2005. Salford Systems.

Horst H.S. (1998). Risk and economic consequences of contagious animal disease introduction. PhD Thesis, Utrecht University, The Netherlands.

Manly B.F.J. (2005). Multivariate Statistical Methods. A primer. Third Edition. Chapman & Hall/CRC. Florida, USA.

Quinlan J.R. (1993). C4.5: Programs for machine learning. Morgan Kaufmann Publishers: San Francisco, USA.

Radostits O.M. (2001). Control of infectious diseases of food-producing animals, in: Herd Health: Food animal production medicine, 3rd Edition, Saunders Company.

Scudamore J. (2004). Animal Health and Welfare Strategy for Great Britain. Chief Veterinary Officer, London: DEFRA.

Van Schaik G., Nielen M., Dijkhuizen A.A. (2001). An economic model for on-farm decision support of management to prevent infectious disease introduction into dairy farms. Prev. Vet. Med. 51:289-305.

Wasserman S, Faust K. (1994). Social Network Analysis. Methods and Applications. Cambridge University Press: Cambridge, UK.

Witten, I. H. and Frank, E. (2005). Data mining - Practical machine learning tools and techniques. Second edition. (Morgan Kaufmann Publishers: San Francisco, USA.)

Zhang B., Valentine I., Kemp P., Lambert G. (2006). Predictive modelling of hill-pasture productivity: integration of a decision tree and a geographical information system. Agricultural Systems 87 (2006) 1-17.

References to published material

9. This section should be used to record links (hypertext links where possible) or references to other published material generated by, or relating to this project.

Van Winden,S. et al (2005): Preliminary findings of a systematic review and expert opinion workshop on biosecurity on cattle farms in the UK. Cattle Practice, Vol 13, Part 2, 135-140.

Further publications are in preparation.