Comparison between Generalized Linear Modelling and Additive Bayesian Network; Identification of Factors associated with the Incidence of Antibodies against Leptospira interrogans sv Pomona in Meat Workers in New Zealand

Pittavino, M ; Dreyfus, A ; Heuer, C ; Benschop, J ; Wilson, P ; Collins-Emerson, J ; Torgerson, P R ; Furrer, R

Abstract: Background Additive Bayesian Network (ABN) is a graphical model which extends Generalized Linear Modelling (GLM) to multiple dependent variables. The present study compares results from GLM with those from ABN analysis used to identify factors associated with Leptospira interrogans sv Pomona (Pomona) infection by exploring the advantages and disadvantages of these two methodologies, to corroborate inferences informing health and safety measures at abattoirs in New Zealand (NZ). Methodology and Findings in a cohort study in four sheep slaughtering abattoirs in NZ, sera were collected twice a year from 384 meat workers and tested by Microscopic Agglutination with a 91% sensitivity and 94% specificity for Pomona. The study primarily addressed the effect of work position, personal protective equipment (PPE) and non-work related exposures such as hunting on a new infection with Pomona. Significantly associated with Pomona were "Work position" and two "Abattoirs" (GLM), and "Work position" (ABN). The odds of Pomona infection (OR, [95% CI]) was highest at stunning and hide removal (ABN 41.0, [6.9-1044.2]; GLM 57.0, [6.9-473.3]), followed by removal of intestines, bladder, and kidneys (ABN 30.7, [4.9-788.4]; GLM 33.8, [4.2-271.1]). Wearing a facemask, glasses or gloves (PPE) did not result as a protective factor in GLM or ABN. Conclusions/Significance The odds of Pomona infection was highest at stunning and hide removal. PPE did not show any indication of being protective in GLM or ABN. In ABN all relationships between variables are modelled; hence it has an advantage over GLM due to its capacity to capture the natural complexity of data more effectively.

DOI: https://doi.org/10.1016/j.actatropica.2017.04.034

Pittavino, M; Dreyfus, A; Heuer, C; Benschop, J; Wilson, P; Collins-Emerson, J; Torgerson, P R; Furrer, R (2017). Comparison between Generalized Linear Modelling and Additive Bayesian Network; Identification of Factors associated with the Incidence of Antibodies against Leptospira interrogans sv Pomona in Meat Workers in New Zealand. Acta Tropica, 173:191-199.
DOI: https://doi.org/10.1016/j.actatropica.2017.04.034

Comparison between Generalized Linear Modelling and Additive Bayesian Network; Identification of Factors associated with the Incidence of Antibodies against *Leptospira interrogans* sv *Pomona* in Meat Workers in New Zealand

M. PITTAVINO[1#] and A. DREYFUS[2#], C. HEUER[3], J. BENSCHOP[3], P. WILSON[3], J. COLLINS-EMERSON[3], P. R. TORGERSON[2], R. FURRER[1]*

(1) Department of Mathematics, University of Zurich, Zurich, Switzerland
(2) Section of Epidemiology, Vetsuisse Faculty, University of Zurich, Zurich, Switzerland
(3) Institute of Veterinary Animal and Biomedical Sciences, Massey University, Palmerston North, New Zealand

***** *Author to whom correspondence should be addressed: Marta Pittavino, Institute of Mathematics, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland, Office +41 44 63 55897, E-Mail: marta.pittavino@math.uzh.ch*

[#] *Joint first authors*

Received xxx 2016; Final revision xxx 2017; Accepted xxx 2017

ABSTRACT

Background: Additive Bayesian Network (ABN) is a graphical model which extends Generalized Linear Modelling (GLM) to multiple dependent variables. The present study compares results from GLM with those from ABN analysis used to identify factors associated with *Leptospira interrogans* sv Pomona (Pomona) infection by exploring the advantages and disadvantages of these two methodologies, to corroborate inferences informing health and safety measures at abattoirs in New Zealand (NZ).

Methodology and Findings: in a cohort study in four sheep slaughtering abattoirs in NZ, sera were collected twice a year from 384 meat workers and tested by Microscopic Agglutination with a 91% sensitivity and 94% specificity for Pomona.

The study primarily addressed the effect of work position, personal protective equipment (PPE) and non-work related exposures such as hunting on a new infection with Pomona. Significantly associated with Pomona were "Work position" and two "Abattoirs" (GLM), and "Work position" (ABN). The odds of Pomona infection (OR, [95% CI]) was highest at stunning and hide removal (ABN 41.0, [6.9-1044.2]; GLM 57.0, [6.9-473.3]), followed by removal of intestines, bladder, and kidneys (ABN 30.7, [4.9-788.4]; GLM 33.8, [4.2-271.1]). Wearing a facemask, glasses or gloves (PPE) did not result as a protective factor in GLM or ABN.

Conclusions/Significance: The odds of Pomona infection was highest at stunning and hide removal. PPE did not show any indication of being protective in GLM or ABN. In ABN all relationships between variables are

modelled; hence it has an advantage over GLM due to its capacity to capture the natural complexity of data more effectively.

Key words: Leptospirosis, MCMC, R, JAGS, risk factors, protective equipment.

INTRODUCTION

The present study compares results from Generalized Linear Modelling (GLM) with those from Additive Bayesian Network (ABN) analysis by exploring the advantages and disadvantages of these two analytical methods while analysing risk factors for occupational leptospirosis in New Zealand (NZ).

A primary objective of many epidemiological studies is to investigate hypothesized relationships between covariates of interest, and one or more outcome variables. To date, a large variety of statistical models is available to analyse epidemiological data (i.e. cross validation criteria, ANOVA), and one of the most popular are GLM [1]. Typically, the biological and epidemiological processes, which generated this data, are highly complex, resulting in multiple correlations/dependencies between covariates and also between outcome variables. Standard epidemiological and statistical approaches have a limited ability to describe such inter-dependent multi-factorial relationships. ABN is a form of probabilistic graphical model that extends the usual GLM to multiple dependent variables, through the representation of the joint probability distribution of random variables. It is a statistical model that allows the analysis of complex data and derives a directed acyclic graph (DAG) from empirical data, describing the dependency structure between random variables as opposed to fixed variables in GLM [2 3]. ABN models comprise two reciprocally dependent parts: a DAG and a set of parameters. A DAG is a graphical representation of the joint probability distribution of all random variables in the data. Each node in the DAG is the equivalent to the dependent variable in a GLM regression model. In a graphical statistical model there is no distinction between covariates and an outcome variable. Hence, while a standard GLM focuses on the association between covariates and a single dependent or outcome variable, an ABN is a multivariate (conditional) regression model, analysing the associations between all covariates with all variables being potentially dependent [4]. Therefore, in a multifactorial complex disease system, interdependencies between risk factors may be revealed in ABN, that may or may not be discovered in GLM, as the latter imposes a linear relationship between covariates and the outcome [4]. By comparing ABN with GLM using identical data, we explore the likely impact of such an analytical difference on the inferences from this study.

The ABN models described here, also if consisting of a DAG, are only related with statistical dependency, and arcs present in such models do not imply any causal relationship. While the identification of a statistical dependency is often a step towards the conclusion of causal mechanisms, it is, however, more demanding to further claim that the given dependency exists within a particular causal web.

In the last decades, Bayesian Network (BN) modelling has been widely used in biomedical science/systems biology [5-13] to analyse multi-dimensional data. However, only in the last few years, it has been applied in the veterinary epidemiology field. A general introduction to BN modelling in veterinary epidemiology is provided by Lewis et al. [14]. Further applications of BN to veterinary studies were described by Ward et al., Wilson et al. and Sanchez-Vazquez et al. [15-17]. Graphical modelling techniques used to analyse epidemiological data were used by Firestone et al., Schemann et al., Lewis et al., Ludwig et al. and McCormick et al. [2 18-23]. Some of these do not compare results from ABN and GLM [18-20], whereas others do [2 21-23]. In the literature, a detailed comparison of these two methodologies can be found in Lewis et al. [23]. However, the aforementioned study was based on simulated (artificial) epidemiological data and differences of results were mainly discussed with graphical outputs (qualitatively), whereas this analysis also compares ORs of parameters directly and indirectly linked to the outcome, focusing on the contrast as well on a quantitative point of view. 'Additive' BN models have the advantage over the 'classical' BN to allow a direct comparison between the reciprocal model parameters. While BN parameters are based on contingency tables, the resulting data counts, ABN refers to regression parameters resulting from the transformation through a link function (here logit) of the cell probability parameters. Hence, ABNs are more appropriate and suitable for the aim of the presented work.

Leptospirosis is a zoonotic disease occurring in many mammals and is caused by a bacterium of the genus *Leptospira* spp. Transmission occurs from exposure to urine or aborted tissues of infected animals, either directly or via contact with contaminated water or soil [24]. Pathogenic leptospires enter the body through mucous membranes or skin abrasions. In humans, infection with *Leptospira* spp. varies from being sub-clinical (asymptomatic), through a mild to a severe acute disease. A mild form with fever and "influenza-like" symptoms appears to be more common in New Zealand [25]. The acute disease is characterized by jaundice, renal failure, hepatic failure, myocarditis, uveitis and/or pulmonary haemorrhage [26 27].

Among temperate developed countries New Zealand (NZ) has a relatively high incidence of notified

human leptospirosis cases of an average annual incidence risk of 2-3 cases per 100,000 population [28 29]. However, under-ascertainment is common and estimated to be 15-65 fold in sheep abattoir workers [25]. The three most common serovars in humans are *Leptospira interrogans* sv Pomona (Pomona) and *Leptospira borgpetersenii* sv Hardjo (Hardjo) a*nd Leptospira interrogans* sv Ballum (Ballum) [29]. The serovar Pomona is highly prevalent in cattle, deer and sheep in NZ [30-32]. Therefore, livestock are a frequent source of human leptospirosis in farmers and meat workers [28] who are most at risk with less than 10% of deer mobs, sheep flocks or beef herds currently vaccinated against leptospirosis [33 34]. Dreyfus et al. [25] found that in 2011 the annual cumulative Pomona incidence risk (%) in sheep abattoir workers was on average 11.9% with a range for four different abattoirs of 8.4-16.4%. The annual risk of confirmed clinical leptospirosis was 0.78% (3/384, 95% CI 0.20-2.46%) and new infections with Pomona increased the risk of illness with 'influenza-like' symptoms 2.1-fold [25].

This study used the data of the study described above [25] with the following two aims: the first aim was to identify factors associated with Pomona infection in sheep abattoir workers in NZ, with two different methodologies GLM and ABN, in order to untangle the web of causality of human infection with Pomona with a real data set. Specifically, we aimed to test the hypothesis of work position being a strong associated variable, to evaluate the role of personal protective equipment (PPE) and non-work related exposures, such as hunting, home slaughtering and farming. If PPE had a protective effect, it would be a good measure to protect workers. If workers were mainly exposed in their work place and not while hunting or home slaughtering, then it becomes clear where the emphasis on their protection should be. The second and equally important aim was to compare the results between GLM and ABN and discuss advantages and disadvantages of the two statistical analyses.

MATERIALS AND METHODS

Case study

A prospective cohort study amongst voluntarily participating meat workers from four purposively selected sheep abattoirs in the North Island of NZ was conducted. Study methods were described in detail by Dreyfus et al. [25]. Participants were blood sampled by certified phlebotomists or nurses and interviewed at the same time by trained researchers using a questionnaire (Supplementary Material). Serum antibodies against Pomona

were analysed by the microscopic agglutination test (MAT) at doubling dilutions from 1:24 to 1:1536 as described previously [35]. Blood samples and data were collected twice at intervals ranging from 50 – 61 weeks in order to estimate the incidence of new infections with Pomona. Study participants of "Abattoir 1" were sampled the first time between February and April 2008 and the second time in April 2009. All other abattoirs were sampled initially in November 2009 - March 2010, and again in November 2010 - May 2011. Hence, one abattoir ("Abattoir 1") was studied twice in two consecutive years and three abattoirs were studied in the second year once. New infection occurred where a worker sero-converted (a sero-negative worker had a MAT titre increase to equal or higher than 1:48) or had an anamnestic response (a sero-positive worker had a MAT titre increase by two or more dilutions) [25].

Data structure

Serological test results and questionnaire information were entered into an Access© database. The longitudinal data resulting from the serological test results and interviews comprised of 384 observations across 13 variables, including the outcome variable (Table 1). There were no missing data. Work positions were categorized in four binary variables: "Work0" included workers with no or presumed low exposure to organs of the urinary tract or to urine and they worked in the office, "boning" room (where the carcass is cut into pieces), "chillers", "freezers" or "blood processing"; "Work1" included workers from areas where organs were handled, such as the "offal"/ "casing"/ "pet food", hide processing positions, also including cleaners, renderers or engineers; "Work2" included workers at the middle and end of the slaughter board, where animals were opened, organs removed and carcasses were inspected; in "Work3" were workers in the yards, where animals were washed and waiting for slaughter and at the beginning of the slaughter board, where animals were stunned, bled and hides were removed.

Workers were asked about the PPE worn for every task in the abattoir. PPE variables were "Facemasks" (mask with movable transparent protective shield covering the whole face), "Safety (= goggles) or normal glasses" and "Gloves on two hands" (made from latex, or similar material or plastic). They were further asked about the frequency PPE was worn. Frequency category 1 was "always or often" and frequency category 0 "sometimes or never". Further variables of interest were number of months worked during the study and in the three preceding seasons, years worked in an abattoir, whether workers went hunting, were farming, home

slaughtering in the study year and the previous three years and personal data such as age and gender. Each of the four participating abattoirs (one participated twice) was a binary variable ("Plant0", "Plant1", "Plant2", "Plant3", and "Plant4"). All variable names and their description are presented in Table 1.

Binary variables of possible multi-categorical ones (i.e. work position or abattoir) have been created, in order to analyze the data with the ABN methodology. The latter approach is implemented in the R package "abn", which so far has a limited functionality to handle categorical variable. To address over parametrization, collinearity and possible overlapping information the variables "Work0" and "Plant0" were omitted from the GLM and ABN model.

Exploratory data analysis (EDA) with correlation matrices (Figure 1 in [36]), captured using the Spearman's correlation coefficients, and parallel coordinate plots evaluation was conducted on the raw data to test the correlation between predictor variables.

Analysis with GLM

Data were analysed using the software R, with the version 3.1.2 [37]. Crude associations between the risk of infection with Pomona and potential risk, protective or confounding factors listed in Table 1 were calculated by univariable analysis. We used multivariable logistic regression (MLR) to test the hypotheses that work position, hunting, slaughtering at home and farming and working in a specific abattoir were risk factors and PPE was a protective factor for new infection with Pomona. We evaluated risk factors and confounding variables by a manual forward stepwise selection in the MLR model, starting with a null model with only an intercept included and then adding one risk factor at a time. A variable was retained if the Likelihood Ratio Test (LRT) was statistically significant at a p-value <0.05 or if its presence changed the OR of another statistically significant variable in the model by more than 15% (= confounder) [38]. Interaction between risk factors was tested by LRT. If the LRT was statistically significant ($p<0.05$) and the interaction term statistically significantly associated with the outcome ($p<0.05$), the interaction term was retained in the model. All possible interactions between the variables "Glov", "Glass", "Mask" and "Work" and between "Work" and "Plant" have been tested

Given the hierarchical structure "abattoir-worker" with potential clustering by abattoir and the fact that 57 workers from Abattoir 1 participated twice in the study, we fitted a multilevel mixed model (GLMM) using

abattoir as a random effect, in order to evaluate the effect of clustering by abattoir on the model outcome. Results of GLMM were compared to the ones of GLM and adjustment only kept if results or precision were altered by GLMM in a fashion that conclusions would change.

The Hosmer-Lemeshow statistic was used to test the distributional assumption and the Pseudo R-square was used to evaluate the overall model fit. Influential covariate patterns and leverage were examined using described methods [39].

Analysis with ABN

All analyses were conducted using the software R, version 3.1.2 [37] and specifically the R package "abn" [40] which is maintained by one of the two principal authors and is available from CRAN (cran.r-project.org) with additional documentation and case studies at "http://www.r-bayesian-networks.org/". The resulting networks were manually created with the programme Xfig.

Prior distributions were defined: all DAG structures were equally supported by a uniform prior in the absence of any data. A uniform prior was used to guarantee that no structure was preferred over the others, to allow a fully data-driven approach. While, uninformative Gaussian priors were applied for the parameters at each node: specifically, independent Gaussian priors with mean zero and variance 1000 for the additive terms, equivalent to beta coefficients in a conventional logistic regression, and a diffuse Gamma distribution with shape and scale of 0.001 for the precision, i.e., the inverse of the variance parameter in the Gaussian nodes.

A three-step procedure was utilized to determine a robust model for the case study data and to estimate the parameters. The first step was to find an optimal model, represented by the DAG, which is a graphical representation of the joint probability distribution of all the random variables where no cycles exist. The best goodness of fit to the available data was computed using the marginal likelihood method, which is the standard goodness of fit metric in Bayesian modelling and includes an implicit penalty for model complexity. This was estimated using the Laplace approximation at each node [41]. The process of identifying an optimal DAG is referred to in the literature as structure learning [42 43]. This was found with an order based exact search method [44], which determines a DAG with goodness of fit being equal to the best possible goodness of fit of any DAG. In order to find the best DAG, the maximum number of parents allowed per node (= number of covariates in each regression model at each node) was increased until the goodness of fit remained constant

and thereby identified the same globally optimal DAG. The model selection procedure started from three possible parents per node and then the parent limit increased gradually until ten possible parents per node (Figure 3 in [36]). A best fitting ABN was identified at the end of this first step, with a maximum number of possible parents per node.

In the second step, the model was adjusted by checking it for over-fitting [45] using Markov chain Monte Carlo (MCMC) simulation implemented in JAGS ('just another Gibbs sampler') [45 46]**.** Before proceeding with this step, it is essential to first visually check the marginal densities estimated from the initial ABN model identified (Figure 2 in [36]), and verify that the area under the curve (posterior density) is one (Figure 4 in [36]). Simulated datasets were generated with MCMC as iterations of an identical size as the original one, from the optimal model found in step one. An identical exact search for an optimal model structure was then performed exactly as in the first step, but applied to the bootstrapped data rather than original data. It was repeated 2560 times (Figure 6 in [36]), a large enough number to get robust results, using the same parent limit per node as the one found in the initial search. Arcs present in less than 50% (dashed lines in Figure 2 in [36]) of the globally optimal DAGs – estimated from the bootstrapped data – were considered not to be robust and removed from the DAG generated in the first step. A threshold of 50% structural support is the usual cut-off in ABN analysis [4]. For sensitivity analysis, the arcs coverage after 640 and 1280 simulations were compared. A most robust ABN model fully adjusted for over-fitting was identified at the end of this second step, equivalent to a multivariate GLM. The R package coda [47] was used to evaluate the mixing of MCMC chain. Both visual and statistical techniques have been used with the Gelman and Geweke diagnostics [48].

In the third step of ABN analysis, the marginal posterior log odds ratio and 95% credible intervals were estimated for each parameter from the posterior distribution (Figure 6 in [36]), expressed by the DAG identified at the second step. Being in a Bayesian statistics framework, the parameters were the maximum likelihood estimates (MLE) based on the joint posterior distribution. With ABN methodology, it is possible to evaluate the association between all variables, including the outcome and hence evaluate all relations present in the data. An arc between two variables in the final ABN model is referred to as a "direct" relationship, whereas an "indirect" relationship is defined as two arcs connecting two variables with an intermediate variable. For example, Figure 1 shows variables "Pomona" (= Pomona infection) and "Mask" (= wearing a

facemask) being “indirectly” linked through the presence of work position (“Work1”, “Work2” and “Work3”) variables that are all “directly” linked to “Pomona”.

In order to estimate the parameters of the linked variables, a specific function (fitabn) of the R package “abn” was used. With the latter, it is possible to compute the odds ratio at each node, connected with an arc in the final model as reported in Figure 1 on top of the link between the variables of interest.

At the end of this third step, the marginal posterior odds ratio of the main variables in the analysis and their 95% credibility intervals were obtained. Data and R codes are available in [36].

RESULTS

At the beginning of the study 567 workers were recruited and blood sampled. The number of participating workers when resampled was 384, and ranged by abattoir from 21-135 (Table 1). The loss to follow up in our cohort was hence 32%. The main reasons were fear of pain at sampling, having already left work for the day, having left employment at the abattoir or been laid off for the season.

The exploratory data analysis revealed the strongest correlation between the continuous variables “Age” and “Time” (c=0.61), the variables “Work3” and “Gender” (c=0.38) and “Work2” and “Mask” (c=0.33). The variable “Pomona”, with 36 positive cases, was mainly linked with variable “Work3” (c=0.2), for all the other variables there was a correlation coefficient < 0.15. A similar pattern in the data was reflected in the results from ABN analysis.

Risk factors for new infection with Pomona analysed by GLM

Statistically significant risk factors in the final GLM model were “Work position” and working in a specific abattoir (Table 2). Workers in the offal room (“Work1”) had 22.1 times (95% CI 2.3-209.8, p=0.001), workers removing the intestines and kidneys, and meat inspectors (“Work2”) had 33.7 times (95% CI 4.2-271.1; p<0.001), and workers stunning, pelting and working in the yards (“Work3”) had 56.96 times (95% CI 6.8-473.3; p<0.001) the odds of infection with Pomona compared to the workers from the other work categories. These associations were independent of working in a specific abattoir. Persons working at abattoir 2 had 4.5 (95% CI 1.9-10.67; p<0.01) times and persons working at abattoir 4 had 3.4 times the odds of infection (95% CI 1.3-8.9; p=0.01) compared to workers working in all the other abattoirs, irrespective work position.

Even though the variable Gender (“Sex”) was not statistically significant and did not improve the

model fit, it was left in the model as a potential confounder, as it changed the work position OR by ≥15% (Table 2). None of the other potential risk or protective factors or confounders were significantly associated with new infection in GLM and did not improve the model fit. None of the tested interactions were statistically significantly associated with Pomona infection, however, "Work2*Plant4" did improve the model fit in the likelihood ratio test and increased the odds ratio of Work position 2 (handling kidneys) from 33.7 to 51.3.

The GLMM reduced the effect of the variables on the outcome (OR) by less than 10% (see Table 3 in [36]. However, the significance and the weighting, such as work position having the strongest association with Pomona, remained the same. Therefore, the GLM was retained as model formulation.

Model diagnostics indicated that the data fitted the logit-normal distribution. One outlier was identified, but its removal and collapsing work position categories zero and one did not change any of the statistical significant model coefficients by more than 8% and hence, did not impact on the inferences.

Association between variables analysed by ABN

The resulting best fitting ABN comprised 30 arcs and a maximum number of seven parents (Figure 3 in [36]), for the variable "Gender". The MCMC revealed a good mixing of the chain, with no evidence of non-convergence toward the stationary distribution resulting from the Gelman and Geweke diagnostics. After the bootstrap analysis, four of the arcs in the globally optimal DAG were only weakly supported. Therefore the number of arcs was reduced from 30 to 26 (Figure 2 in [36]). Identical results were obtained in the sensitivity analysis, where we started with 2560 bootstraps (Figure 5 in [36]), which was a large enough number to generate robust results, and then performed half (1280) and a quarter (640) of the bootstrap analyses, suggesting a robust conclusion. The final globally optimal additive Bayesian network model after adjustment for over-fitting is shown in Figure 1. The ABN models considered here are concerned only with statistical dependency, and arc direction in such networks has no epidemiological interpretation. Therefore, the graphical models are presented without arc direction.

In the final ABN model shown in Figure 1, the only variables directly linked to Pomona infection were work positions. More specifically, people working in stunning, pelting and yards had the highest odds of infection, compared to those who were not working in this particular category, as the odds of infection with Pomona was 41.0 (95% CI 6.9-1044.1) times higher than in workers not working in these positions. Workers removing the intestines and kidneys and meat inspectors had 30.7 (95% CI 4.9 -788.4) times the odds of

infection with Pomona compared to workers not working in these positions. Workers removing offal and pet food had 18.3 (95% CI 2.2 – 506.7) times the odds of infection with Pomona compared to workers not working in these positions (Table 2). As illustrated in the final DAG, work positions were strongly inter-dependent with PPE.

DISCUSSION

In the last four decades, four cross-sectional studies investigated *Leptospira* sero-prevalence in meat workers in NZ [49-52] estimating sero-prevalences against Pomona, Hardjo, and/or *Leptospira borgpetersenii* sv Tarassovi of being between 4.1% and 31%. One longitudinal study investigating risk factors for *Leptospira* incidence risk in abattoirs has been conducted recently [53]. However, in the latter study, the *Leptospira* serovars contributing to "new infection" were the two serovars Hardjo and Pomona as a combined outcome. Since a study found that risk factors for Hardjo and Pomona infection in livestock varied substantially [30], and since Pomona was associated with the majority of new infections in workers in all abattoirs and with more signs of flu-like illness as opposed to Hardjo, we omitted Hardjo infection from our analysis outcome in this analysis. The analysis of risk factors for new infection with Pomona is therefore novel and has not been done in the former study [53].

Had serovar Hardjo been associated with more than 13 new infections, we would have incorporated it in the analysis as a variable, as ABN could have demonstrated the dependencies between Hardjo, Pomona and all other variables, differentiating the roles of these two serovars in the risk factor scenario at sheep abattoirs for leptospirosis. However, given the few sero-positive cases, one third with respect to Pomona new infection, it would have resulted in a poorer model fit. Further, in GLM it would have been nonsensical to include Hardjo infection as a risk factor for Pomona infection in the outcome. Hence, an inclusion of Hardjo was not possible for comparing the two methods.

The objective of the presented analysis was to identify risk factors for new infection with Pomona in sheep abattoir workers and to compare results from GLM with those from ABN. GLM and ABN confirmed the hypothesis that work position was the strongest risk factor for new infection with Pomona in sheep abattoir workers (Table 2). Hence, both methods appeared to be appropriate for identifying strong associations. ABN models are multidimensional multivariate regression models and analyse associations between all variables at the same time [4]. Therefore, ABN and GLM are likely to identify the same risk factors when associations are strong and highly significant.

The work position variables ("Work1", "Work2", "Work3" in Figure 1) were all significantly related to Pomona infection ("Pomona") in ABN and in GLM. However, while showing the same trend (OR of Work3>Work2>Work1) in ABN the odds ratios were generally lower. The lower value of the OR in ABN is due to the holistic structure of the model, which is considering all the variables at the same time while in GLM only a selection of them is considered. The multivariate ABN can be viewed as a collection of multivariable models (GLM) along the arcs of the DAG, hence the parameter estimates are expected to be identical given the same explanatory variables. However, compared with the GLM, the ABN model is regarded as more flexible since each level of a categorical variable can potentially have different sets of dependencies to other variables, whereas GLM only associates independent variables with a single outcome. This flexibility was apparent with the variable 'wearing gloves' ("Gloves"), which was only connected to the removal of intestines/kidneys/meat inspection ("Work2") in ABN, suggesting that the risk attribution to wearing the PPE depended on work position. This hidden dynamic was not detected while using GLM also when interaction terms were considered. Hence, ABN has an advantage over GLM because it can disentangle the complex nature of the data, stratifying further the internal mechanisms present between the variables.

As already discussed by Dreyfus et al. [53], the highest odds of infection in workers at the beginning of the slaughter board may be explained by contact with contaminated droplets due to frequent urination of stunned sheep. The relatively high odds of infection during removal of kidneys and at meat inspection may be attributable to direct exposure of workers to Pomona residing in the genital-urinary system. Kidneys pass through the offal room, possibly explaining the odds of infection in that working area.

Wearing PPE at the work place (gloves, facemasks and glasses) were not statistically significantly associated with Pomona in GLM analysis. Hence, they did not result as protective factors in the GLM (Table 2). Descriptive analysis supported the lack of protection of Pomona infection by PPE: of 12 workers infected with Pomona in "work position 3" (stunning, hide removal) (Table 1), ten, four and 11 workers reported to always or often have worn gloves, and/or facemasks and or safety or normal glasses, respectively (Table 1 and 2 in [36]. This may be biologically plausible because workers wearing safety goggles or facemasks reported they sweat and presumably wipe their eyes with potentially contaminated hands, which eliminates the protective effect. We recommend research to clarify whether this is actually true (e.g. by detecting *Leptospira* DNA in facemasks and glasses by PCR). Our findings about PPE should be interpreted with some caution, as there is a possibility of differential misclassification bias. When responding to questions about wearing PPE,

participants may have overstated the use of PPE and not admitted non-compliance to the employment policy enforcing the use of PPE, despite a clear statement that interviews were confidential. This may have led to an overstatement of wearing PPE by meat workers in exposed work positions, reducing the chance of determining a protective effect of PPE in the analysis. Nevertheless, we believe that such bias was small because workers handling kidneys were, contrary to belief, less likely to sero-convert than workers at stunning/pelting.

Hunting, farming or slaughter of animals at home were not associated with Pomona infection in the GLM and only indirectly linked to Pomona infection through three to four arcs in the ABN model. This is an indication that in this study population exposure to Pomona was more likely occurring in the abattoir than through contact with livestock at times off-work. This finding underlines the role of leptospirosis as an occupational hazard in sheep abattoirs in NZ. These findings were confirmed in the study on sero-prevalence/incidence and risk factors by Dreyfus et al. [52 53], but contrast with the findings of Heuer et al. [54], where home slaughter was found to be a risk factor for sero-prevalence of Hardjo or Pomona, where Hardjo titres were 5-fold more frequent than Pomona titres among workers of one abattoir (A1).

An advantage of ABN is the illustration of the dependencies between all variables by the graphical model (Figure 1), compared with GLM which only shows the dependencies between risk factors and outcome (Pomona infection). Hence, the GLM only identified variables that were directly associated with Pomona infection, and was restricted to a limited model space ignoring indirect relationships. Conversely, ABN considered all variables jointly allowing arcs to be present between any variables. The DAG illustrates that farming was dependent on slaughter of animals at home, which was associated with working at the yards, stunning and pelting ("Work1"). Hence, persons working in these positions were more likely to slaughter at home and farm. Hunting was associated with the variables "Age" and "Time", meaning that older, long time workers were more likely to go hunting.

The Yule-Simpson paradox [55] states that taking a narrow univariate (single dependent variables/multivariable regression) approach to risk factor analysis will, in general, not give the same result as a joint and truly multivariate approach [4]. In this study, ABN and GLM methodology did not produce exactly the same results: whereas work position was the only directly dependent variable upon new infection with Pomona in both ABN and GLM, the GLM found workers in Abattoir 2 and 4 to be at higher odds of infection than workers from the other participating abattoirs. As shown in Figure 1, ABN suggested that the odds of infection in Abattoirs was indirectly linked to the outcome "Pomona infection" through wearing normal or

safety glasses (“Plant1”, “Plant2” and “Plant3”) or wearing a facemask (“Plant 4”) and all three work positions. The results from the ABN method suggest that Pomona infection occurred more often in abattoirs in association with the use of the above mentioned specific PPE. GLM had not detected these associations. Hence, while GLM only established a direct association between two abattoirs and Pomona infection, ABN identified a network of inter-dependent factors linked to the outcome. Here ABN was more informative about potential causal pathways in the disease system than GLM [4]. This potential advantage of the ABN method would specifically be useful for observational studies with large number of variables, where causal and time relationships are often unknown.

The technical foundations of ABN modelling lie within the machine learning and data mining literature [42 43 56-58]. The main obstacle of using this methodology in practice is that it can be computationally rather demanding: determining the best model for a given data set has been shown to be NP-hard [59], the most difficult class of computational problem. This means that finding an optimal model must be done using heuristic search algorithms [42 43 56-58], rather than brute force computation. Hence, another limitation of ABN modelling is the restriction of the number of variables. To date, exact structure discovery with bootstrapping is only feasible for around 20 variables [44]. Heuristic searches [43] and inexact order-based searches [58] for globally optimal DAGs offer an alternative, however they are less ideal approaches, because they are local techniques and not exact methods. In the future, advances in either technology or statistical methods will make larger computationally intensive analysis more feasible. An example could be that presented by [60]. Moreover, another drawback of the current ABN methodology is the unfeasibility to take into account possible interactions. However, we checked possible effect modifications for the GLM model, as clarified in the previous sections, but they revealed to be not significant. Therefore, although if this possible limitation is present, it does not harm our analysis and the two methodologies are still comparable due to absence of significant effect modifications.

The credible intervals in both models were very high, due to the few sero-positive cases and in particular the right-side intervals were wider in the ABN than in the GLM results. This is due to the model nature, where all the variables are considered due to the joint mathematical model formulation, despite the use of GLM techniques in the estimation process.

The stepwise algorithm used to select the best GLM model is not always recognized as the standard procedure for model selection [23]. Nevertheless, in this context, the stepwise approach was appropriate, as

the choice of the frequency of variables put into the model was based on hypotheses, formulated with knowledge from former studies [51 53], with knowledge of infection pathways and the epidemiology of leptospirosis.

One abattoir ("Abattoir 1") was studied twice in two consecutive years and three abattoirs were studied in the second year once. However, the data of one of the abattoirs studied twice was omitted in the GLM and ABN analysis. Since only 14.8 % of the study population were sampled repeatedly in "Abattoir 1" and since sero-conversion and anamnestic responses (= Pomona incidence) were measured, and not sero-prevalence, clustering was expected to be at very low level. This was confirmed when we conducted the GLMM model and the results changed by less than 10%. Moreover, since only four abattoirs participated, working in a specific abattoir was a hypothesized risk factor, and interaction between work positions and abattoirs was regarded as potentially important, we preferred keeping the simpler GLM over a GLMM.

Although the abattoirs were not fully representative of the whole of NZ, being solely located in the west and east of the North Island, the animals originated from all areas of the North Island. Furthermore, as demonstrated in [52], the study population was recruited from almost 20% of the total sheep abattoir worker population.

Since participation was voluntary, it was likely a sampling bias had been introduced. A parallel analysis revealed that workers from high exposed work positions were more likely to participate [61]. But this did not affect the results from the multivariable logistic regression analysis where working area was included as a covariate.

The MAT titre cut-off of 1:48 is appropriate to determine exposure to leptospires in humans, but is generally not recommended as a cut-off for diagnosing clinical disease [35 62]. Hence, a two-fold increase can determine new infection due to exposure to leptospires, but is not necessarily appropriate to diagnose clinical disease. The latter requires, by WHO definition, a single MAT antibody titre ≥800 or a four-fold increase in the convalescent blood sample. However, this definition has been challenged recently [30 63], as infection with certain serovars seem to lead to clinical disease with lower antibody responses. Seroconverting meat workers in this study had a two-fold risk of influenza-like symptoms compared to workers not seroconverting [25]. Fang et al. [64] modelled the association between *Leptospira* sero-positivity and risk factors in meat workers of one sheep abattoir for different MAT cut-offs. While the percentage of sero-positive meat workers reduced by approximately 40%, when choosing a MAT titre cut-off of 1:96 rather than 1:48, the conclusions on risk

factors did not change. In many countries a wide range of serologically related serovars is prevalent introducing the problem of cross reactivity in the MAT. However, the prevalence of six endemic serovars in NZ, which belong to different serogroups, should reduce the problem of cross reactivity [65]. The MAT in the NZ context is therefore very specific and false positives should not represent a problem in this study context. In a study evaluating the MAT sensitivity and specificity of acute (MAT cut-off 1:100) and convalescent (MAT cut-off not mentioned) sera in an urban setting in Brazil [66], the MAT testing of convalescent sera had a sensitivity of 91% to 100% and specificity of 94% to 100%. If we assumed that the MAT in our study had a 91% sensitivity and 94% specificity, the tested incidence in meat plants was likely under-estimated. However, since we used a MAT titre cut-off of 1:48 and tested for the serovars Hardjobovis and Pomona, which are less likely to be encountered in a urban setting, where serovar Copenhageni is predominant [66], it is possible that the sensitivity and specificity of the MAT in NZ are not the same as in Brazil.

In conclusion, this study demonstrated that workers were at highest odds of new infection when working at the beginning of the slaughter (stunning and hide removal), followed by those removing intestines, bladder and kidneys, and workers in the offal/pet food area. PPEs like facemasks and safety glasses did not show any indication of being protective in GLM and ABN and descriptive analysis supported the lack of protection of Pomona infection by PPE. Further, other means of protection might be considered, like vaccination of farmed livestock or slaughter procedure changes. ABN has an advantage over GLM due to its capacity to capture and illustrate graphically the natural complexity of data more effectively. In ABN, all relationships between variables are modelled, which appears to be more explanatory in view of the inter-dependencies between study variables in complex disease systems.

ACKNOWLEDGEMENTS

The authors are indebted and grateful to study participants, managers and health and safety workers of the participating abattoirs, nurses and phlebotomists, without whom the study would have been impossible. The ABN methodology was part of the PhD project of Marta Pittavino. She thanks Fraser Ian Lewis to have introduced her to this new methodology and Gilles Kratzer for the interesting scientific exchanges about additive Bayesian networks approach.

Marta Pittavino expresses her deep gratitude to the Foundation ‘‘Franco and Marilisa Caligara per l’Alta Formazione Interdisciplinare’’ for its support before and during the PhD studies.

We thank Sarah Moore for tireless support of sampling and data management, occupational health physicians John Reekie & John Kerr for advice and communication with abattoirs, Heather Duckett for helping to organize sampling, Christine Cunningham and Wendy Maharey for administrative support, Brian O`Leary, Masood Sujau and Simon Verschaffelt for help developing the database, Fang Fang, Prakriti Bhattarai, Rayon Gregory, Claire Cayol and Emilie Vallee for interviewing, Neville Haack and Rae Pearson for MAT testing Roger Lentle for advice for the Massey University Human Ethics Committee application and Lesley Stringer and Sarah Rosanowski for analytical and software support. Further, we thank and the Department of Labour NZ for support.

FINANCIAL SUPPORT

We gratefully acknowledge funding by Rural Woman New Zealand, and commissioned by the Tertiary Education Commission (TEC) via the Institute of Veterinary, Animal and Biomedical Sciences, Massey University (TEC #RM12703 (2008)), and the Swiss National Science Foundation (PBBEBS-124186, SNF138562 and SNF144973).

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

ETHICAL STANDARDS

All procedures were approved by the Massey University Human Ethics Committee in 2008 and 2009 (HEC: Southern A, Application 05/123 and 09/08, Slaughter carcasses as a source for human infection with Leptospira serotypes Hardjo, Pomona and Ballum at abattoirs in New Zealand).

REFERENCES

1. McCulloch CE, Searle S. R., Neuhaus JM. *Generalized, Linear, and Mixed Models*: Wiley, 2008.
2. Sivasundaram S, ed. Bayesian Networks as a tool for Epidemiological Systems Analysis. 9th international conference on mathematical problems in engineering, aerospace and sciences (ICNPAA 2012); 2012; Melville, NY 11747-4501 USA. American Institute of Physics.
3. Rijmen F. Bayesian networks with a logistic regression model for the conditional probabilities. International Journal of Approximate Reasoning 2008;**48**(2):659-66
4. Lewis FI, McCormick BJJ. Revealing the complexity of health determinants in resource-poor settings. Am. J. Epidemiol. 2012;**176**(11):1051-9 doi: 10.1093/aje/kws183[published Online First: Epub Date]|.
5. Lycett SJ, Ward MJ, Lewis FI, Poon AFY, Pond SLK, Brown AJL. Detection of Mammalian Virulence Determinants in Highly Pathogenic Avian Influenza H5N1 Viruses: Multivariate Analysis of Published Data. Journal of Virology 2009;**83**(19):9901-10 doi: 10.1128/jvi.00608-09[published Online First: Epub Date]|.
6. Poon AFY, Lewis FI, Frost SDW, Pond SLK. Spidermonkey: rapid detection of co-evolving sites using Bayesian graphical models. Bioinformatics 2008;**24**(17):1949-50
7. Poon AFY, Lewis FI, Pond SLK, Frost SDW. Evolutionary interactions between N-linked glycosylation sites in the HIV-1 envelope. Plos Computational Biology 2007;**3**(1):e11
8. Poon AFY, Lewis FI, Pond SLK, Frost SDW. An evolutionary-network model reveals stratified interactions in the V3 loop of the HIV-1 envelope. Plos Computational Biology 2007;**3**(11):e231
9. Dojer N, Gambin A, Mizera A, Wilczynski B, Tiuryn J. Applying dynamic Bayesian networks to perturbed gene expression data. Bmc Bioinformatics 2006;**7**:249
10. Hodges AP, Dai DJ, Xiang ZS, Woolf P, Xi CW, He YQ. Bayesian Network Expansion Identifies New ROS and Biofilm Regulators. Plos One 2010;**5**(3):e9513
11. Jansen R, Yu HY, Greenbaum D, et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science 2003;**302**(5644):449-53
12. Needham CJ, Bradford JR, Bulpitt AJ, Westhead DR. A primer on learning in Bayesian networks for computational biology. Plos Computational Biology 2007;**3**(8):e129
13. Djebbari A, Quackenbush J. Seeded Bayesian Networks: Constructing genetic networks from microarray data. Bmc Systems Biology 2008;**2**:57
14. Lewis FI, Brulisauer F, Gunn GJ. Structure discovery in Bayesian networks: An analytical tool for analysing complex animal health data. Prev. Vet. Med. 2011;**100**(2):109-15
15. Ward MP, Lewis FI. Bayesian Graphical modelling: Applications in veterinary epidemiology. Prev. Vet. Med. 2013;**110**(1):1-3
16. Wilson AJ, Ribeiro R, Boinas F. Use of a Bayesian network model to identify factors associated with the presence of the tick Ornithodoros erraticus on pig farms in southern Portugal. Prev. Vet. Med. 2013;**110**(1, SI):45-53
17. Sanchez-Vazquez M, Nielen M, Edwards S, Gunn G, Lewis F. Identifying associations between pig pathologies using a multi-dimensional machine learning methodology. BMC Veterinary Research 2012;**8**(1):151
18. Firestone SM, Lewis FI, Schemann K, Ward MP, Toribio J-ALML, Dhand NK. Understanding the associations between on-farm biosecurity practice and equine influenza infection during the 2007 outbreak in Australia. Prev. Vet. Med. 2013;**110**(1, SI):28-36
19. Firestone SM, Lewis FI, Schemann K, et al. Applying Bayesian network modelling to understand the links between on-farm biosecurity practice during the 2007 equine influenza outbreak and horse managers' perceptions of a subsequent outbreak. Prev. Vet. Med. 2014;**116**(3, SI):243-51
20. Schemann K, Lewis FI, Firestone SM, et al. Untangling the complex inter-relationships between horse managers' perceptions of effectiveness of biosecurity practices using Bayesian graphical modelling. Prev. Vet. Med. 2013;**110**(1, SI):37-44

21. Ludwig A, Berthiaume P, Boerlin P, Gow S, Léger D, Lewis FI. Identifying associations in Escherichia coli antimicrobial resistance patterns using additive Bayesian networks. Prev. Vet. Med. 2013;**110**(1):64-75
22. McCormick BJJ, Sanchez-Vazquez MJ, Lewis FI. Using Bayesian networks to explore the role of weather as a potential determinant of disease in pigs. Prev. Vet. Med. 2013;**110**(1):54-63
23. Lewis F, Ward MJ. Improving epidemiologic data analyses through multivariate regression modelling Emerging Themes in Epidemiology 2013;**10:4**
24. Hartskeerl RA, Collares-Pereira M, Ellis WA. Emergence, control and re-emerging leptospirosis: dynamics of infection in the changing world. Clinical Microbiology and Infection 2011;**17**(4):494-501 doi: 10.1111/j.1469-0691.2011.03474.x[published Online First: Epub Date]|.
25. Dreyfus A, Heuer C, Wilson P, Collins-Emerson J, Baker MG, Benschop J. Risk of infection and associated influenza-like disease among abattoir workers due to two Leptospira species. Epidemiol Infect 2014:1-11 doi: 10.1017/s0950268814002477[published Online First: Epub Date]|.
26. Adler B, de la Pena Moctezuma A. Leptospira and leptospirosis. Veterinary Microbiology 2010;**140**(3-4):287-96 doi: 10.1016/j.vetmic.2009.03.012[published Online First: Epub Date]|.
27. Bharti AR, Nally JE, Ricaldi JN, et al. Leptospirosis: a zoonotic disease of global importance. Lancet Infectious Diseases 2003;**3**(12):757-71
28. Thornley CN, Baker MG, Weinstein P, Maas EW. Changing epidemiology of human leptospirosis in New Zealand. Epidemiol. Infect. 2002;**128**(1):29-36 doi: 10.1017/s0950268801006392[published Online First: Epub Date]|.
29. Institute of Environmental Science and Research (ESR). Annual surveillance summary 2006-2010 Secondary Annual surveillance summary 2006-2010 2006-2010. https://surv.esr.cri.nz/surveillance/annual_surveillance.php.
30. Dreyfus A. Leptospirosis in humans and pastoral livestock in New Zealand. Massey University, 2013.
31. Marshall RB, Manktelow BW. Fifty years of leptospirosis research in New Zealand: a perspective. New Zealand Veterinary Journal 2002;**50**(3):61-63
32. Ayanegui-Alcerreca M, Wilson PR, Mackintosh CG, et al. Regional seroprevalence of leptospirosis on deer farms in New Zealand. New Zealand Veterinary Journal 2010;**58**(4):184-89 doi: 10.1080/00480169.2010.68863[published Online First: Epub Date]|.
33. Wilson P, Glossop JC, van der Kroef JW, Heuer C, Stringer L. Disease and deer farm productivity and profitability. Proceedings of the Deer Branch of the New Zealand Veterinary Association. Palmerston North, New Zealand: The Deer Branch New Zealand Veterinary Association, 2008:22-29.
34. Keenan B. Leptospirosis: reducing the impact on New Zealand workplaces. . Wellington, New Zealand: Department of Labour, 2007:55.
35. Faine S, Adler B, Bolin C, Perolat P. *Leptospira and Leptospirosis*. Melbourne, Australia: MediSci, 1999.
36. Pittavino M and, Dreyfus A, Heuer C, et al. Data in Brief: Data on Factors associated with the Incidence of Antibodies against *Leptospira interrogans* sv Pomona in Meat Workers in New Zealand, "submitted". Acta & Tropica 2017
37. R Development Core Team. *R: A Language and Environment for StatisticalComputing*. Vienna, Austria, 2015
38. Dohoo I, Wayne M, Stryhn H. Veterinary Epidemiologic Research. Veterinary Epidemiologic Research. Charlottetown, Canada: VER inc., 2010.
39. Hosmer DW, Lemeshow S. Assessing the fit of the model. Applied Logistic Regression Second ed. New York: John Wiley & Sons Inc, 2000:143-67.
40. Lewis F, Pittavino M, Furrer R. abn: Data Modelling with Additive Bayesian Networks [program], 2014.

41. Tierney L, Kadane JB. Accurate Approximations For Posterior Moments and Marginal Densities. Journal of the American Statistical Association 1986;**81**(393):82-86
42. Friedman N, Goldszmidt M, Wyner A. Data analysis with Bayesian networks: A Bootstrap approach, 1999:196-205.
43. Heckerman D, Geiger D, Chickering DM. Learning Bayesian Networks - The Combination of Knowledge And Statistical Data. Machine Learning 1995;**20(3)**(3):197-243
44. Koivisto M, Sood K. Exact Bayesian structure discovery in Bayesian networks. Journal of Machine Learning Research 2004;**5**:549-73
45. Babyak MA. *What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models*: Psychosomatic Medicine, 2004.
46. Plummer M. JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. Hornik, K., Leisch, F., Zeileis, A. (Eds.), Proc 3rd Int Work Dist Stat Comp (DSC 2003). Vienna, Austria, pp. 20–22. 2003
47. Plummer M, Best N, Cowles K, Vines K. CODA: Convergence Diagnosis and Output Analysis for MCMC. R News 2006;**6**(1):7-11
48. Cowles M, Carlin B. Markov chain Monte Carlo convergence diagnostics: A comparative review. Journal of the American Statistical Association 1996;**91**(434):883-904
49. Blackmore DK, Bell L, Schollum L. Leptospirosis in Meat Inspectors - Preliminary-Results of a Serological Survey. New Zealand Medical Journal 1979;**90**(648):415-18
50. Blackmore DK, Schollum L. The Occupational Hazards of Leptospirosis in the Meat Industry. New Zealand Medical Journal 1982;**95**(712):494-97
51. Benschop J, Heuer C, Jaros P, Collins-Emerson J, Midwinter A, Wilson P. Sero-prevalence of leptospirosis in workers at a New Zealand slaughterhouse. The New Zealand Medical Journal 2009;**122**(1307):39-47
52. Dreyfus A, Benschop J, Collins-Emerson J, Wilson P, Baker M, Heuer C. Sero-Prevalence and Risk Factors for Leptospirosis in Abattoir Workers in New Zealand. International Journal of Environmental Research and Public Health 2014;**11**(2):1756-75
53. Dreyfus A, Wilson P, Collins-Emerson J, Benschop J, Moore S, Heuer C. Risk factors for new infection with Leptospira in meat workers in New Zealand. Occupational and environmental medicine 2014 doi: 10.1136/oemed-2014-102457[published Online First: Epub Date]|.
54. Heuer C, Dreyfus A, Wilson PR, et al. Epidemiology and control of leptospirosis in New Zealand. In: Alban L, Kelly LA, eds. Society for Veterinary Epidemiology and Preventive Medicine. Proceedings, Nantes, France, 24-26 March, 2010, 2010:174-85.
55. Hand DJ, McConway KJ, Stanghellini E. Graphical models of applicants for credit. IMA Journal of Management Mathematics 1997;**8**(2):143-55
56. Buntine W. *Theory refinement on Bayesian networks. In Proceedings of Seventh Conference on Uncertainty in Artificial Intelligence, Los Angeles, CA, USA, pages 52-60*: Morgan Kaufmann, 1991.
57. Cooper GF, Herskovits E. A Bayesian Method For the Induction of Probabilistic Networks From Data. Machine Learning 1992;**9**(4):309-47
58. Friedman N, Koller D. Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. Machine Learning 2003;**50**(1-2):95-125
59. Chickering DM, Heckerman D, Meek C. Large-sample learning of Bayesian networks is NP-hard. Journal of Machine Learning Research 2004;**5**:1287-330
60. Parviainen P, Koivisto M, editors. Exact structure discovery in Bayesian networks with less space. Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence; 2009; Arlington, Virginia, United States. AUAI Press.
61. Dreyfus A, Benschop J, Collins-Emerson J, Wilson P, Moore S, Heuer C. Adjusting the leptospirosis sero-prevalence of New Zealand abattoir workers for sampling bias [presentation], 2010.
62. Shivakumar S, Krishnakumar B. Diagnosis of leptospirosis--role of MAT. The Journal of the Association of Physicians of India 2006;**54**:338-39

63. Goris. M. Diagnostic tests for human leptospirosis. Second meeting of the European Leptospirosis Society on Leptospirosis and other rodent borne hemorrhagic fevers. Royal Tropical Institute, Amsterdam, The Netherlands: European Leptospirosis Society 2015.
64. Fang F, Heuer C, Collins-Emerson J, Wilson P, Benschop J, editors. Effect of antibody titer cut points on inferences from a cross-sectional sero-prevalence study of leptospirosis in meat workers. Leptocon 2009, the 6th Annual Scientific Meeting of International Leptospirosis Society; 2009; Cochin, India.
65. Fang F, Heuer C, Collins-Emerson J, Wilson P, Benschop J, editors.Hathaway SC. Leptospirosis in New Zealand: an ecological view. New Zealand Veterinary Journal 1981;**29**(7):109-12 doi: 10.1080/00480169.1981.34815[published Online First: Epub Date]|.
66. McBride AJA, Santos BL, Queiroz A, et al. Evaluation of four whole-cell Leptospira-based serological tests for diagnosis of urban leptospirosis. Clinical and Vaccine Immunology 2007;**14**(9):1245-48 doi: 10.1128/cvi.00217-07[published Online First: Epub Date]|.

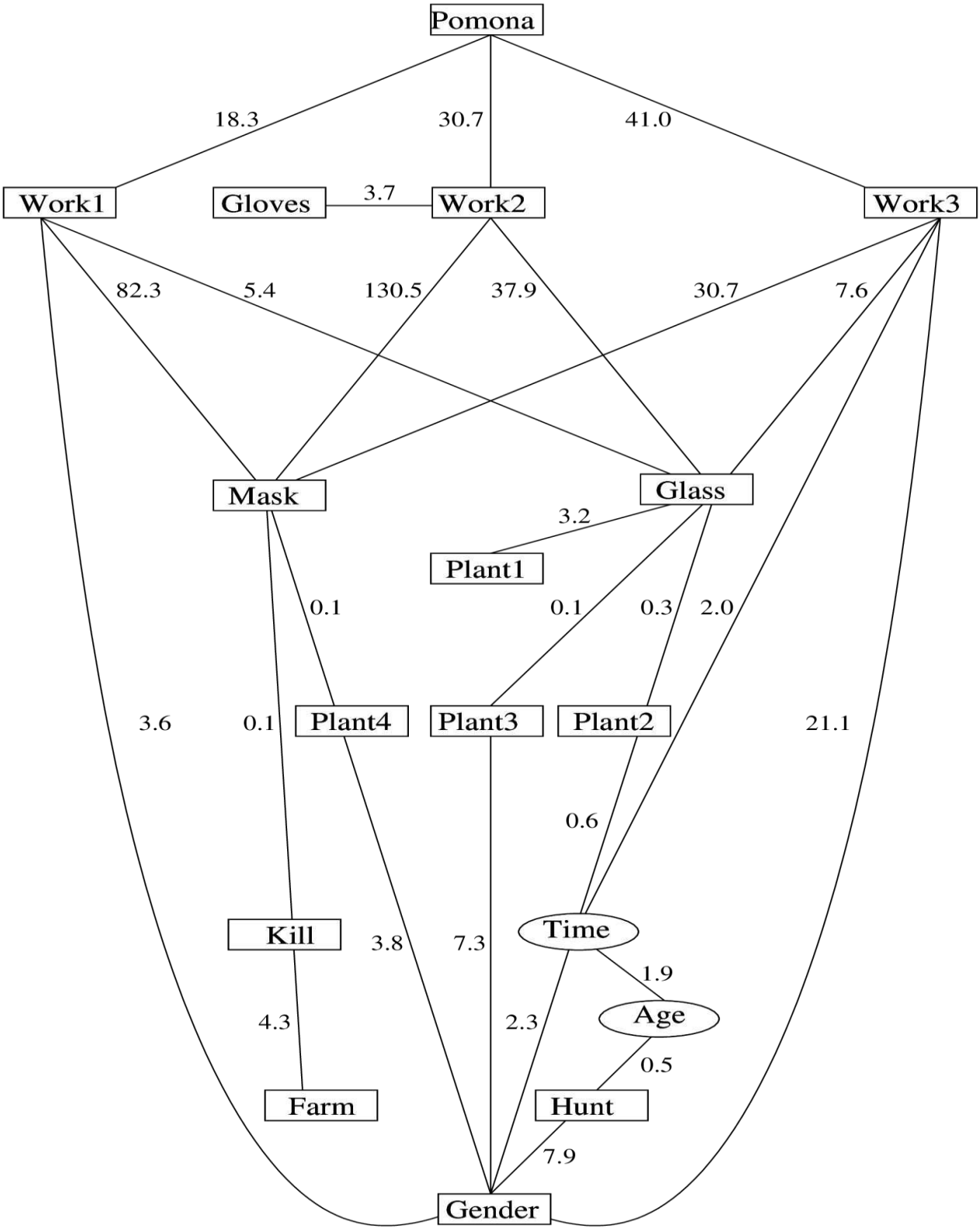

Figure 1: Final globally optimal additive Bayesian Network (ABN) model, after adjustment for over-fitting, evaluating factors linked with the odds of new infection with *Leptospira interrogans* sv Pomona ("Pomona") in sheep abattoir workers (n=384) in New Zealand. Directly dependent variables were various work positions ("Work1", "Work2", and "Work3"). Binary variables are shown as squares and continuous variables as ovals. Numbers represent odds ratios of significant directly dependent variables in ABN model, as reported in Table 2. Arc direction is omitted to not create confusion with the usual epidemiological DAGs, which imply causality and possible variables intervention, absent in ABN model where only statistical dependency are relevant.

Table 1: Frequencies of potential determinants of new infection with *Leptospira interrogans* sv Pomona in sheep abattoir workers (*n*=384) in New Zealand and odds ratios (OR), 95% confidence intervals (**95% CI**) and p-value from univariable GLM analysis.

Variables and Categories in GLM and ABN (Node label)	**% Workers (n)**	**% New infection (n)**	**OR**	**95% CI**	**P-value**
Work position 0 (Omitted[1])					
0 Not working in Boning, chillers, office					
1 Working in Boning, chillers, office	37 (142)	2.8 (1)	-	-	-
Work position 1 (Work1)					
0 Not working in offal removal, pet food					
1 Working in offal removal, pet food	11.5 (44)	9.1 (4)	14.1	(2.0-280.0)	0.019
Work position 2 (Work2)					
0 Not removing intestines or kidneys, not inspecting meat					
1 Intestines or kidney removal, meat inspection	22.9 (88)	14.8 (12)	22.3	(4.3-409.5)	0.003
Work position 3 (Work3)					
0 Not working in yards, not stunning or pelting					
1 Working in yards, stunning or pelting	28.7 (110)	23.6 (19)	29.4	(6.0-533.5)	0.001
Abattoir 1 (A1) (Omitted[1])					
0 Not working in Abattoir 1 (A1)[2]					
1 Working in Abattoir 1 (A1)[2]	35.2 (135)	16.7 (6)	-	-	-
Abattoir 1 (A2) (Plant1)					
0 Not working in Abattoir 1 (A2)[2]					
1 Working in Abattoir 1 (A2)[2]	21.4 (82)	22.2 (8)	2.3	(0.8- 7.3)	0.132
Abattoir 2 (Plant2)					
0 Not working in Abattoir 2					
1 Working in Abattoir 2	17.7 (68)	33.3 (12)	4.6	(1.7-13.8)	0.004
Abattoir 3 (Plant3)					
0 Not working in Abattoir 3					
1 Working in Abattoir 3	5.5 (21)	5.6 (2)	2.3	(0.3-10.7)	0.338
Abattoir 4 (Plant4)					
0 Not working in Abattoir 4					
1 Working in Abattoir 4	20.3 (78)	22.2 (8)	2.5	(0.8-7.7)	0.108
Gender (Gender)					
0 Female	33.3 (128)	27.8 (10)	**Ref**		
1 Male	66.7 (256)	72.2 (26)	1.3	(0.6-3.0)	0.459
Hunter of goats, pigs & or deer (Hunt)					
0 No	92.4 (355)	94.4 (34)	**Ref**		
1Yes	7.6 (29)	5.6 (2)	0.7	(0.1-2.5)	0.636
Slaughter of sheep, goats, pigs, beef & or deer at home (Kill)					
0 No	83.3 (320)	86.1 (31)	**Ref**		
1 Yes	16.7 (64)	13.9 (5)	0.8	(0.3-2.0)	0.639
Owning a farm with pigs, goats, sheep, beef cattle, alpaca & or deer (Farm)					
0 No	83.9 (322)	88.9 (32)	**Ref**		
1 Yes	16.1 (62)	11.1 (4)	0.6	(0.2-1.7)	0.392
Wearing normal or safety glasses (Glass)					
0 Sometimes/ never	43.2 (166)	27.8 (10)	**Ref**		
1 Always/ often	56.8 (218)	72.7 (26)	2.1	(1.0-4.7)	0.053
Wearing gloves on both hands (Gloves)					
0 Sometimes/ never	34.9 (134)	22.2 (8)	**Ref**		
1 Always/ often	65.1 (250)	77.8 (28)	2.0	(0.9-4.8)	0.099
Wearing a facemask (Mask)					
0 Sometimes/ never	83.3 (320)	80.6 (29)	**Ref**		
1 Always/ often	16.7 (64)	19.4 (7)	1.2	(0.5- 2.8)	0.639
Months worked in the meat industry (Time)					
Continuous	216.6 (9-636)[3]	146.6[3]	1.0	(0.9-1.0)	0.279
Age (Age)					
Continuous	48.1 (19-73)	11.9	1.0	(0.9-1.0)	0.794

[1]Omitted; [2]Abattoir 1 (A1) took part in the study in two consecutive years Abattoir 1 (A2), with 57 of initial 135 participants being resampled; [3]For continuous variables, mean, range and standard deviation are given.

Table 2: Odds ratios (OR) and confidence intervals (CI) of significant covariates and confounders in generalized linear modelling (GLM) (left) and DIRECTLY[1] dependent covariates in Additive Bayesian Network analysis (ABN) (right) for new infection with *Leptospira interrogans* sv Pomona in abattoir workers processing sheep (*n*=384) in New Zealand.

Covariates with categories (Node label)	**GLM[2]**		**ABN**	
	OR	**95% CI**	**OR**	**95% CI[3]**
Work position 1 (Work1)				
0 Not working in offal removal, pet food				
1 Working in offal removal, pet food	22.1	2.3-209.8	18.3	2.2-506.7
Work position 2 (Work2)				
0 Not removing intestines or kidneys, not inspecting meat				
1 Intestines or kidney removal, meat inspection	33.8	4.2-271.1	30.7	4.9-788.4
Work position 3 (Work3)				
0 Not working in yards, not stunning or pelting				
1 Working in yards, stunning or pelting	57.0	6.9-473.3	41.0	6.9-1044.2
Abattoir 2[2] (Plant2)				
0 Not working in abattoir 2				
1 working in abattoir 2	4.5	1.9-10.7		
Abattoir 4[2] (Plant4)				
0 Not working in abattoir 4				
1 working in abattoir 4	3.4	1.3-8.9		
Gender[2,4] (Gender)				
Female				
Male	0.5	0.2-1.4		

[1]One arc between variables; [2]not directly associated with Pomona in ABN; [3]ABN methodology does not generate p-values because of the joint mathematical formulation;[2]Statistically not significant, but kept in the model due to a confounding effect on the work position variable;