Use of Bayesian networks in predicting contamination of drinking water with *E. coli* in rural Vietnam.

David C. Hall[1], DVM, PhD and Quynh B. Le, MD, PhD
Dept. Ecosystem and Public Health, Faculty of Veterinary Medicine, University of Calgary, Calgary, Alberta, Canada T2N 4N1
1. dchall@ucalgary.ca +1(403)210-7589

Abstract
A Bayesian Belief Network (BBN) was designed to describe association between various deterministic and probabilistic variables gathered from 600 small scale integrated (SSI) farmers in Vietnam. The variables relate to E. coli content of their drinking water, sourced on-farm from wells and rain water. Sensitivity analysis of the model revealed that choice variables were particularly likely to influence endpoint values, reflecting the highly variable and impactful nature of preferences, attitudes, and beliefs relating to mitigation strategies. This BBN model of SSI farming in Vietnam is helpful to understanding the complexity of small scale agriculture as well as for identifying and estimating impact of policy options, particularly where combined with other analytical and policy tools. With appropriate knowledge translation, the model results will be particularly useful for helping SSI farmers understand their options for engaging in water public health mitigation strategies that do not disrupt their chosen livelihoods.

Keywords Bayesian networks, water public health, small scale farming,

1 Introduction

Vietnam is a country of 90.7 million people (World Bank, 2016), of whom close to 80% rely on small scale farming for some element of household income (Thanh, 2010). The VAC[1] integrated small scale integrated (SSI) farming system, a bio-intensive model that incorporates garden, animal, and fisheries production, has been promoted and adopted as an important part of Vietnamese rural livelihoods for centuries (Luu, 2001). Animal manure is used as fertilizer on crops and recycled in fish ponds where it acts as a feed source and promotes algal growth which also feeds the fish, animal and human waste is used in biogas production, and crops are used to supplement animal feeding.

According to the Government of Vietnam (MoC and MARD, 2000), more than 70% of small-scale integrated (SSI) farmers in Vietnam use water contaminated with waste of animal, human, or industrial origin. Many of these farms have no access to hygienic human waste management systems, and awareness of water quality and environmental sanitation is very limited. In contrast to the VAC system where livestock and human waste is recycled in fish ponds and rice fields, some non-sustainable farm management practices contribute to the issue of water contamination. Part of the problem is due to non-sustainable farm waste management practices such as dumping of animal waste directly into rivers or canals has increased the risk of bioaccumulation of hazardous materials in the wider environment, leading to increased risk of water related zoonotic disease (WRZD) transmission (Gilbert, *et al.*, 2008; Nguyen, *et al.*, 2011; Pfeiffer, *et al.*, 2009).

Water supply and water quality have a profound influence on public health (WHO, 2011). It is also estimated that water is the vehicle for about 80% of all WRZDs such as Salmonellosis and *E. coli* O157:H7 in people (Ford and Colwell, 1996). Although, water is considered the central of development in the Red

[1] The name VAC is derived from the Vietnamese words Vuon (garden or orchard), Ao (fish pond), and Chuong (animal housing).

River and Mekong River Deltas in Vietnam, these locations were also the major foci for the highly pathogenic avian influenza (HPAI) epidemic waves that took place between 2003 and 2005 (Gilbert *et al.*, 2008). HPAI viruses can survive for extended period in water, and the distribution of early HPAI outbreaks during the 2003-5 HPAI epidemic in Vietnam was thought to be associated with bodies of water on farms (Morris and Jackson, 2005).

With the participation of 600 farmers and more than six research and government institutes in North and South Vietnam during the summer months of 2013, we examined the quality of water on SSI farms as well as farmers' attitudes and preferences relating to prevention of WRZDs (Le, *et al.*, 2016a; Le, *et al.*, 2016b). One of our key objectives was to understand, based on risk factors, what was the likelihood of having unacceptable levels of coliforms in on-farm drinking water. Understanding of practices that contribute to risk of water contamination would help to identify those most requiring intervention to reduce likelihood of WRZD transmission. Part of the research addressed improved understanding of farmers' beliefs which influence their willingness to adopt and/or change such practices. This paper addresses development of a Bayesian Belief Network (BBN) (Pearl, 1988) using data from our earlier research cited above to model SSI farms in Vietnam in order to estimate likelihood of contamination with *E. coli* and related WRZD pathogenic agents. BBNs are described below. The benefits of such a model include use in identification of key risk factors for emerging infectious diseases including WRZDs, use in directing attention to modifying behaviour that contributes to those risk factors, and support for science based reasoning in developing policy options to support clean water and public health in rural areas of Vietnam.

2 Methodology

Data were gathered from a total of 600 SSI farms in the two river deltas of Vietnam: Thai Binh Province in the Red River Delta in the north of Vietnam and An Giang Province in the Mekong River Delta in the south of Vietnam. These provinces were selected because they fit well with the standard concept in Vietnam of VAC SSI farmers, because the extended research teams were familiar with the communities from previous studies, and because the communities had previous experience with control and prevention of HPAI outbreaks (Hoang, 2006; Rushton *et al.*, 2006). In each of the two provinces we selected 300 SSI farms from 10 communes to participate in the field data collection using a random selection approach. A two-stage cluster random sampling method was used to select target communes as clusters; farm households within each cluster were selected to participate in the study.

Using a cross-sectional approach, our methods included data gathering by (1) questionnaire which included both quantitative (*e.g.*, respondent's age, number of chickens) and qualitative variables (*e.g.*, farmers' subjective responses to questions relating to preferences of water management) as well as (2) qualitative methods such as in-depth interviews and focus groups. The latter helped to confirm or clarify data gathered by questionnaire, as well as engage farmers more fully in the research process. Data were gathered to populate variables describing 1) demographics, 2) livestock production, and 3) perceptions of risk factors for WRZD focusing on HPAI, diarrhea, coliform bacteria, and parasites. Examples of the latter include perceived threat of HPAI to health and wellbeing, and perceived ability to prevent WRZD by managing on-farm water storage.

Samples of both drinking and domestic[2] water sources were collected, stored, and transported in 200 ml units on-farm by trained technicians from national microbiological laboratories, using both the Vietnamese national technical regulations (MoH, 2009; TCVN, 2011) and World Health Organization (WHO, 2011) regulations for identifying factors and for testing of drinking and domestic water quality. Factors tested included pH, turbidity, and presence of *E. coli*. Water contamination with *E. coli* was assessed using the membrane filtration (MF) method, and the number of *E. coli* colony forming units (cfu's) in each 100ml of sample water was calculated using a standard formula[3] used with the MF technique.

Bayesian belief networks

A Bayesian belief network (BBN) is a directed acyclic graph (a particular sort of influence diagram) describing likely cause-and-effect relationships between various factors or nodes (Korb and Nicholson, 2011; Shachter, 1986). Recently they have been used to model aquatic species management (Herring, Stinson, and Landis, 2015), health risk of noncarcinogenic substances (Liu, *et al.*, 2012), and minimizing the risks from spilled oil (Carriger and Barron, 2011). This is the first application of which we are aware of a BBN to examine the risk of *E. coli* contamination in rural drinking water sources. In a BBN, originating or parent nodes are connected by arrows to receiving or child nodes, with the implication that the factor represented by the parent node has direct influence on the outcome (and possibly a factor in the next step) represented by the child node. This directional description of a probabilistic network allows consideration of the relationship between factors and consequences, identifying important variables and pathways to outcomes. This can be particularly helpful to policy makers when considering options for impactful policies that target desirable results through efficient means. Thus, following the philosophy of Bayes' theorem, previous decisions influence future outcomes.

Methods for discretizing variables in BBNs are contentious. There are no particularly correct or widely accepted approaches, although some approaches such as the use of a Likert scale have been widely adopted. When developing BBNs, despite justification of approach in discretization, the number of discrete classes as well as the absolute value of each class can have impact on the accuracy of the model. Larger, complicated models with greater numbers of classes per discretized variable (*i.e.*, a more finely defined model) tends to result in improved precision (Uusitalo, 2007), but the trade-off is there is also an increasing need for data to sufficiently populate the variables for model resolution. Numbers of levels used in discretized variables commonly range from two to ten.

Modeling Vietnamese Small Scale Integrated Farms

We used Netica[4] (version 5.22) to build BBNs modeling the interactions of various factors on small scale farms in Vietnam. The purpose was to examine the pathways, influence, and impact of various factors on

[2] Domestic water is defined by the Government of Vietnam as "water used for domestic purposes but not for direct drinking or processing food" (MoH, 2009), such as washing dishes or bathing.

[3] C (cfu's in V ml) =(A ×N)/B where: C = number of *E. coli* cfu's confirmed in 100ml of sample water; V = volume of filtered sample water; A = number of presumptive *E. coli* cfu's positive with Indole test; N = number of presumptive *E. coli* cfu's on filtered incubated membranes with Lactose TTC agar medium; B = number of presumptive *E. coli* cfu's transferred for further incubation in Trypone Soy Agar medium.

[4] Version 5.22 (Norsys, 2016).

the level of *E. coli* found in drinking water sourced from wells and rain water and housed in storage units on-farm. All of the data used to inform the model came from the dataset described above, gathered from 600 SSI farms in Vietnam. Factors can roughly be categorized into four groupings: demographics (*e.g.*, years of formal education, age); animals raised on farm (*e.g.*, pigs, fish); mitigation strategies; and beliefs and preferences (*e.g.*, covering wells impacts is important for keeping drinking water clean). These factors and their values based on the database are summarized in Tables 1 and 2. Some continuous variables in the dataset (*e.g.*, age, pigs) were discretized before using in the BBN; others (*e.g.*, sex, respondent's self-evaluation of livestock management skills) were gathered as discrete variables in the field study. Note that Netica calculates mean values based on the value of the end of a range within a discretized variable, and thus the mean reported in Netica diagrams (reported in Figure 1) may not be the same as that calculated using standard methods. The latter are reported in Table 1.

Not all possible decision making factors are included in the BBN. The scenario represented in Figure 1 captures key factors we have identified for the purpose of understanding better the structure of SSI farms in Vietnam. We were particularly interested in the influence that presence of on-farm livestock can have on the level of *E. coli* in drinking water, and how levels of pathogens in drinking water might be mitigated. Thus we included two key outcomes as the final nodes in Figure 1: level of *E. coli* in drinking water found in wells and level of *E. coli* in drinking water found in storage tanks collecting rain water.

3 Results and Discussion

Our BBN is illustrated in Figure 1. Using field data, it was possible to set up the BBN assuming the demographics, livestock ownership, and stated preferences and attitudes provided by the 600 SSI farmers in Thai Binh and An Giang, Vietnam (Table 1 and 2). Using this approach allowed us to incorporate case file data to generate the BBN and subsequently conduct sensitivity analysis of the two key outcome variables, level of *E. coli* in drinking water sourced from rain and wells. The sensitivity analysis is reported below and in Table 3. The BBN also allowed scoring of accuracy of future predicted cases or scenarios given data on the same variables. This is not reported here for brevity.

Table 3 reports the variance reduction of the two nodes that accounted for *E. coli* in drinking water from wells and rain on SSI farms (in the BBN in Figure 1, these are labelled "DrinkWtrWellEColi" and "DrinkWtrRainEColi" respectively). The variance reduction reported in Table 3 can be interpreted as the expected reduction in variance of the *E. coli* nodes (or variables based on the dataset) due to a single finding in other nodes. The five nodes variables that impact most on variance reduction are listed in the table for each *E. coli* node. For *E. coli* in drinking water from wells, the variance of *E. coli* was most reduced by single observations at the nodes capturing farmers' stated response to having a mitigation strategy relating to livestock health ("MitigStratLvstkMgmt"), stated belief that covering stored drinking water has an impact on water quality ("ImpCvrWtrSrc"), and income. Of the top five most impactful nodes, one is quantifiably verifiable (Income) while four are stated beliefs or preferences. Similarly, for drinking water sourced from rain water, three of the most impactful nodes are based on respondents' beliefs or preferences, while income and cattle are quantifiably verifiable. Three nodes are similarly identified for the two *E. coli* variables but having different impact on variance.

Building a network from a selection of integrated components that make up a much larger integrated system of elements allows investigation at a more focused level, identifying particularly sensitive variables that might require further data gathering in order to improve model accuracy, and investigation of

potential impact from changes to specific values of variables. This BBN was built using known values for all nodes based on data from 600 SSI farms. We had expected a fairly robust network that did not show high sensitivity to single node changes, but our sensitivity analysis revealed some particularly large reductions in variance were possible of the nodes capturing probability of *E. coli* in water storage units for drinking water from wells and rain. It is noteworthy that of the top five nodes to which each of the *E. coli* count nodes were most sensitive, most were of a non-quantifiably verifiable nature; that is, they captured expressed beliefs, preferences, and attitudes. Preferences for choice variables such as these can be difficult to elicit while quantifiable variables such as age and number of chickens are less prone to inconsistencies across respondents. Thus it is not surprising to see the greater impact on variance of the choice variables.

A second feature of our BBN that deserves critical attention is the selection of our nodes. With more than 150 variables collected from our study, we chose variables that were a good mix of deterministic and probabilistic variables as well as quantifiably verifiable and choice type variables. Furthermore, we deliberately selected variables that we felt reflected best the main attributes and preferences of the communities with which we participated that were of relevance to on-farm mitigation of waterborne zoonotic diseases, based on discussions during forums and on-farm visits. The particular selection of nodes in the BBN heavily influences the robustness of the endpoint nodes. Further modeling of various combinations is warranted with examination of impact on variance and predictive accuracy of results. This will include identification of lower and higher impact scenarios and impact analysis of potential rural water policy implementation directed at improving public health.

5 Conclusions

We were able to build a BBN describing a cause and effect relationship between various deterministic and probabilistic variables gathered from 600 small scale integrated farmers in Vietnam on *E. coli* content of their drinking water which was sourced on-farm from wells and rain water. Sensitivity analysis of the model revealed that choice variables were particularly likely to influence endpoint values, reflecting the highly variable and impactful nature of preferences, attitudes, and beliefs relating to mitigation strategies. This BBN model of SSI farming in Vietnam and the influence that not just subjective choice variables but also deterministic factors including on-farm livestock, income, education, and years of farming can have on the likelihood of presence of *E. coli* in on-farm drinking water is a valuable addition to understanding the complexity of small scale agriculture as well as for identifying and estimating impact of policy options. Attention to rural drinking water management policies in Vietnam, particularly with respect to public health outcomes in agricultural communities, has been limited in scope. With the use of this BBN and other policy tools we will continue to examine those and related issues further, with particular attention to the role policy can play in helping SSI farmers understand options for engaging in water public health mitigation strategies that do not disrupt their chosen livelihoods.

Table 1. Demographic and coliform content variables used in a Bayesian Belief Network of *E. coli* content of drinking water on 600 small-scale integrated farms in Vietnam.

Factor	**Thai Binh Province**			**An Giang Province**		
	Sample mean (s.d.)		*n*	*Sample mean (s.d.)*		*n*
Age (years)	47.5	(11.9)	299	44.4	(10.5)	298
Years farming	11.7	(8.7)	300	7.6	(7.2)	298
Income (Vietnam dong[5], millions)	15.96	(12.11)	294	10.4	(9.12)	296
Education (years)	8.2	(2.7)	298	5.6	(3.2)	298
Chickens (n)	35.2	(36.0)	273	9.3	(18.9)	298
Ducks (n)	187.4	(210.4)	186	26.4	(215.1)	298
Pigs (n)	30.3	(34.2)	203	2.6	(6.7)	298
Cattle (n)	0.4	(2.4)	300	2.7	(13.4)	298
Fish (kg/year)	1,773.3	(2,391.2)	241	430.6	(1,109.1)	51
E. coli in well water (cfu)	24.0	(39.2)	150	27.3	(63.1)	6
E. coli in rain water (cfu)	61.8	(82.6)	106	11.1	(20.9)	44

[5] At the time of data gathering (July and August 2013), one US dollar was equivalent to 21,000 Vietnamese dong.

Table 2. Discrete variables used in Bayesian Belief Network of *E. coli* content of drinking water on 600 small-scale integrated farms in Vietnam.

Dummies (short name[6])	Brief explanation	Categories
RespLvstk	Sex; main manager[7] of livestock	Male; female
RespFamHlth	Sex; main manager for family health matters	Male; female; both
GndrDecMd	Sex; decision maker[6] for livestock (Dummy variable)	0 Male; 1 Female
GndrDecFd	Sex; decision maker for family health (Dummy variable)	0 Male; 1 Female
MitigStratLvstkMgmt	Has a mitigation strategy relating to livestock health	Yes; No
MitigStratSrcMgmt	Has a mitigation strategy for protecting water quality at source	Yes; No
MitigStratWtrStor	Has a mitigation strategy for protecting water quality at storage	Yes; No
ImpCvrWtrSrc	Impact on water quality of covering stored drinking water	0 None; 1 Very low; 2 Low; 3 Medium; 4 High; 5 Very high; 6 Extremely high
SuscHPAIDrkWtr	Susceptibility to avian influenza from drinking water	1 High; 2 Moderate;
SelfEvalLvstkMgmt	Self-evaluation of livestock management skills	1 Poor; 2 Below average; 3 Average; 4 Above average; 5 Outstanding
SelfEvalWtrStor	Self-evaluation of water storage management skills	1 Poor; 2 Below average; 3 Average; 4 Above average; 5 Outstanding
GndrDecMd	Sex of decision maker for livestock (Dummy variable)	0 Male; 1 Female
GndrDecFd	Sex of decision maker for family health (Dummy variable)	0 Male; 1 Female

[6] Short names are from figure 1.

[7] In the study questionnaire, a manager carries out the directives of a decision maker; they are not necessarily the same person, hence two variables were used to capture management *vs.* decision making.

Figure 1. Bayesian Belief Network of risk of *E. coli* contamination of drinking water on 600 small-scale integrated farms in Vietnam.

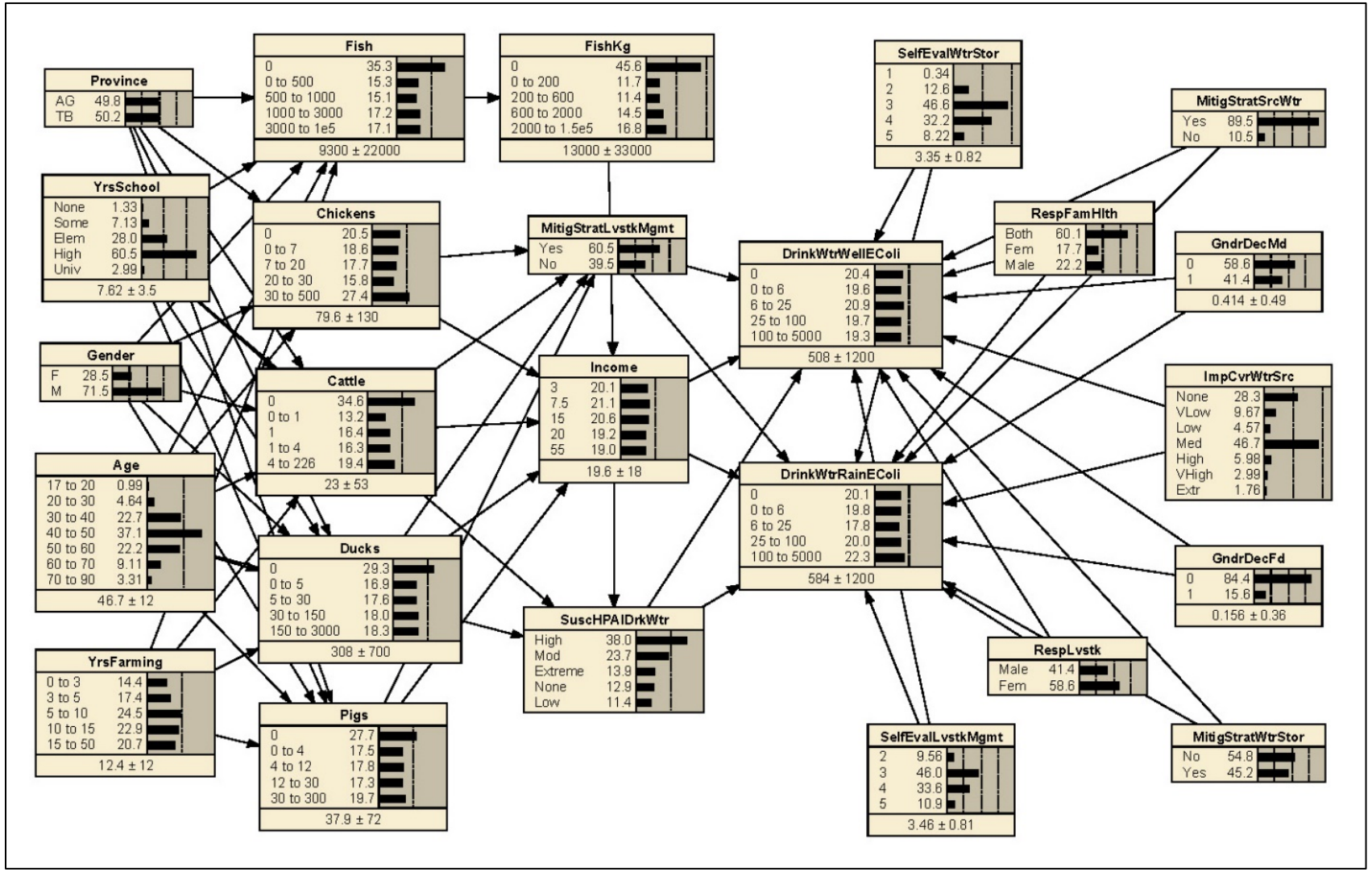

Table 3. Sensitivity of nodes capturing *E. coli* content of drinking water sourced from wells and rain water in a Bayesian Belief Network using data from 600 small scale integrated farms in Vietnam.

'DrinkWtrWellEColi'		
Node	*Variance reduction*	*Percent*
MitigStratLvstkMgmt	343.2	0.0248
ImpCvrWtrSrc	265.4	0.0192
Income	248.9	0.0180
SuscHPAIDrkWtr	163.0	0.0118
RespFamHlth	96.9	0.0070
'DrinkWtrRainEColi'		
Income	2560	0.1650
ImpCvrWtrSrc	1768	0.1140
SuscHPAIDrkWtr	1195	0.0768
MitigStratLvstkMgmt	1048	0.0674
Cattle	47.8	0.0031

References

Carriger J.F. and M.G. Barron. (2011) Minimizing risks from spilled oil to ecosystem services using influence diagrams: The Deepwater Horizon spill response. *Environ. Sci. Technol.* 45:7631–7639.

Ford, T.E. and Colwell, R. (1996) A Global Decline in Microbiological Safety of Water: A Call for Action. American Academy of Microbiology.

Gilbert, M., X. Xiao, D.U. Pfeiffer, M. Epprecht, S. Boles, and C. Czarnecki. (2008) Mapping H5N1 highly pathogenic avian influenza risk in Southeast Asia. Proceedings of the National Academy of Sciences. 105:4769-4774.

Herring, C.E., J. Stinson, and W.G. Landis. (2015) Evaluating nonindigenous species management in a Bayesian networks derived relative risk framework for Padilla Bay, WA, USA. *Integrated Environmental Assessment and Management*. 11(4):640-652.

Hoang, T.D. (2006) HPAI epidemic and lessons learned in Viet Nam. Asia-Pacific Economic Cooperation.

Korb, K.B. and A.E. Nicholson. (2011) Bayesian Artificial Intelligence, Second Edition. CRC Press, London.

Le, Q.B., S.C. Cork, and D.C. Hall. (2016) Microbial and related quality indicators of domestic water on small-scale integrated farms in Vietnam. *Under review.*

Le, Q.B., D.C. Hall, and S.C. Cork. (2016) Microbial and related quality indicators of drinking water on small-scale integrated farms in Vietnam. *Under review.*

Liu, K.F., C. Lu, C. Chen, and Y. Shen. (2012) Applying Bayesian belief networks to health risk assessment. *Stochastic Environmental Research and Risk Assessment*. 26:451-465.

Luu, L.T. (2001) The VAC system in Northern Viet Nam. In: Integrated agriculture-aquaculture. FAO Fisheries Technical Paper 407. FAO/IRRI/World Fish Centre.

MOC and MARD. (2000) National Rural Clean Water Supply and Sanitation Strategy Up to the Year 2020. Hanoi, Vietnam. Ministry of Construction (MOC) and Ministry of Agriculture and Rural Development (MARD) of the Government of Vietnam.

MoH (2009). National technical regulation on domestic water quality (QCVN 01 and 02: 2009/BYT). Hanoi: Department of Preventive Medicine & Environment, Ministry of Health (MoH), Socialist Republic of Vietnam.

Morris, R.S. and Jackson, R. (2005) Epidemiology of H5N1 Avian Influenza in Asia and Implications for Regional Control. FAO-UN. Rome.

Nguyen, H.V., T.A. Vuong, D.P. Pham, and V.T. Vu. (2011) Safe use of wastewater in agriculture and aquaculture (Rep. No2). The National Centre of Competence in Research (NCCR) North-South.

Norsys. (2016) Netica version 5.22. Norsys Software Corp. Vancouver, Canada.

Pearl, J. (1988) Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kauffmann, San Francisco.

Pfeiffer, J., R.M. Pantin, T.L. To, T. Nguyen, and D.L. Suarez (2009) Phylogenetic and biological characterization of highly pathogenic H5N1 avian influenza viruses (Vietnam 2005) in chickens and ducks. *Virus Research*. 142:108-120.

Shachter, R.D. (1986) Evaluating influence diagrams. *Operations Research*. 34(6):871-882.

Rushton, J., A. McLeod, and J. Lubroth. (2006) Managing transboundary animal disease. In: Livestock Report, 2006. FAO-UN. Rome.

TCVN. (2011) TCVN 6663 2011 - ISO 5667-1:2006: Water quality - Sampling - Part 1: Guidance on the design of sampling programmes and sampling techniques. Directorate for Standards, Meteorology and Quality, Ministry of Science and Technology of the Socialist Republic of Vietnam [On-line]. Available: http://www.tcvn.vn/sites/head/en/tim-kiem-tieu-chuan.aspx?sk=&cat=tcvn

Thanh, P.V. (2010) VAC integrated system with entire energy chain in Vietnam. The Center for Rural Communities (CCRD), Hanoi.

Uusitalo, L. (2007) Advantages and challenges of Bayesian networks in environmental modelling. *Ecological Modelling.* 203(3-4):312-318.

WHO. (2011) Guidelines for drinking-water quality. (Fourth ed.) World Health Organisation.

World Bank. (2016) Accessed online February 14, 2016. http://data.worldbank.org/country/vietnam