Quantitative Risk Assessment from Farm to Fork and Beyond: a global Bayesian approach concerning food-borne diseases

Abstract

A novel approach to the quantitative assessment of food-borne risks is proposed. The basic idea is to use Bayesian techniques in two distinct steps: first by constructing a stochastic *core model* via a Bayesian network based on expert knowledge, and secondly using the *data* available to improve this knowledge. Unlike the *Monte Carlo simulation* approach as commonly used in quantitative assessment of food-borne risks where data sets are used independently in each module, our consistent procedure incorporates information conveyed by data throughout the chain. It allows "back calculation" in the food chain model, together with the use of data obtained "dowstream" in the food chain. Moreover the expert knowledge is introduced more simply and consistently than with classical statistical methods. Other advantages of this approach include the clear framework of an iterative learning process, considerable flexibility enabling the use of heterogeneous data, and a fully justified method to explore the effects of variability and uncertainty.

As an illustration, we present an estimation of the probability of contracting a campylobacteriosis as a result of broiler contamination, from the standpoint of quantitative risk assessment. Although the model thus constructed is oversimplified, it clarifies the principles and properties of the method proposed, which demonstrates its ability to deal with quite complex situations and provides a useful basis for further discussions with different experts in the food chain.

Key words: Risk assessment, Bayesian network, Bayesian statistics, campylobacteriosis, broiler, food-borne disease.

1 Introduction

Quantitative risk assessment (QRA) regarding food-borne pathogens is growing in importance, for both public health and trade purposes. International organizations such as the WHO and FAO, and various national institutions, carry out such studies concerning the principal hazards. The Codex Alimentarius has established certain methodological rules which have standardized (or at least included in an identical framework) these different studies.[1] A QRA model for microbial hazard in food is a model made up of modules each of which is itself a model that needs to be populated with values. It is a modular approach called Process Risk Modeling[2] (PRM) or Modular Process Risk Modeling (MPRM)[3,4] which splits the farm-to-fork chain into modules that logically and sequentially progress in a similar order to that of the food system. The starting points for the model and specific modules into which the system is classified, needs to be determined by the assessor by means of data or/and expert knowledge. In order to estimate the risk, there is basically one or two parameters to consider: the hazard prevalence or/and the level of contamination. This approach has the advantage of assessing the impact of the various elements in the pathway in terms of their contribution to the overall risk to human health. This explained why most models start by modeling confined modules of the chain before attempting to link them. As a result, only part of the chain is modeled in most cases, and the final output may be the probability of eating a contaminated meal.

We explore the problem starting with a global modeling. Contrary to most studies on QRA where specific modules of the food chain are modeled in detail, our aim is to tackle the problem in reverse order, *i.e.* starting with a simplified model which would

encompass the entire chain, and could then be improved gradually at certain points while remaining consistent with the global model. The hope is that mutual consolidations of the modules can be obtained by a complete construction. The Bayesian framework allows "back calculation" together with the use of data obtained "dowstream" in the food chain model. Information could spread through the different parts of the model.

Bayesian approaches in the context of Monte Carlo simulation and risk analyses have been considered in particular by Patwardan and Small[5] and Brand and Small.[6] In the first paper the authors make a link between Monte Carlo simulations and the Bayesian approach (using also Monte Carlo simulations) and in the second paper they study more precisely how the Bayesian approach can take into account the difference between variability and uncertainty, through the estimation of different indices. In this paper, our aim is slightly different since we propose a general procedure for modeling global and complex phenomena in food risk assessment and for estimating such models using a Bayesian approach. This can be used in different contexts, such as prediction or evaluation of the risk associated with different behaviors.

We illustrate the proposed approach with the assessment of campylobacteriosis in France caused by home consumption of chicken meat. As emphasized in previous published reports on campylobacter,[7–9] this task is not simple for several reasons: the global chain is highly complex (numerous important factors must be taken into account), large areas are quantitatively almost unknown (for example, campylobacter concentrations at the point of consumption) and pertinent data are frequently absent (for example, large samples of campylobacter enumeration data on the birds at the entrance and at the end of the industrial processing). Although the model we propose for this assessment is quite crude and needs to be refined in many ways, we believe that it is well suited as an illus-

tration of a Bayesian approach for a quantitative risk assessment study. This hazard is a major food-borne pathogen and receives much attention in QRA. We establish here the basis of a global Bayesian modeling of this risk. This example should receive a thorough development to present the French situation concerning this hazard but it corresponds to this context of scarce data, where a Bayesian approach is typically quite powerful, since it enables the combination of data and other types of knowledge of very different nature.

The proposed methodology is built on the definition of a core model using Bayesian network and then on the integration of available data by a Bayesian statistical approach.

2 Outline of the proposal

The proposed methodology comprises two main successive steps, each being composed of substeps. Step **M** defines a stochastic model of the phenomenon under study by constructing a Bayesian network. Step **D** improves the model with the help of available data obtained through Bayesian statistics. In this paper we consider as data, all available data which is treated as such, and which has not been summarized in some way. Other types of data, such as those coming from expert knowledge or summary obtained in the literature are used to construct informative priors and are not denoted here as *data*.

We propose here a consistent formalism whose theoretical properties are demonstrated to construct a QRA model. The construction of the model and the risk estimation are based on the Bayes' rule:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A|B)P(B)}{P(A)} \propto P(A|B)P(B), \tag{1}$$

where A and B are random variates. From (1), we can fully specify the joint distribution

of the model using conditional independencies between some model parameters (Bayesian network's principle). Then the parameters of interest are estimated from their posterior distributions using again Bayes' rule. Let $Data$ designates the data (see above) and θ any unknown or unobserved quantity (typically parameters):

$$P(\theta|Data) \propto P(Data|\theta)P(\theta),$$

where $P(\theta|Data)$, $P(Data|\theta)$ and $P(\theta)$ are respectively the posterior distribution of the parameter θ, the likelihood function depending on the $Data$ and on the parameter and the prior distribution of θ.

In this section, we list all these steps which are then illustrated in the case of campylobacteriosis originating from broiler chickens in France.

2.1 Step M: core model through a Bayesian network

In a first step, data are not introduced. The aim is to model current (or prior) knowledge stochastically using only expert opinions and relevant scientific literature. The result is a predictive tool.

M1 Make the inventory of the variates of interest (vi) to characterize the system under study. These are the variates in which experts are spontaneously interested. There is no need for them to be observable. Because of uncertainty and/or variability, most of them will be considered as random variates. The distinction between these two notions can be introduced as described in Pouillot *et al.*.[10]

M2 If necessary, link the variates by means of conditional dependencies, constructing a directed acyclic graph.[11] The conditional distributions given the parent variates

must be expressed in the simplest possible manner. Most often, to achieve this, it is necessary to introduce some extra nodes in the graph: either complementary variates (cv) or ancestor variates (without parents) that will be assimilated as parameters (pa).

M3 Specify the marginal distributions of the parameters, and the conditional distributions of the other variates.

At this point, the core model is defined. It is constructed so that the theory ensures that the full joint probability distribution of all defined variates is given simply by the product of marginal and conditional distributions.[12] Note however that it is necessary to check whether the global behavior of the model is consistent with the expert knowledge.

M4 Run the model to check that all variates behave sensibly. This must be done first of all by generating the joint distribution of all variates using Monte Carlo methods and then plotting the marginal and bivariate distributions of variates or pairs of variates of interest. It is also advisable to produce some conditional distributions, corresponding to specific cases well known by the experts. If the results are not convincing, carefully reconsider steps **M1**, **M2** and **M3** until agreement is obtained. It may be that, faced with an inconsistency, the expert will revise his/her initial opinions. This is a validation step based on the global behavior of the locally constructed system.

This process produces a tool similar to that produced using a *Monte Carlo simulation* approach. As the data are not yet incorporated, the above quantitative assessment is based on available prior knowledge only. From the perspective of Bayesian statistical analysis, it represents the prior distribution. When too little is known about some phenomenon,

improper *priors* can be used if the *posterior* distribution is proper. In such a case M4 is no longer possible (an example of such a possibility is given in the farm and broiler modules in the campylobacter example (see section 3.1)). Another option may be to use some relevant data to obtain conditional proper distributions at this level, see also section 3.1.

If one does not want to follow the Bayesian statistics paradigm, it is possible to decide at step **M3** not to give marginal distributions to the parameters and adopt a classical approach to infer the parameters from available data, for instance by maximizing the likelihood. The drawbacks to such a decision are not only the non-availability of software to perform the analysis when the model is very complex but also the probable occurrence of non-identifiabilities, from which the Bayesian statistical approach is protected by the use of prior distributions. In addition, a classical approach does not provide a distribution on unknown parameters, only a point estimate unless one refers to asymptotic theory to construct asymptotical confidence intervals but in such complicated models the asymptotic is not relevant (small number of data compared to the number of parameters). We therefore follow a Bayesian approach which enables us in particular to construct coherent credible regions together with point estimates, using sampling algorithms such as Markov Chain Monte Carlo (MCMC) algorithms. As noted previously, such approaches are particularly helpful in complex situations and in situations where data are sparse.

2.2 Step D: Incorporating data within a Bayesian paradigm

This second step is devoted to extracting from available data as much relevant information as possible to improve the core model, following a standard Bayesian approach, as has been described, for instance, by Robert,[13] and Gelman *et al.*,[14], and Gilks *et al.*.[15]

D1 When relevant data are selected, they are represented by new random variates (designated by da) whose distributions depend on the core model variates. It is often necessary to introduce more complementary variates (cv), and even new parameters requiring new prior distributions. This does not modify the full joint prior distribution of the core model. Let us designate the model obtained as the *augmented model*. In fact, in the setting of the Bayesian statistical approach, this substep corresponds to the definition of the likelihood, the parameter of interest arising from the Bayesian network specification, possibly with some additional nuisance parameters (pa).

D2 As is usual, the posterior distribution of non observed quantities (either variates or parameters) is calculated conditionally on the observed data. We can then focus on the posterior distribution of the quantities of interest by marginalizing over the other variates.

D3 A round similar to that of **M4** is then necessary with the posterior distribution of the core model. Again, if no agreement is obtained, the process must be revised starting from **M1**.

2.3 Computation

Numerical computations can be performed easily using Jags[16] and WinBUGS[17] softwares, both of which used similar codes to describe the model. We found Jags to be much better for almost degenerate distributions (such as Beta(0.024,0.011) used in the campylobacter's example below), although mixing was better under WinBUGS, certainly because it comprised a broader range of samplers, including Metropolis-Hastings sampling. We finally ran the calculations using WinBUGS, with a $[10^{-5}, 1-10^{-6}]$ truncation on the

degenerate Beta distribution. The code necessary to run our example on campylobacteriosis is shown in Appendix §A, along with some of the data sets. Simulated data produced by WinBUGS were processed using the CODA[18] statistical **R**[19] software package.

It should be noted that MCMC simulations could also be performed directly in **R** or in other languages, requiring that the practitioner actually writes the MCMC code.

3 Campylobacteriosis and broiler chickens

Although massive outbreaks can arise from the consumption of contaminated water,[20] campylobacteriosis is not, in most cases, a dramatic disease; it is rarely reported in the media and the illness usually resolves within a few days without any after-effects. Some serious complications such as Guillain-Barre syndrome may nevertheless occur. However, it is a subject of increasing concern in industrialized countries, and is now deemed as a major public health problem[21] with considerable socioeconomic effects, *e.g.* the number of work days lost. It is accepted that one of the principal sources of human contamination is the consumption of poultry, either when the meat is undercooked or most often as a result of cross-contamination between ready-to-eat food and raw chicken meat.[22,23] Major studies by experts in different fields and countries have addressed the problem of *Campylobacter* spp. and broiler chicken.[7–9] The model, presented herewith is the continuation of the expert review carried out at the French Agency for Food Safety (AFSSA),[24] the main aim of which is to determine the incidence of campylobacteriosis in the French population, with particular emphasis on cases caused by the home consumption of chicken meat. Because of their microaerophilic nature, campylobacters, unlike other pathogenic food borne bacteria, do not proliferate on food; this characteristic of campylobacter simplifies the modeling process.

3.1 Defining the core model (step M)

The first objective of our modeling is to construct a QRA model leading to model all the variates of interest listed in Table I with their variability and uncertainty. It is essential to see how the variability and uncertainty of the food system propagates in the model to propose new modeling or control measures. We want to see how these variates interact to better understand the system and to possibly assess the impact of various elements in the pathway. Also we want to see how the information produced by the data propagates into the model to propose new modeling or in some cases to consider new data which would be of importance in the global modeling.

The entire food chain describing the different steps between production and illness can be broken down into modules corresponding to well-identified substeps. These are: "chicken farm" ending at the production of chicken flocks, "broiler production" ending with the production of chicken carcasses for the consumer,"consumption" ending with the numbers of broilers consumed by households in the specified population, "hygiene" ending with the probability of exposure if a contaminated broiler enters the household,"exposure" during a defined period and finally the"illness" module, ending with the targeted hazards. For each of the modules, one or two variates of interest (vi) were defined: they are presented in Table I.

The joint probabilistic distribution of these seven variates of interest was specified by successive conditional distributions according to the influence graph depicted in Figure 1; this implied the definition of complementary variates (cv) and parameters (pa). All these distributions are presented in Table II.

For shortness sake, we do not provide all details on the determinants of the core model defined above, but the principal points are outlined:

Chicken farm and broiler production modules:

- In the chicken farm and broiler production modules, a hierarchical model has been considered to take into account the overdispersion or variability of p_f, the probability of a farm chicken being contaminated, and p_b, the probability of a carcass being contaminated, between the farms and the slaughterhouses around the country. More precisely, we set $\text{logit}(p_f)$ and $\text{logit}(p_b)$ as normal random variables centered at m_f and m_b respectively, with the constraint $m_f \leq m_b$. The constraint in mean was deduced from the experts' opinion that on average a carcass from a contaminated flock was contaminated and that cross-contamination between carcasses from different flocks could occur during slaughterhouse processing. Parameters were tuned to obtain the magnitudes accepted by experts; that is a prior belief around 0.5 for p_f and p_b (see Figure 3). We have also considered an improper prior on (m_f, m_b, s_f, s_b) where the latter two represents the standard errors of the Normal random variables, in the form:

$$\pi(m_f, m_b, s_f, s_b) \propto \mathbf{1}_{m_f \leq m_b} \frac{1}{s_f.s_b}.$$

  This improper prior is often considered as a noninformative prior (apart from the constraint on m_f and m_b). The posterior distribution of the contamination module under this prior is proper, given the available data on the contamination of flocks. We cannot in this case produce the M4 step since the prior model does not follow a proper distribution, but we can still produce a modified M4 step, taking into account the data on flock and slaughterhouse contaminations.

  Hygiene module:

- Although very little is known about what happens in household kitchens, we believe

it is important to distinguish between the cross-contamination process (p_{hc}) and the poor hygiene effect (p_{hh}), in particular since some more data could be obtained on either or both of these aspects. The prior on p_{hc} was assessed using the two cross-contamination models described in the FAO/WHO report[8] and discussed by Luber *et al.*[22] which lead to a probability of transfer around 1/3 and 2/3 that we translated into a beta(8,8) as a prior on p_{hc} (mean= 0.5, CI(95%)=[0.27,0.73]); the prior p_{hh} was assessed from Yang *et al.*[25] and the surveys on consumer habits mentioned by Christensen *et al.*[7] who estimated a probability of bad hygiene between 10% and 37% of the population that we translated into a beta(8,28) (mean=0.22, CI(95%)= [0.10,0.37]).

- p_h, the probability of cross-contamination from a contaminated broiler in a household is taken as the product of p_{hc} times p_{hh}.

Consumption module:

- Consumption data took the form of a series of household purchases of raw chicken meat over four-week periods, modeled as Poisson random variables with intensity λ_c. Again a hierarchical model is retained to take into account the variability of λ_c over the French population, which is modeled as a Gamma(ab, b) distribution for λ_c.

Exposure module:

- Conditionally on the number, say X, of times some raw chicken has been brought into the household, the number of times the household is exposed is taken as a Binomial random variable $\mathcal{B}(X, p)$, where $p = p_h \times p_b$, which corresponds to assuming that for each household the possible contaminations are independent. This allows us to represent the probability of exposure during a four-week period as $p_e = 1 -$

$\exp(-\lambda_e)$, where $\lambda_e = \lambda_c \times p_h \times p_b$. From this we deduce the probability of exposure for a household during one year as $p_{ey} = 1 - (1 - p_e)^{13}$. The risk is determined on the basis of cross-contamination process assuming for example, that people in the household, like children for example, not eating chicken can still be infected with the pathogen through the entrance in the household of a contaminated chicken.

- Note that (λ_e, p_e, p_{ey}) are equivalent parameterizations for exposure (no stochastic relationships implied and one-to-one mappings).

Illness module:

- The formulae used to deduce the global probability of illness from the attributable fraction (p_{iq}), the probability of illness from poultry (p_{ib}) and the probability of exposure to poultry (p_{ey}) were based on the assumption that the dose response effect was identical for poultry and other routes. p_{iq} was assessed from control studies: Evans *et al.*,[26] Friedman *et al.*[27] and Nauta *et al.* (Table 2.2)[9] which leads to a population risk having mean 23% and lying between 10% and 33% that we translated into a beta(9,30) (mean=0.23, CI(95%)=[0.11,0.37]).

- For $p_{in,}$ the probability of illness conditional on infection, we adopted the same strategy as in the CARMA project[9] where p_{in} is fixed based on Black[28] because its uncertainty is not well quantified in the literature.

- The probability of being ill conditional on exposure, p_{ie}, is the product of p_{in} times $(1-(1-p_{ne})^d)$, the probability of infection due to the ingestion of d campylobacters where p_{ne} is the probability of infection by one ingested campylobacter. We assumed that the distribution of microorganism-host survival probability is given by the Beta

distribution proposed by Teunis *et al.*,[29] *i.e.* a beta(0.024,0.011), including the variability and uncertainty of the parameter.

- d is here the actual ingested dose; it is variable and so is represented by a distribution, which we have taken as the empirical distribution given in section 3.2.7 of the CARMA report.[9]

3.2 Resulting prior distributions

It is worth pointing out that the core model previously defined could be used in the same way as when simulating the food chain using the Monte Carlo simulation. It is a probabilistic model and using WinBUGS,[17] for instance, without data, it is simple to simulate all distributions of the random variates involved, or couples of these variates. This provides another view of the construction and its initial form can then be modified. Figure 3 shows the marginal densities of the variates of interest.

3.3 Coupling the data sets (step D)

Data sets are available in addition to the information provided by experts. These are not always directly linked to the core model and additional modeling is necessary to enable their inclusion. In this case, we decided to incorporate (i) the 16 relevant surveys included in the AFSSA report[24] on chicken flock contamination, (ii) the 14 relevant surveys (also in the AFSSA report) concerning chicken carcass contamination, (iii) purchase data on raw chicken meat from the TNS Worldpanel database[30] (for 2001) concerning 4770 households over several four-week periods spread throughout the year, and (iv) an epidemiological study carried out in the United Kingdom[31] which determined the number of campylobacteriosis cases for the equivalent of 4026 person-years. The notations are

shown in Table III.

These data were introduced in the model through their distribution conditionally on parameters or variates present in the core model. Table IV and Figure 2 show the relevant details.

For example, the use of a Poisson distribution for the purchase numbers (n_{csb}) for each household led to the Gamma-Poisson model, because a Gamma distribution was used for its parent parameter (λ_{cs}) (representing the variability between households).

3.4 Posterior distributions

Posterior distributions were computed using WinBUGS.[17] A burn in of 10^4 iterations was followed by 10^5 iterations thinned at a proportion of 1 to 100. In the same way as with prior distributions, we used the posterior distribution as a Monte Carlo simulator. The marginal posterior densities of the seven variates of primary interest are shown in Figure 3. Modeling needs to be reconsidered if the experts do not agree with the posterior distribution, but of course in such cases, outputs of other variates can be proposed to facilitate any reconsideration.

3.5 Comparing prior and posterior distributions

The differences between prior and posterior distributions are the result of introducing data sets into the model. Then, the differences provide information on agreement or disagreement between the expert opinion and the data. Systematic and detailed comparisons between prior and posterior distributions need to be performed; this is not an easy task because the information is highly multivariate. Marginal distributions must therefore not be the only ones considered. Figure 3 reveals some major differences, particularly regard-

ing λ_c, p_{ey}, p_{ib} and p_{it}. One sees clearly with these four parameters that the data, provided in the entire model, have strongly modified the prior model. One can see that posterior distributions have different shape and are narrower. Nevertheless one observes that these distributions are still very large. They cannot yet constitute a satisfying response to the risk question, since the estimates are not accurate enough for decision makers. However they reflect all the variability and uncertainty of the entire system and point out the need for more relevant data in the analysis, for instance in the hygiene module. It is also to be noted that from this analysis, we can reduce the variability on the risk response by splitting the population into groups of individuals having similar behavior. Such a refinement can be done directly from the crude model presented in the paper.

By comparison, p_f and p_b were little modified; the same applied for p_h, which is not surprising since no data were added to the corresponding module. For a more direct examination, we present bivariate distributions in Figure 4. Striking differences are then revealed. This is for instance the case for (p_{ey}, p_{ib}) and (p_{ib}, p_{it}). Note that such bivariate visual comparisons required many more simulated points than univariate comparisons.

3.6 Sensitivity analysis

In order to determine the stability of the results obtained, different components in the construction were altered to assess their marginal importance. The tests performed are summarized in Table V. They mainly concerned prior distributions on parameters (*e.g.* piq0) and the inclusion of data sets in the analysis (*e.g.* wed). But attempts were also made to assess the effect on a small part of the core model (*e.g.* pie0). For prior distributions, the modifications could be of different types (shift, less or more precise, etc...). The effects were mainly assessed on variates of interest but in some cases it was

useful to detail changes on other nodes of the Bayesian network in order to better clarify their nature. The conclusions of these tests produced incentives and suggestions for model improvements, or indications concerning the weakness of the information available, thus highlighting directions to be investigated as a priority.

Using diagrams similar to those shown in Figures 3 and 4, we were able to assess which variates of interest were mainly affected by sensitivity trials, at both the prior and posterior levels. It was shown that very similar relative effects on variates of interest were obtained during the four sensitivity trials. These are summarized in Table VI, and Figure 5 illustrates certain cases. The adjectives slight, moderate, and strong have been chosen subjectively.

- In Table VI, when comparing priors (a), the effect was irrelevant when the distribution of a variate was not affected by the modification or was the variate itself.
- p_h was strongly affected by the ph0 trial. This is a consequence of the lack of data in this module.
- p_{ey} was slightly modified, except during the ph0 trial where it was markedly affected.
- λ_c was not affected by the modification of prior distributions. However, it was unexpectedly affected by epidemiological data.
- p_f and p_b behaved similarly.
- p_{ib} and p_{it} were sensitive to the prior distributions but became stable when the epidemiological data were introduced.

4 Discussion

4.1 The campylobacter-broiler application

As explained in the introduction, our aim was to adopt a global approach. In doing so, every subpart of our modeling can be considered as crude. The objective was not to reproduce reality in a mechanical way, but to determine whether the level of approximation we achieved was sufficient for the model to be useful, which is probably not yet the case. A second round with experts in the field is now necessary, in order to improve the knowledge available on each variate of interest. However, we consider it is useful to propose a framework for this global approach and to demonstrate that conceptual and numerical tools are available to enable this kind of application. Moreover, by doing so, further discussions with experts are simplified since this work has shown where more information is definitely needed, and to which extent it is needed.

The global model we have proposed for the campylobacter-broiler problem can be improved in numerous ways. In particular, we did not introduce the modeling of contamination levels, and only considered the presence/absence of bacteria. In this specific context, we argue that (i) most information is to be found at this level, and when it is not, the correspondence between the different methods used to measure levels of contamination remains an unsolved problem, (ii) binary variates could be associated with a given threshold, not only naught or positive.

There is a desperate need for data on the hygiene module. As can be seen from the sensitivity analysis hygiene has a strong influence on exposure. Most efforts have focused on what we called the *illness* module where two goals can be highlighted: (1) improvements to the concept of dose consumption via broilers or other routes when an

individual is exposed, and (2) the development of more efficient methods to link epidemiological studies with the modeling of quantitative risk assessment in a food chain; for obvious reasons, experimentation on humans for dose response is not possible, although it is unfortunately the principal way to improve such quantitative assessments. As can be seen from the sensitivity analysis epidemiological data has an influence, which is reliant upon this relationship. Note that we included the epidemiological data assuming an error between p_{it}, the probability of a person suffering from campylobacteriosis in France within one year, whatever the source and p_{its}, the probability of a person suffering from campylobacteriosis in a similar country within one year whatever the source, because the available epidemiological data were English data. So, it seemed reasonable to include a possible difference between the two countries. A fixed standard deviation has been given to this error term ($s_s = 1$). This choice has been made after different sensitivity analyses not shown here. The choice of $s_s = 1$ allows an introduction of these data in a reasonable manner for us (not too soft, not too hard). Surely, this choice has an important impact on the influence of these data, and a more thorough analysis on this should be considered.

Another key point is the integration in the model of the different spatial and temporal scales at which the modules behave. Improvements to this integration will be very difficult, but essential if more precise assessments are to be achieved.

4.2 The methodology proposed

Software programs such as WinBUGS[17] and Jags[16] are available and allow simple implementation of these approaches. However, a cautious approach must be adopted towards the convergence of algorithms. Specific programming will enable improved control of the sampling algorithms; for example, monitoring of the Metropolis-Hastings step can be

better adapted, and in most cases algebraic simplifications can be introduced.

The assembly of modules raises new problems, including the relative flexibility of different modules. If a prior distribution is too strongly peaked, it will not be influenced by the other modules. But if a module is too flexible, it may absorb variation (even of a random nature) from other modules. This was the case for λ_c, which was sensitive to the removal of epidemiological data. A balance must be introduced between different components in the chain. This is certainly more difficult when insufficient data are available, a key trait of quantitative risk assessment concerning food-borne diseases. The relative strength of the links between the principal modules thus deserves further investigation.

There are several incentives to developing the approach we propose. Clearly, the first is the theoretical possibility of being able to address highly complex situations through consistent use of the data sets and expert knowledge available. Classical statistical methods are inefficient in such situations, and standard Monte Carlo approaches cannot use data sets downstream of other data sets. There are no such limitations to probabilistic modeling defined using a Bayesian network and exploited with Bayesian statistics, because back propagation is the consequence of the full joint probability distribution conditioned by the quantities observed.

This means that it is possible to tackle the phenomenon globally, in the hope that information will spread efficiently throughout the Bayesian network and thus benefit all modules. Study of the appropriateness of information dissemination throughout the network is both necessary and instructive and can be achieved via sensitivity analysis, as proposed in this paper.

Assuming the model is satisfactory, it is possible to gain a clearer understanding of the effect of interventions on outcomes. By altering the prior distribution of a variate to

mimic the implementation of a new strategy, the joint posterior probability distributions will be conditioned on the new prior and the model output will thus reflect the outcome that would be anticipated following the implementation of the new strategy.

As was studied in the case of epidemiological data, it is relatively simple to remove data sets in order to understand their effect on the construction of joint distributions and their importance to the conclusions reached. It is even possible to introduce data substitutes into the system, solely to determine the importance of their presence. The addition of new data is relatively straightforward, and does not necessarily require new modeling.

Finally, a major advantage of this two-steps (Model-Data) methodology is that it enables a clear distinction between expert ideas and the information provided by data sets, with an opportunity in steps **M4** and **D3** to interact with the model construction. Nevertheless, it must be acknowledged that the splitting between expert knowledge and formal data is not so clear. For instance, a way to construct modules or priors from experts' mind is asking them to propose likely data. Then, some variation can be introduced at this level. It should also be noted that inputs and outputs are joint probability distributions on the same set of variates, a satisfactory consequence of the Bayesian approach retained: if experts are able to provide prior distributions, they are able to interpret posterior distributions.

ACKNOWLEDGMENTS

The authors would like to thank both referees for their judicious comments on the first draft of this article and the members of the AFSSA workshop[24]. The authors are grateful to Christine Boizot-Szanti, from INRA-CORELA, for the discussion on consumption data and for their extraction from the TNS Woldpanel database.

References

1. Codex Alimentarius, FAO/WHO (2006).

   http://www.codexalimentarius.net/web/standard_list.do

2. M.H. Cassin, A.M. Lammerding, E. C. Todd, W. Ross and R. S. McColl (1998). Quantitative risk assessment for *Escherichia coli* O157:H7 in ground beef hamburgers, *Int. J. Food Microbiology*, 41(1): 21-44.

3. M.J. Nauta (2001). *A modular process risk model structure for quantitative microbiological risk assessment and its application in an exposure assessment of* Bacillus cereus *in a REPFED*, Rivm Report 149106 007, Bilthoven, The Netherlands.

4. M.J. Nauta (2001). Modelling bacterial growth in quantitative micobiological risk assessment: is it possible?, *Int. J. Food Microbiology*, 73(2-3): 297-304.

5. A. Patwardhan and M. J. Small (1992). Methods for Model Uncertainty Analysis with Application to Future Sea Level Rise, *Risk analysis*, 12(4): 513-523.

6. K.P. Brand and M.J. Small (1995). Updating Uncertainty in an Integrated Risk Assessment: Conceptual Framework and Methods, *Risk analysis*, 15(6): 719-731.

7. B. Christensen, H. Sommer, H. Rosenquist and N. Nielsen (2001). *Risk assessment on* Campylobacter jejuni *in chicken products,* The Danish veterinary and food administration, Institute of food safety and toxicology, Division of microbiological safety, 145pp.

8. FAO/WHO (2001). *Hazard identification, hazard characterization and exposure assessment of* Campylobacter spp. *in broiler chickens; preliminary report,* Joint

FAO/WHO activities on risk assessment of microbiological hazards in food, 143 pp.

9. M. J. Nauta, W. F. Jacobs-Reitsma *et al.* (2005). *Risk assessment of Campylobacter in the Netherlands via broiler and other routes*, Rivm Report, Bilthoven, The Netherlands, 128pp.

10. R. Pouillot, I. Albert, M. Cornu and J.-B. Denis (2003). Estimation of uncertainty and variability in bacterial growth using Bayesian inference. Application to Listeria monocytogenes, *Int. J. Food Microbiology*, 81(2): 87-104.

11. C.M. Bishop (2006). *Pattern Recognition and Machine Learning*, Chap. 8, Springer.

12. S.L. Lauritzen, A.P. Dawid, B.N. Larsen and H.-G. Leimer (1990). Independence properties of directed Markov fields, *Networks*, 20(5): 491-505.

13. Ch. Robert (2001). *The Bayesian choice*, Texts in Statistics, Springer, second edition.

14. A. Gelman, J. B. Carlin, H. S. Stern and D. B. Rubin (2004). *Bayesian data analysis*, Texts in Statistical Science, Chapman & Hall/CRC, second edition.

15. W. R. Gilks, S. Richardson and D. J. Spiegelhalter (1996). *Makov Chain Monte Carlo in practice*, Chapman & Hall, London.

16. M. Plummer (2005). Jags, *http://www-fis.iarc.fr/~martyn/software/jags/*

17. D. Spiegelhalter, A. Thomas *et al.* (2003). *WinBUGS version 1.4 User manual*, MRC Biostatistics Unit, Cambridge, United Kingdom.

18. M. Plummer, N. Best, K. Cowles and K. Vines (2005). *The coda package,* http://www.R-project.org

19. **R** Development Core Team (2004). *R: A language and environment for statistical computing*, http://www.R-project.org

20. M. Kuusi, J.P. Nuorti, M.L. Hanninen, M. Koskela, V. Jussila, E. Kela, I. Miettinen and P. Ruutu (2005). A large outbreak of campylobacteriosis associated with a municipal water supply in Finland, *Epidemiol Infect*, 133(4): 593-601.

21. Anonymous (2004). Opinion of the scientific panel of biological hazards on the request from the commission related to *Campylobacter* in animals and foodstuffs, *The EFSA Journal*, 173: 1-10.

22. P. Luber, S. Brynestad, D. Topsch, K. Scherer and E. Bartelt (2006). Quantification of Campylobacter Species Cross-Contamination during Handling of Contaminated Fresh Chicken Parts in Kitchens, *Appl. Envir. Microbiol.*, 72(1): 66-70.

23. A. Wingstrand, J. Neimann, J. Engberg, E. Møller Nielsen, P. Gerner-Smidt, H. C. Wegener and K. Mølbak (2006). Fresh Chicken as Main Risk Factor for Campylobacteriosis, Denmark. *Emerg Infect Dis.*, 12: 280-284.

24. F. Mégraud, C. Bultel *et al.* (2004). *Appréciation des risques alimentaires liés aux campylobacters*, Rapport Afssa, Maisons-Alfort, France, 96pp.

25. S. Yang, M.G. Leff, D. Mctague, K.A. Horvath, J. Jackson-Thompson, T. Murayi, G.K. Boeselager, T.A. Melnik, M.C Gildemaster, D.L. Ridings, S.F. Altekruse and F.J. Angulo (1998). Multistate surveillance for food-handling, preparation, and consumption behaviors associated with food-borne diseases: 1995 and 1996 BRFSS Food-Safety Questions, *Mor. Mortal. Wkly. Rep. CDC Surveill,* 47(SS-4): 33-54.

26. M. R. Evans, C. D. Ribeiro and R. L. Salmon (2003). Hazards of healthy living: bottled water and salad vegetables as risk factors for campylobacter infection, *Emerging Infectious Diseases,* 9(10): 1219-1225.

27. C. R Friedman, R. M. Hoekstra, M. Samuel, R. Marcus, J. Bender, B. Shiferaw, S. Reddy, S. D. Ahuja, D. L. Helfrick, F. Hardnett, M. Carter, B. Anderson, R. V. Tauxe *et al.* (2004). Risk factors for sporadic Campylobacter infection in the United States: a case-control study in FoodNet sites, *Clin. Infect. Dis.,* 38(3): 285-296.

28. R.E. Black, M.M. Levine, M.L. Clements, T.P. Hughes and M.J. Blaser (1988). Experimental Campylobacter jejuni infection in humans, *J. Infect. Dis.,* 157: 472-479.

29. P. Teunis, W. van den Brandhof, M. Nauta, J. Wagenaar, H. van den Kerkhof and W. van Pelt (2005). A reconsideration of the Campylobacter dose-response relation, *Epidemio. Infect.,* 133: 583-592.

30. TNS Worldpanel (2006). *http://www.secodip.fr*

31. J.G. Wheeler, D. Sethi *et al.* (1999). Study of infectious intestinal disease in England: rates in the community, presenting to general practice, and reported to national surveillance, *British Medical Journal*, 318: 1046-1050.

A Bugs code

```
model  { # starting the model ........................
### CORE MODEL ########################################
### CHICKEN FARM MODULE ................................
p.f <- exp(lp.f)/(1+exp(lp.f));                      #vi
lp.f ~ dnorm(m.f, tau.f);                            #..
m.f ~ dnorm(0.0,20.66);                              #pa
s.f ~ dunif(0,0.2); tau.f <- 1/(s.f*s.f);            #pa
```

```
### BROILER PRODUCTION MODULE ..........................
p.b <- exp(lp.b)/(1+exp(lp.b));                        #vi
lp.b ~ dnorm(m.b, tau.b);                              #..
m.b <- m.f + d.bf;                                     #cv
d.bf ~ dnorm(0.1,206.6)I(0,);                          #pa
s.b ~ dunif(0,0.2); tau.b <- 1/(s.b*s.b);              #pa
### HYGIENE MODULE .....................................
p.h <- p.hc * p.hh;                                    #vi
p.hc ~ dbeta(8,8);                                     #pa
p.hh ~ dbeta(8,28);                                    #pa
### CONSUMPTION MODULE .................................
lambda.c ~ dgamma(ab.c,b.c);                           #vi
ab.c <- a.c*b.c;                                       #..
a.c ~ dgamma(4,4);                                     #pa
b.c ~ dgamma(10,10);                                   #pa
### EXPOSURE MODULE ....................................
p.ey <- 1 - pow((1-p.e),13);                           #vi
p.e <- 1 - exp(-lambda.e);                             #cv
lambda.e <- p.b * p.h * lambda.c;                      #cv
### ILLNESS MODULE .....................................
p.ib <- p.ey * p.ie;                                   #vi
p.it <- p.ib / (1 - (1-p.iq)*(1-p.ey));                #vi
p.ie <- (1-pow(1-p.ne,d)) * p.in;                      #cv
d <- vd[c.d];                                          #cv
p.ne ~ dbeta(0.024,0.011)I(0.00001,0.999999);          #pa
p.in <- 0.33;                                          #pa
p.iq ~ dbeta(9,30);                                    #pa
c.d ~ dcat(c.di[]);                                    #..
### AUGMENTED MODEL FOR DATA INCORPORATION ##############
### CHICKEN FARM MODULE ....................................
for (i.fs in 1:n.fst) {                                #..
  g.fs[i.fs] ~ dbin(p.fs[i.fs],n.fs[i.fs]);            #da
  p.fs[i.fs] <- exp(lp.fs[i.fs]) /                     #cv
                  (1 + exp(lp.fs[i.fs]));              #..
  lp.fs[i.fs] ~ dnorm(m.f,tau.f);                      #..
}                                                      #..
### BROILER PRODUCTION MODULE ..........................
for (i.bs in 1:n.bst) {                                #..
  g.bs[i.bs] ~ dbin(p.bs[i.bs],n.bs[i.bs]);            #da
  p.bs[i.bs] <- exp(lp.bs[i.bs]) /                     #cv
                 (1 + exp(lp.bs[i.bs]));               #..
  lp.bs[i.bs] ~ dnorm(m.b,tau.b);                      #..
}                                                      #..
### HYGIENE MODULE .....................................
### CONSUMPTION MODULE .................................
for (i.cs in 1:n.cst) {                                #..
  n.csb[i.cs] ~ dpois(lambda.csb[i.cs]);               #da
  lambda.csb[i.cs] <- n.cs[i.cs]*lambda.cs[i.cs];      #cv
  lambda.cs[i.cs] ~ dgamma(ab.c,b.c);                  #cv
}                                                      #..
### EXPOSURE MODULE ....................................
### ILLNESS MODULE .....................................
g.its ~ dbin(p.its,n.its);                             #da
logit(p.its) <- logit(p.it) + err;                     #cv
s.s <- 1; tau.s <- 1/(s.s*s.s);                        #pa
err ~ dnorm(0,tau.s);                                  #..
} # ending the model ###################################
list( # starting the doses ..............................
```

```
c.di=c(0.500,0.163,0.222,0.097,0.018), vd=c(1,2,10,100,300))
list( # starting the data incorporated ..........................
n.fst=16,
n.fs=c(287,79,100,8911,75,59,125,155,403,398,112,62,187,100,22,450),
n.bst=14,
n.bs=35,82,2925,410,251,691,203,79,50,97,180,708,994,858),
n.cst=4770,
n.cs=c(.....),
n.its=4026,
g.fs=c(77,29,40,3787,32,29,63,80,228,226,64,50,153,90,22,294),
g.bs=c(28,18,1082,157,99,333,163,70,43,78,146,291,338,80),
n.csb=c(..........),
g.its=32)
```

Table I: variates of interest (vi) in the six modules.
The framework for the application of variates must be clearly specified. In this case, the framework is mainland France for one year, and consumption refers to that of raw chicken meat at home.

Module	**Variate**	**Definition**
Chicken farm	p_f	probability of a chicken flock being contaminated
Broiler production	p_b	probability of a chicken carcass being contaminated
Consumption	λ_c	intensity of chicken consumption in a household over a four-week period
Hygiene	p_h	probability of cross-contamination from a contaminated broiler in a household
Exposure	p_{ey}	probability of a person being exposed during a period of one year
Illness	p_{ib}	probability of a person suffering from campylobacteriosis due to broiler meat within one year
Illness	p_{it}	probability of a person suffering from campylobacteriosis within one year, whatever the source

Figure 1: Graph description of the core model.
All random variates of the core model are included in the diagram. Variates of interest are circled. Parameters are indicated by triangles; their marginal distributions must be specified. The distribution of other nodes (variables of interest and complementary variables) are specified conditionally on their parent variates. Parent variates are indicated by means of arrows: *e.g.* the parents of p_{it} are p_{ey}, p_{ib} and p_{iq}. Thick arrows indicate a functional relationship, otherwise this is probabilistic (defined in Table II).

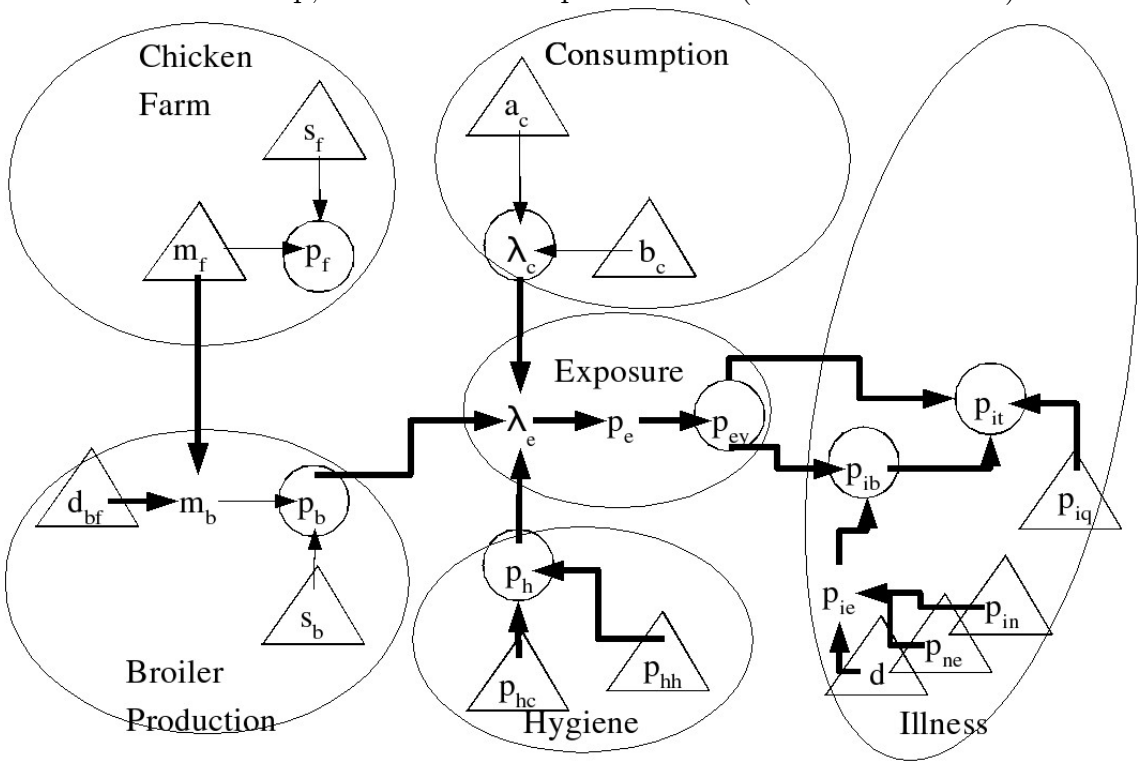

Table II: Prior distribution of the core model.

Each line of the table defines a variate. This may involve some parents and can result from either a logical relationship (=) or a distribution ($\sim$). $N(m, s)$ stands for the normal distribution with an expectation m and variance s^2. Possible truncation of the distributions are indicated by inequalities. Except for the normal distribution, the parameters defining distributions are those used in WinBUGS[17] and Jags[16] software programs. The variates are classified according to their status: variate of interest (vi), complementary variate (cv) and parameter (pa).

Class	**Variate**	**Parent(s)**	**Distribution / Relationship**
vi	p_f	m_f, s_f	$logit(p_f) \sim N(m_f, s_f)$
vi	p_b	m_b, s_b	$logit(p_b) \sim N(m_b, s_b)$
vi	p_h	p_{hc}, p_{hh}	$= p_{hc}p_{hh}$
vi	λ_c	a_c, b_c	$\sim Gamma(a_c b_c, b_c)$
vi	p_{ey}	p_e	$= 1-(1-p_e)^{13}$
vi	p_{ib}	p_{ey}, p_{ie}	$= p_{ey}p_{ie}$
vi	p_{it}	p_{ey}, p_{ib}, p_{iq}	$= p_{ib}/\left(1-(1-p_{iq})(1-p_{ey})\right)$
cv	m_b	m_f, d_{bf}	$= m_f + d_{bf}$
cv	λ_e	p_b, λ_c, p_h	$= p_b\lambda_c p_h$
cv	p_e	λ_e	$= 1-\exp(-\lambda_e)$
cv	p_{ie}	d, p_{ne}, p_{in}	$= \left(1-(1-p_{ne})^d\right) p_{in}$
pa	m_f	-	$\sim N(0, 0.22)$
pa	s_f	-	$\sim Uniform(0, 0.2)$
pa	d_{bf}	-	$\sim N(0.1, 0.0696) > 0$
pa	s_b	-	$\sim Uniform(0, 0.2)$
pa	p_{hc}	-	$\sim Beta(8, 8)$
pa	p_{hh}	-	$\sim Beta(8, 28)$
pa	a_c	-	$\sim Gamma(4, 4)$
pa	b_c	-	$\sim Gamma(10, 10)$
pa	d	-	$\sim$ (1,2,10,100,300) with p = (0.5,0.163,0.222,0.097,0.018)
pa	p_{ne}	-	$\sim Beta(0.024, 0.011)$
pa	p_{in}	-	$= 0.33$
pa	p_{iq}	-	$\sim Beta(9, 30)$

Table III: The four data sets included in the statistical analysis

Module	**Variate**	**Definition**
Chicken farm	g_{fs}	number of chicken flocks contaminated out of n_{fs} in sample $s = 1, ..., 16$.
Broiler production	g_{bs}	number of chicken carcasses contaminated out of n_{bs} in sample $s = 1, ..., 14$.
Consumption	n_{csb}	number of broilers purchased by a given household during n_{cs} periods of four weeks in sample $s = 1, ..., 4770$.
Illness	g_{its}	number of people suffering from campylobacteriosis in a sample of size n_{its} during a year in a similar country.

Table IV: Likelihood and completion of the prior distribution.
The same conventions are used as in Table II, except that data (da) replace the variate of interest. A column giving the size (number of scalar components or observations) has also been added, together with constants (co) associated to these sizes.

Class	**Variate**	**Parent(s)**	**size**	**Distribution / Relationship**
da	g_{fs}	p_{fs}	16	$\sim Binomial\,(n_{fs}, p_{fs})$
da	g_{bs}	p_{bs}	14	$\sim Binomial\,(n_{bs}, p_{bs})$
da	n_{csb}	λ_{csb}	4770	$\sim Poisson\,(\lambda_{csb})$
da	g_{its}	p_{its}	1	$\sim Binomial\,(n_{its}, p_{its})$
cv	p_{fs}	m_f, s_f	16	$logit(p_{fs}) \sim N(m_f, s_f)$
cv	p_{bs}	m_b, s_b	14	$logit(p_{bs}) \sim N(m_b, s_b)$
cv	λ_{cs}	a_c, b_c	4770	$\sim Gamma(a_c b_c, b_c)$
cv	λ_{csb}	λ_{cs}	4770	$= n_{cs}\lambda_{cs}$
cv	p_{its}	p_{it}, s_s	1	$logit(p_{its}) \sim N(logit(p_{it}), s_s)$
pa	s_s	-	1	$= 1$
co	n_{fs}	-	16	...
co	n_{bs}	-	14	...
co	n_{cs}	-	4770	...
co	n_{its}	-	1	$= 4026$

Figure 2: Augmented model.
Variates and parameters necessary to describe the available data have been added to the core model described in Figure 1.

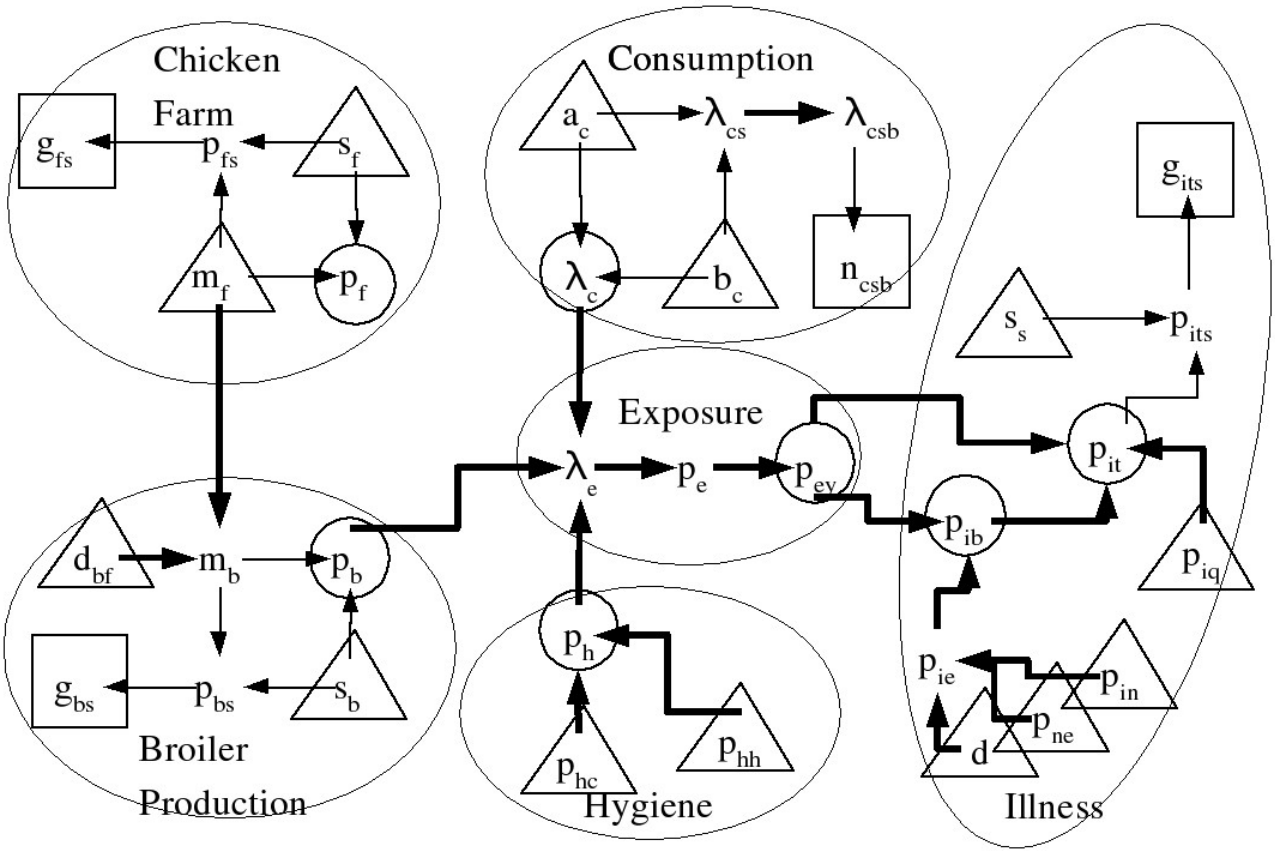

Figure 3: Posterior marginal densities of the seven variates of interest defined in Table I. These posterior distributions (solid lines) should be compared with the prior distributions indicated here using dotted lines.

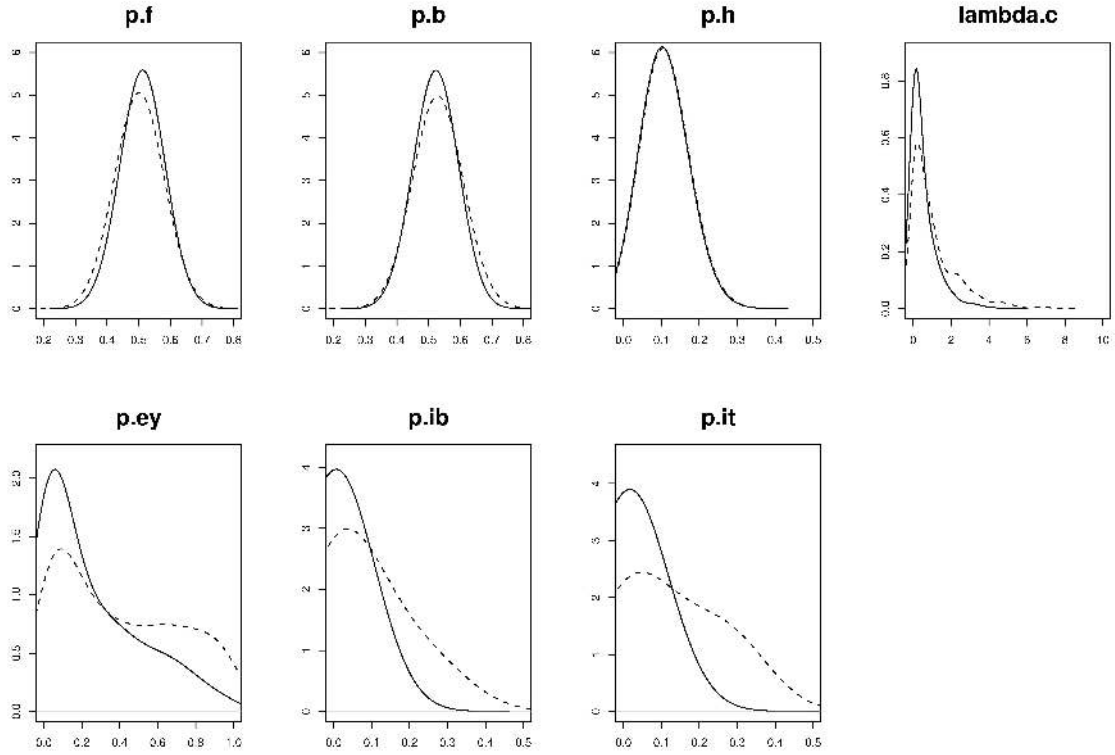

Figure 4: Comparisons between prior (a-b-c) and posterior (d-e-f) distributions. Diagrams (a) and (d) display the variates p_h and p_{ey}; diagrams (b) and (e) display the variates p_{ey} and p_{ib}; diagrams (c) and (f) display the variates p_{ib} and p_{it}. The distributions are represented by means of 1000 thousands simulated parameter vectors from the prior and posterior distributions.

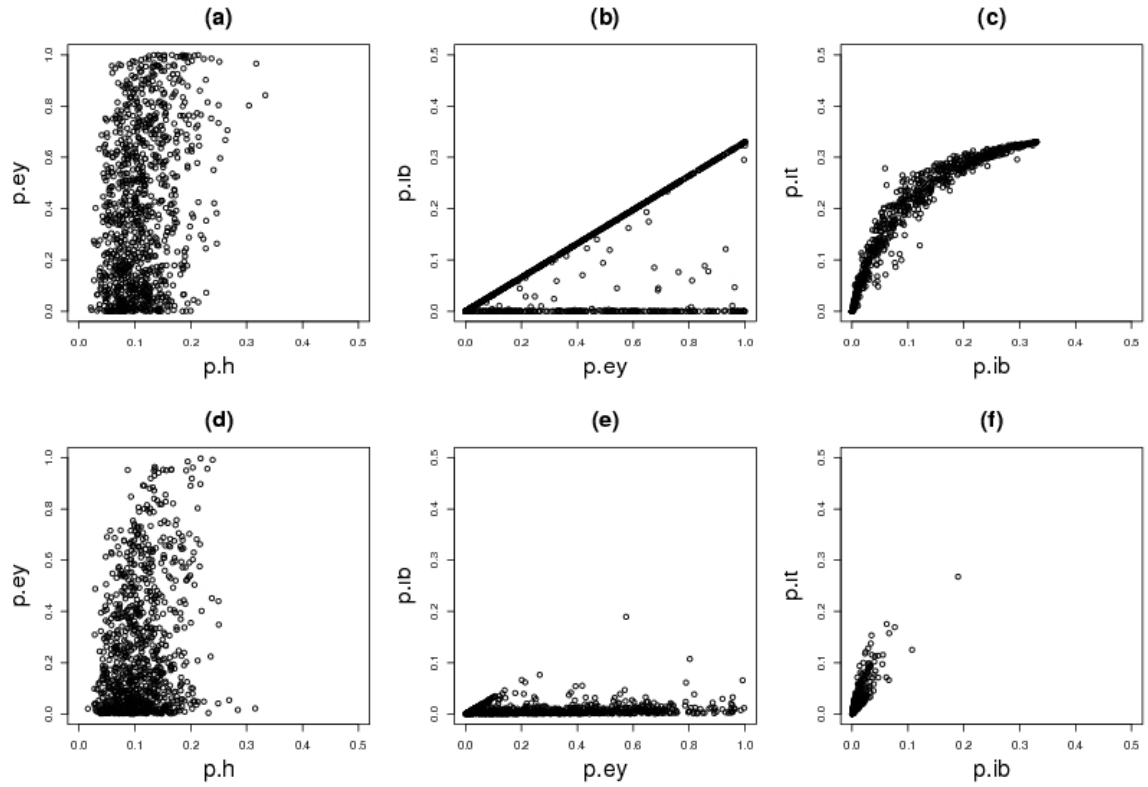

Table V: Sensivity trials.

Each line of the table is associated with a sensitivity trial and provides the code name used to designate it, the aim and the difference between the basic model (described in Table II) and the trial.

name	**aim**: see the effect of	**basic case**	**modification**
ph0	prior information on p_h	$Beta(8,8)$*$Beta(8,28)$	$Beta(\frac{1}{2},\frac{1}{2})$
pie0	prior information on p_{ie}	(see Table II)	$Beta(\frac{1}{2},\frac{1}{2})$
piq0	prior information on p_{iq}	$Beta(9,30)$	$Beta(\frac{1}{2},\frac{1}{2})$
wed	epidemiological data from a similar country	with g_{its} data	without g_{its} data

Table VI: Synthetic result of the sensitivity analysis.
For each sensitivity trial the magnitudes of the effect on the variates of interest are reported: irrelevant (-), no effect (0), very slight effect (1), slight effect (2), moderate effect (3) and strong effect (4). Three types of effect are distinguished : (a) modification to the prior, (b) modification to the posterior, (c) amount of data included [prior=no data / wed=data without epidemiological data / posterior= all available data]. The 6 cases, in bold type, are illustrated in Figure 5.

	p_f	p_b	p_h	λ_c	p_{ey}	p_{ib}	p_{it}
a: ph0	-	-	-	-	4	2	1
a: pie0	-	-	-	-	-	2	**3**
a: piq0	-	-	-	-	-	-	0
b: ph0	0	0	4	0	**4**	0	0
b: pie0	0	0	0	0	1	0	**0**
b: piq0	0	0	0	0	1	0	0
c: prior / wed	1	1	0	**1**	2	1	1
c: wed / posterior	0	0	0	0	**2**	**3**	4

Figure 5: Comparisons of marginal distributions in certain sensitivity trials.
a-b-c cases refer to the caption to Table VI; the trial is indicated between parentheses (see Table V); the variate is indicated between brackets. Dotted lines represent the basic computation and solid lines the sensitivity cases.