A bayesian evidence synthesis for estimating campylobacteriosis prevalence

Isabelle I. Albert, E. E. Espié, H. H. de Valk, Jean-Baptiste J.-B. Denis

Risk Analysis, Vol. 31, No. 7, 2011 DOI: 10.1111/j.1539-6924.2010.01572.x

A Bayesian Evidence Synthesis for Estimating Campylobacteriosis Prevalence

Isabelle Albert,[1,*] Emmanuelle Espié,[2] Henriette de Valk,[2] and Jean-Baptiste Denis[3]

Stakeholders making decisions in public health and world trade need improved estimations of the burden-of-illness of foodborne infectious diseases. In this article, we propose a Bayesian meta-analysis or more precisely a Bayesian evidence synthesis to assess the burden-of-illness of campylobacteriosis in France. Using this case study, we investigate campylobacteriosis prevalence, as well as the probabilities of different events that guide the disease pathway, by (i) employing a Bayesian approach on French and foreign human studies (from active surveillance systems, laboratory surveys, physician surveys, epidemiological surveys, and so on) through the chain of events that occur during an episode of illness and (ii) including expert knowledge about this chain of events. We split the target population using an exhaustive and exclusive partition based on health status and the level of disease investigation. We assume an approximate multinomial model over this population partition. Thereby, each observed data set related to the partition brings information on the parameters of the multinomial model, improving burden-of-illness parameter estimates that can be deduced from the parameters of the basic multinomial model. This multinomial model serves as a core model to perform a Bayesian evidence synthesis. Expert knowledge is introduced by way of pseudo-data. The result is a global estimation of the burden-of-illness parameters with their accompanying uncertainty.

KEY WORDS: Bayesian evidence synthesis; burden-of-illness; campylobacteriosis; expert opinion; multinomial model; surveillance data

1. INTRODUCTION

Assessing the burden of foodborne infectious diseases is a challenging statistical and epidemiological objective for public health and world trade decisions. Good estimates of burden-of-illness are needed to improve the quality of quantitative risk assessment (QRA) models. In particular, the use of available human data regarding studied food disease can provide better accuracy in dose-response models.[1]

In a case study in France, Albert *et al.*[2] demonstrated that including Campylobacter—one of the most common causes of acute bacterial gastroenteritis in industrialized countries—epidemiological data in a global Bayesian approach may improve the QRA model. However, only a unique foreign survey[3] examined the prevalence of the disease (the number of campylobacteriosis cases (32 cases) for the equivalent of 4,026 person-years determined by this U.K. survey was introduced in the illness module of the food chain modeling). This was not satisfactory; it would be better to include all related human studies of the disease, i.e., to use a collection of data sources

[1]INRA-Unité Met@risk, Paris France.
[2]Institute de Veille Sanitaire, Saint Maurice, France.
[3]INRA-Unité de recherche MIA, Jouy-en-Josas, France.
*Address correspondence to Isabelle Albert, 16 rue Claude Bernard Paris 75231 cedex 05 France; isabelle.albert@paris.inra.fr.

not possibly directly linked to the prevalence of the disease but that could inform it.

In this article, we estimate the campylobacteriosis burden in France by taking advantage of all known available and pertinent data. There is a large amount of missing data because for many people, this illness is not a dramatic event, and many cases go undiagnosed or unreported. In this context, a common practice is to use point estimates from each of the available surveys and combine them into a surveillance or burden-of-illness pyramid to assess the number of cases.[4,5] This pyramid describes the chain of events occurring from any illness episode. Each step of the pyramid corresponds with a surveillance step (e.g., seeking medical care, providing a stool specimen, and getting laboratory testing, etc.), which is associated with conditional probabilities (e.g., the probability of asking for a stool culture from people seeing a physician due to acute gastroenteritis). Some links can be missing, however; in such cases, estimates are often taken from similar diseases or similar countries.[5] Also, as not all cases are included in surveillance systems (e.g., ill persons do not seek medical care, specimens are not obtained, etc.), various surveillance multipliers are applied to obtain the true number of cases in the general population from the number of registered cases. At best, a sequence of negative binomial distributions is used in Monte Carlo simulations to estimate the number of cases missed at each step of the pyramid.[6] However, uncertainties or variability associated with the estimates are not consistently conveyed because different point estimates are used in the Monte Carlo simulations without regarding the size of the different data sets involved: for example, a discrete uniform distribution was used to choose between three point estimates of the number of cases per person-years.

In this article, a Bayesian meta-analysis or more precisely a Bayesian evidence synthesis is proposed to assess burden-of-illness parameters in human studies through the chain of events occurring in an illness episode. Goubar *et al.*[7] recently proposed a similar approach for HIV infection, an illness that has received much more attention. In that case, the context was favorable because there are different active surveillance systems to detect inconsistencies between redundant data sets.[8] The term "meta-analysis" refers to a broader class of analyses encompassing results from studies that have addressed the same question in a similar way and therefore provide information on common parameters of interest (one level of parameters).[9] Whereas the term "evidence synthesis" refers to analyses encompassing diverse results from diverse sources that inform indirectly the parameters of interest through parameters that are functions of them (two levels of parameters). The synthesis of diverse sources of evidence on particular quantities of interest is increasingly employed in epidemiology as a means of exploiting all available information, even from studies of differing designs.[10–12] In a Bayesian setting, priors are assumed on parameters of the upper level giving a joint prior distribution and each piece of evidence contributes to the likelihood through its likelihood function. The likelihood of the model is then the product of the likelihood functions. The posterior distribution of the parameters of interest is obtained via Bayes's theorem. The Bayesian evidence synthesis we proposed is a complex evidence structure compared to usual evidence synthesis because it uses sometimes three levels of parameters and expert opinions are added in the synthesis to exploit even better all available information.

The objective of this study is to provide a consistent estimate of campylobacteriosis prevalence for 1 year in France given the present conditions. We only use human surveys about the disease selected by the epidemiologists in charge of this disease in France. Other estimated parameters from the proposed model are indicators of the disease burden (e.g., the probability of having a campylobacteriosis and consulting a doctor). We obtained estimates of all parameters associated with the modeled chain of events. The proposed Bayesian evidence synthesis retains and conveys uncertainties and variability of all model parameters. In the spirit of Bayesian updating, this evidence structure can be improved by adding new data.

Section 2 provides the available data for the case study and the suggested method to estimate illness burden parameters. Section 3 presents the results, which are discussed in Section 4.

2. MATERIAL AND METHODS

2.1. Data and Modeling

We gathered publications and reports that provided data that could contribute to the assessment of the situation concerning campylobacteriosis in France (see Table I). These included reports from national and regional Campylobacter surveillance systems, reports from the national reference laboratory for Campylobacter, reports from the sentinel general practitioners (GP) surveillance system for acute gastroenteritis, studies on laboratory practices and

Table I. Data Sets and Related Modeling (See Table II for the Definitions of the Qs)

Source	Provided Information	Modeling
Vaillant *et al.*[13]	$r_{1,1}$ SC[a] out of $n_{1,1}$ people (Charente-Maritime region, 1996)	$r_{1,1} \sim \text{Bin}(n_{1,1}, Q_1)$
	$r_{1,2}$ SC[a] out of $n_{1,2}$ people (Mayenne region, 1998)	$r_{1,2} \sim \text{Bin}(n_{1,2}, Q_1)$
	$r_{1,3}$ SC[a] out of $n_{1,3}$ people (Mayenne region, 1999)	$r_{1,3} \sim \text{Bin}(n_{1,3}, Q_1)$
	$r_{1,4}$ SC[a] out of $n_{1,4}$ people (Mayenne region, 2000)	$r_{1,4} \sim \text{Bin}(n_{1,4}, Q_1)$
	$r_{1,5}$ SC[a] out of $n_{1,5}$ people (CNAMTS[b], 2004)	$r_{1,5} \sim \text{Bin}(n_{1,5}, Q_1)$
	r_5 positive CSC[c] out of n_5 CSC[c] (Epicop survey, 1997)	$r_5 \sim \text{Bin}(n_5, Q_5)$
Gallay and Mégraud[14]	$r_{6,1}$ positive CSC[c] and $r_{7,1}$ CSC[c] out of $n_{6,1}$ SC (hospital's labs, 2000)	$(r_{6,1}, r_{7,1}-r_{6,1}, n_{6,1}-r_{7,1}) \sim \text{Multinomial}_3(n_{6,1}, Q_6, Q_7, 1-Q_6-Q_7)$
	$r_{6,2}$ positive CSC[c] and $r_{7,2}$ CSC[c] out of $n_{6,2}$ SC (private labs, 2000)	$(r_{6,2}, r_{7,2}-r_{6,2}, n_{6,2}-r_{7,2}) \sim \text{Multinomial}_3(n_{6,2}, Q_6, Q_7, 1-Q_6-Q_7)$
Surveillance network data	r_6 positive CSC[c] out of n_6 SC[a] (period: 2002–2005)	$r_6 \sim \text{Bin}(n_6, Q_6)$
	$r_{6,3}$ positive CSC[c] and $r_{7,3}$ CSC[c] out of $n_{6,3}$ SC[a] (period: 2005–2006)	$(r_{6,3}, r_{7,3}-r_{6,3}, n_{6,3}-r_{7,3}) \sim \text{Multinomial}_3(n_{6,3}, Q_6, Q_7, 1-Q_6-Q_7)$
CNAMTS[b]	$r_{1,6}$ SC[a] out of $n_{1,6}$ people (CNAMTS[b], 2006)	$r_{1,6} \sim \text{Bin}(n_{1,6}, Q_1)$
Sentinelles' network	r_9 people having an AGE[d] and seeing a doctor out of n_9 exposed individuals (year: 2006)	$r_9 \sim \text{Bin}(n_9, Q_9)$
Wheeler *et al.*[3]	r_2 campylobacteriosis out of n_2 individuals over 1 year	$r_2 \sim \text{Bin}(n_2, Q_2)$
	$r_{3,1}$ people having an AGE[d] out of $n_{3,1}$ exposed people over 1 month	$r_{3,1} \sim \text{Bin}(n_{3,1}, Z_3)$ $Z_3 = 1-(1-Q_3)^{1/12}$
	Incidence estimate of AGE[d] over 1 year, Inc_1 and its 95% confidence interval	$r_{3,Inc1} \sim \text{Poisson}(l_1 n_{3,Inc1})$[e] $l_1 = n_{3,Inc1}\ln(1/(1-Q_3))$
	r_4 people having a campylobacteriosis and seeing a doctor out of n_4 exposed people	$r_4 \sim \text{Bin}(n_4, Q_4)$
Frosst *et al.*[15]	$r_{3,2}$ people having an AGE[d] out of $n_{3,2}$ exposed people over 1 month	$r_{3,2} \sim \text{Bin}(n_{3,2}, Z_3)$ $Z_3 = 1-(1-Q_3)^{1/12}$
	Incidence estimate of AGE[d] over 1 year, Inc_2 and its 95% confidence interval	$r_{3,Inc2} \sim \text{Poisson}(l_2 n_{3,Inc2})$[e] $l_2 = n_{3,Inc2}\ln(1/(1-Q_3))$
Anonymous[16]	$r_{3,3}$ people having an AGE[d] out of $n_{3,3}$ exposed people over 1 month	$r_{3,3} \sim \text{Bin}(n_{3,3}, Z_3)$ $Z_3 = 1-(1-Q_3)^{1/12}$
	Incidence estimate of AGE[d] over 1 year, Inc_3 and its 95% confidence interval	$r_{3,Inc3} \sim \text{Poisson}(l_3 n_{3,Inc3})$[e] $l_3 = n_{3,Inc3}\ln(1/(1-Q_3))$
	$r_{8,1}$ people having an AGE[d] and seeing a doctor out of $n_{8,1}$ people having an AGE[d]	$r_{8,1} \sim \text{Bin}(n_{8,1}, Q_8)$
Anonymous[17]	Incidence estimate of AGE[d] over 1 year, Inc_4 and its 95% confidence interval	$r_{3,Inc4} \sim \text{Poisson}(l_4 n_{3,Inc4})$[e] $l_4 = n_{3,Inc4}\ln(1/(1-Q_3))$
	$r_{8,2}$ people having an AGE[d] and seeing a doctor out of $n_{8,2}$ people having an AGE[d]	$r_{8,2} \sim \text{Bin}(n_{8,2}, Q_8)$
Kuusi *et al.*[18]	$r_{8,3}$ people having an AGE[d] and seeing a doctor out of $n_{8,3}$ people having an AGE[d]	$r_{8,3} \sim \text{Bin}(n_{8,3}, Q_8)$

[a]SC = stool cultures.
[b]CNAMTS = French Health Insurance Fund.
[c]CSC = stool cultures of Campylobacter.
[d]AGE = acute gastroenteritis.
[e]For the incidence estimate, Inc_i ($i = 1, \ldots, 4$), a Poisson distribution was retained because some Inc_i were greater than 1. From the value of the estimate and its associated confidence interval, the size of the sampled population, $n_{3,Inci}$, was numerically retrieved and the number of cases, $r_{3,Inci}$, deduced.

results of stool cultures, and outbreak reports. We also consulted the national health insurance database on reimbursement of stool cultures. The studies had been selected by the InVS (French Institute for Public Health Surveillance), on criteria of data validity and representativeness to estimate the burden of gastrointestinal illness in France.[13] Foreign surveys have been used when no French data existed and when the foreign context was analogous to a French context (similar food hygiene, similar consumption habits, etc.). Table I describes the selected data and the associated models; Table II summarizes and provides the probabilities of events ($Q_k, k = 1, \ldots, 9$) directly informed by the selected data sets. We carefully selected models considering the representativeness of the data and their associated variability. Simple models were often chosen (nonhierarchical binomial or multinomial) due to data poverty. Nevertheless, this modeling is a key feature of the synthesis of multiple data sources.

For example, several sources provide information on the probability of having a stool culture (Q_1 in Tables I and II) derived from data on stool cultures collected from different French regions, from the number of stool cultures registered by the French National Health Insurance Fund (CNAMTS), and over different years. Assuming data independence and that they are representative of the current population (omitting spatial and temporal hierarchical modeling in agreement with the epidemiologists), each data set provides "cases" ($r_{1,i}$) from a sample of the population ($n_{1,i}$). The first index of r and n is 1 because it refers to Q_1 and the second index refers to the data set i ($i = 1, \ldots, 6$), which informs Q_1. The second index is only introduced in Table I when several sources inform the same Q.

For most of the data, a binomial model was retained. We do not report all the decisions made that led to the chosen model in this article; rather, we merely underline that these decisions constitute the model's foundation and must be made in close collaboration between epidemiologists and statisticians. Some assumptions may appear strong, but are intended to shed light on the French situation using all the data of interest.

Q_2 is the probability of having a campylobacteriosis within 1 year, which represents the main objective of this study. The data set associated directly with Q_2 comes from an English study and cannot account for the unique source of data to produce an estimation of a French Q_2. Thus, our main objective is to estimate Q_2 for France, taking into account all the available data sets gathered in Table I, along with the associated uncertainty linked to sample size. In Section 2.2.2 we show how to introduce expert opinions to take both the retained data and expert knowledge into account.

2.2. Method: Estimating the Burden-of-Illness Parameters and Associated Uncertainty

The proposed approach comprises three main steps: (i) the definition of a target-population partition associated with a multinomial model covering indirectly the independent likelihoods of all data sets; (ii) the introduction of expert knowledge as pseudo-data into the model to take all information into account improving unobserved data parameter estimation, and recalibrating the retained data; (iii) the Bayesian inference produced on the data and pseudo-data using poorly informative priors through a Dirichlet distribution for the probability parameters of the basic multinomial model.

2.2.1. Synthesis

To link the probabilities informed by the data sets in the same framework, the population of interest (i.e., the population from which probability of illness is inferred) is split based on the characteristics (events) informed by the data sets (shown in italics in Table II). This partitioning of the target population is built on the individual level of health status and disease investigation. The partition provides a formal framework that represents the population of interest and the probabilities associated with each basic event. The partition is defined in Table III as the combination of the health status (C, O, or R; C for "having a Campylobacteriosis," O for "having an-Other acute gastroenteritis" and R for "Remaining possibilities") associated with the level of investigation (y, n, s, d, or r; y for "having a positive stool culture to Campylobacter," n for "having a Negative stool culture to Campylobacter," s for "having a Stool culture without Campylobacter research," d for "having consulted a Doctor but no stool culture taken" and r for "Remaining possibilities"). Any person in the target population belongs to only one class of the partition. It means that an individual of the French population is only in one class of the partition (defined in Table III) in 1 year, which seems to be reasonable given the small size of each class and the fact that both events (acute gastroenteritis and campylobacteriosis) are relatively rare. In Table III, we define the probabilities (Ps) that a person from the French population is in each class of the

Table II. Qs probabilities and Related Data Sets or/and Expert Opinion; Qs in Terms of Ps (See Table III) and Posterior Credible Intervals at 95%

Q_k	Definition	Related Data Sets and/or Expert Opinion	Combination of Ps	$CI_{95\%}$
Q_1	Probability of having a stool culture	French health statistics[13] and CNAMTS data (year 2006)	$P_{y+} + P_{n+} + P_{s+}$	[8.90‰; 8.94‰]
Q_2	Probability of having a campylobacteriosis	English national surveillance data and expert opinion ($n_{2,p} = 685, r_{2,p} = 13$)	P_{+C}	[7.15‰; 12.68‰]
Q_3	Probability of having an acute gastroenteritis	English national surveillance data,[3] Canadian study,[15] Irish survey,[16] and Australian national gastroenteritis survey[17]	$P_{+C} + P_{+O}$	[34.80%; 36.85%]
Q_4	Probability of having a campylobacteriosis and consulting a doctor (for an acute gastroenteritis)	English national surveillance data[3]	$P_{yC} + P_{nC} + P_{sC} + P_{dC}$	[3.74‰; 4.60‰]
Q_5	Probability of having a positive stool culture to Campylobacter on the condition of having a stool culture to Campylobacter	French health statistics[13]	$P_{y+}/(P_{y+} + P_{n+})$	[4.22%; 4.34%]
Q_6	Probability of having a positive stool culture to Campylobacter on the condition of having a stool culture	French laboratory survey,[14] French surveillance network (years 2002–2005), and French surveillance network (years 2005–2006)	$P_{y+}/(P_{y+} + P_{n+} + P_{s+})$	[2.55%; 2.62%]
Q_7	Probability of having a negative stool culture to Campylobacter on the condition of having a stool culture	French laboratory survey,[14] French surveillance network (period: 2002–2005), and French surveillance network (period: 2005–2006)	$P_{n+}/(P_{y+} + P_{n+} + P_{s+})$	[57.62%; 58.01%]
Q_8	Probability of consulting a doctor on the condition of having an acute gastroenteritis	Irish survey,[16] Australian national gastroenteritis survey,[17] Norwegian survey,[18] and expert opinion ($n_{8,p} = 87, r_{8,p} = 21$)	$(P_{yC} + P_{nC} + P_{sC} + P_{dC} + P_{yO} + P_{nO} + P_{sO} + P_{dO})/(P_{+C} + P_{+O})$	[23.42%; 24.99%]
Q_9	Probability of consulting a doctor and having an acute gastroenteritis	French Sentinelles' network (year 2006)	$P_{yC} + P_{yO} + P_{nC} + P_{nO} + P_{sC} + P_{sO} + P_{dC} + P_{dO}$	[8.53%; 8.88%]
Q_{10}	Probability of having a negative stool culture to Campylobacter on the condition of having a stool culture to Campylobacter and having a campylobacteriosis	Expert opinion ($n_{10,p} = 79, r_{10,p} = 23$)	$P_{nC}/(P_{yC} + P_{nC})$	[20.80%; 40.87%]
Q_{11}	Probability of not having an acute gastroenteritis on the condition of having a stool culture	Expert opinion ($n_{11,p} = 419, r_{11,p} = 31$)	$(P_{yR} + P_{nR} + P_{sR})/(P_{y+} + P_{n+} + P_{s+})$	[5.40%; 10.65%]

Note: all probabilities are over 1 year. The notation "+" in the subscript indicates a summation over the row or the column of the partition given in Table III; for instance, $P_{y+} = P_{yC} + P_{yO} + P_{yR}$.

Table III. *P*s Probabilities (Over 1 Year) Associated with the Population Partition of Interest Using the Combination of Health Status (*C*, *O*, and *R*) and the Level of Investigation Partitions (*y*, *n*, *s*, *d*, and *r*)

	C	*O*	*R*
y	P_{yC}	P_{yO}	P_{yR}
n	P_{nC}	P_{nO}	P_{nR}
s	P_{sC}	P_{sO}	P_{sR}
d	P_{dC}	P_{dO}	P_{dR}
r	P_{rC}	P_{rO}	P_{rR}

Note: *C* is for "having a Campylobacteriosis"; *O* is for "having anOther acute gastroenteritis"; *R* is for "Remaining possibilities"; *y* is for "having a positive stool culture to CampYlobacter"; *n* is for "having a Negative stool culture to Campylobacter"; *s* is for "having a Stool culture without Campylobacter research"; *d* is for "having consulted a Doctor but no stool culture taken"; *r* is for "Remaining possibilities"; and the *P*s are the probabilities for a person sampled from the target population belonging to the subpopulation defined by the associated row and column. The sum over the 15 probabilities is 1 because each individual belongs to one, and only one, of the15 defined categories.

partition. As required, all the probabilities provided by the available data sets in Table II are expressed in terms of *P*s: Table II provides the *Q* probabilities in terms of *P*s. Regarding the models in Table I, the data sets inform the *Q*s, and the *P*s via the *Q*s. The repartition of any subset of people could be modeled as a multinomial distribution. From this core model, the model likelihood is the product of the likelihoods of each available data set. Fig. 1 provides a schematic representation of the model showing how the data sets are linked to the parameters, either directly on *Q*s or indirectly (on functions of *Q*s) and indirectly on *P*s.

2.2.2. *Introduction of Expert Opinions*

To better estimate burden-of-illness probabilities, it is natural to introduce expert opinions into the model. Expert knowledge can inform the quantities related to the partition defined above; therefore, expert knowledge adds confidence with respect to the assessment of *P*s, or their mappings as new *Q*s. Expert opinions have to be independent of the data used and related to the population to draw inferences. Specifically, in this case study, an epidemiological viewpoint of the people in charge of French disease surveillance was used (i) to simplify the model and (ii) to provide information on the situation in France.

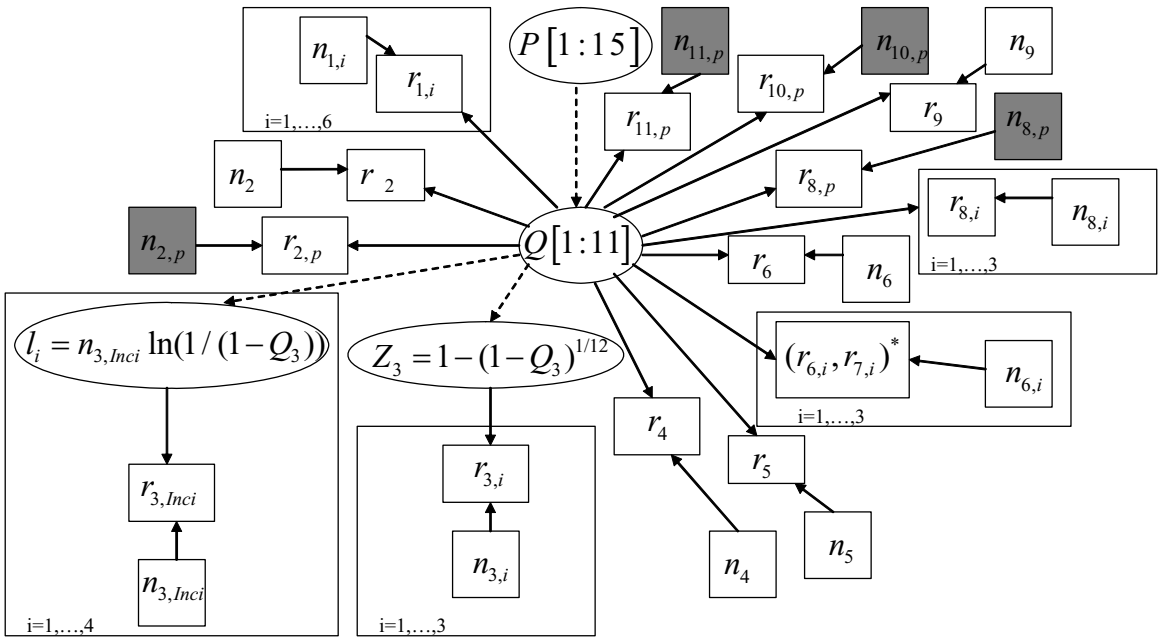

Fig. 1. Graphical representation of the model structure. A dashed line indicates a logical link and a solid line indicates a stochastic link (see Table I column: modeling). Ellipses indicate the parameters and rectangles indicate data (the pseudo-data are in gray; see Tables I and II). Surrounding rectangles, with $i = 1, \ldots, 3, 4$, or 6 at the bottom, indicate a series of similar models of data/parameters indexed by *i*. The *P*s (see Table III) derive the *Q*s (see Table II) and all data sets are modeled with the help of *Q*s (see Table I column: modeling).

(i) We supposed that

- it is possible to neglect positive Campylobacter stool cultures when having another acute gastroenteritis; that is, to assume: $P_{yO} = 0$; then, the probability of having a Campylobacter infection concomitantly with an acute gastroenteritis due to another pathogen is considered to be zero. This assumption is justified by the fact that for an individual a Campylobacter infection and acute gastroenteritis are relatively rare events.
- it is possible to neglect positive Campylobacter stool cultures when not having an acute gastroenteritis; that is, to assume $P_{yR} = 0$; healthy carriers exist but usually patients have a stool culture only in case of symptoms. Therefore, we consider that there is no person with a positive Campylobacter stool culture without symptoms of acute gastroenteritis.
- it is possible to neglect consultations for acute gastroenteritis without having an acute gastroenteritis; that is, to assume $P_{dR} = 0$; we assume that persons with no symptom of acute gastroenteritis do not consult for acute gastroenteritis.

This reduces the parametric dimension of the model to 11 (because additionally *P*s sum to one).

(ii) Also, we had expert opinions about Q_2, the probability of having a campylobacteriosis; Q_8, the probability of consulting a doctor on the condition of having an acute gastroenteritis; Q_{10}, the probability of having a negative stool culture to Campylobacter and having a campylobacteriosis; and Q_{11}, the probability of not having an acute gastroenteritis on condition of having a stool culture (see also Table II). For each of these four probabilities, the expert opinion was expressed as a confidence interval, interpreted as a 95% credibility interval. Assuming that the underlying random variable followed a beta distribution, this interval was translated into pseudo-data based on the conjugacy between binomial and beta distributions. More precisely, r successes out of n trials were retained such that the 2.5th and 97.5th quantiles of a beta(r, $n-r$) be the elicited interval. This choice is consistent with the assumed prior Dirichlet distribution for the basic P multinomial parameters. Indeed, the distributions entailed for the Qs are beta distributions since when a vector $(X_1, \ldots, X_p)$ follows a Dirichlet distribution and A and B are subsets of $\{1, \ldots, p\}$, then the ratio

$$\sum_{i \in A} X_i \Big/ \sum_{i \in \{A \cup B\}} X_i$$

follows a Beta distribution. Pseudo-data were obtained numerically with the help of an iterative algorithm programmed in R (R Development Core Team, 2008). Once the two parameters (*alpha* and *beta*) of the beta distribution (rounded to the nearest integer) followed by the elicited quantity (e.g., Q) combination of *P*s were obtained numerically from the expert confidence interval, the information is introduced independently into the model as pseudo-data, which is the equivalent of a data set comprising *alpha* successes (cases) for *alpha* + *beta* trials. That is:

$$alpha \sim \text{Bin}((alpha + beta), Q).$$

Table II provides the pseudo-data introduced in the synthesis. In the framework of the Bayesian paradigm, the basic idea is to generate pseudo-data whose integration via statistical analysis leads to estimates and uncertainties approaching the ideas of the experts.

2.2.3. Model Inference and Computation

We adopted a Bayesian approach.[19] In the first step, so-called prior probability distributions are placed on the P parameters. Due to the relationships indicated in Fig. 1 and specified in Table II, prior distributions for the Q probabilities are implicitly deduced. Taking account of the observed values of the data sets, the inference step produced posterior distributions by conditioning the parameters on the observed data. The difference between the prior and the posterior can be interpreted as the modification of previous knowledge provided by the data. In this case study, an analytical solution of the posterior distribution does not exist despite the conjugacy of the Dirichlet prior and the multinomial model because the likelihood is not directly multinomial. Indeed the data are not introduced in the multinomial model, but are incorporated through the *Q*s, which are sums, ratios, or both of the *P*s. Nevertheless, Bayesian inference can be made through Markov chain Monte Carlo (MCMC) algorithms that simulate the posterior distribution of the model parameters.[20] MCMCs are iterative algorithms that produce, after convergence, simulated values of the parameters according to the posterior distribution

and are generally autocorrelated. In practice, the first simulations (called the "burn in") are eliminated, and only a fraction of equally spaced simulations are retained (thinning option) to avoid autocorrelations.

Bayesian inference has been performed on pseudo-data to produce prior modeling using only expert knowledge and poorly informative priors on the P parameters of the multinomial distribution not equal to zero (that is, 12 Ps). A Dirichlet distribution of order 12 with parameters $\alpha_i = 1, i = 1, \ldots, 12$ was chosen as a vague prior distribution for the Ps. The posterior distributions of the model parameters were produced using the modeling and the data sets presented in Table I. Prior and posterior distributions were obtained using MCMC algorithms from Jags 1.0.3[21] and OpenBUGS 3.0.3[22] software packages.

A burn-in period of 2.10^5 iterations was followed by 10^6 iterations (thinned by 1/100). We produced two chains using different initial values in each software package. The results from both software packages are consistent when the Dirichlet distribution is based on the gamma distributions—that is, when the formula defining the prior:

$$\mathbf{P} \sim \text{Dir}(\mathbf{1}_{12})$$

is replaced with:

$$P_i = G_i \Big/ \sum_{i=1}^{12} G_i \quad \forall i = 1, \ldots, 12,$$

where

$$G_i \sim \text{Gamma}\,(1, 1) \quad \forall i = 1, \ldots, 12 \text{ (independent)}.$$

Both parameterizations are mathematically equivalent, but they do not lead to the same sampling algorithms inside the MCMC procedure and for both software packages, the multivariate Dirichlet method generates problems (e.g., too many iterations detected in the nonconjugate Dirichlet sampling algorithm in OpenBugs and nonconvergence of the two chains in Jags). We considered that similar results obtained by two independent software packages from the gamma parameterization reinforced the plausibility of convergence to the posterior distribution.

2.2.4. *Relevance of the Data for Estimating the* P *Parameters*

Roughly speaking, we wanted to estimate 11 P parameters from data associated with 11 Q parameters that are mappings of them. One can hope that there is equivalence between Ps and Qs but this has to be investigated since some redundancy can occur within the Q parameters that are not independent. This is related to the so-called identification point in statistics. Even if, in the Bayesian framework, the existence of a proper prior on the model parameters relieves of such difficulty (if the prior is defined, the posterior is defined), it is important to know which mappings of the parameters are informed or not by the data sets. For instance, this could bring help about which new data sets would be necessary to get a complete inference about the model. To investigate the point, we looked for a minimal set of new parameters (denoted later Rs), equivalent to the Qs and defined as simple functions of the Ps.

3. RESULTS

It has been checked that nine simple sums of the Ps, the R parameters defined in Table IV, were equivalent to the Qs; this is a minimum set since they are linearly independent. As the parametric dimension of the Ps is 11, this means that some part of them is not given from the Qs. Besides it can be checked that all the sums by row and by column of the partition in Table III are linear combinations of the Rs; this is a good point since they correspond to the two generating partitions of the complete partition and hence they are informed by the Qs. To be able to derive all the Ps from the Rs (equivalently from the Qs), only two well-chosen Ps are sufficient. There are only 15 couples of this kind among the 66 possible, for instance (P_{sC}, P_{sR}) or (P_{sO}, P_{nO}).

Fig. 2 shows the one-dimensional prior distributions (gammas(1, 1) + pseudo-data) of the Q and P parameters. Fig. 3 gives the posterior distributions of the Q and P parameters. Comparing priors and

Table IV. Rs Probabilities in Terms of Ps (See Section 2.2.4)

R_i	Sums of Ps
R_1	P_{yC}
R_2	P_{nC}
R_3	$P_{sC} + P_{dC}$
R_4	P_{rC}
R_5	$P_{nO} + P_{sO} + P_{dO}$
R_6	P_{rO}
R_7	$P_{nR} + P_{sR}$
R_8	$P_{nO} + P_{nR}$
R_9	$P_{sC} + P_{sO} + P_{sR}$

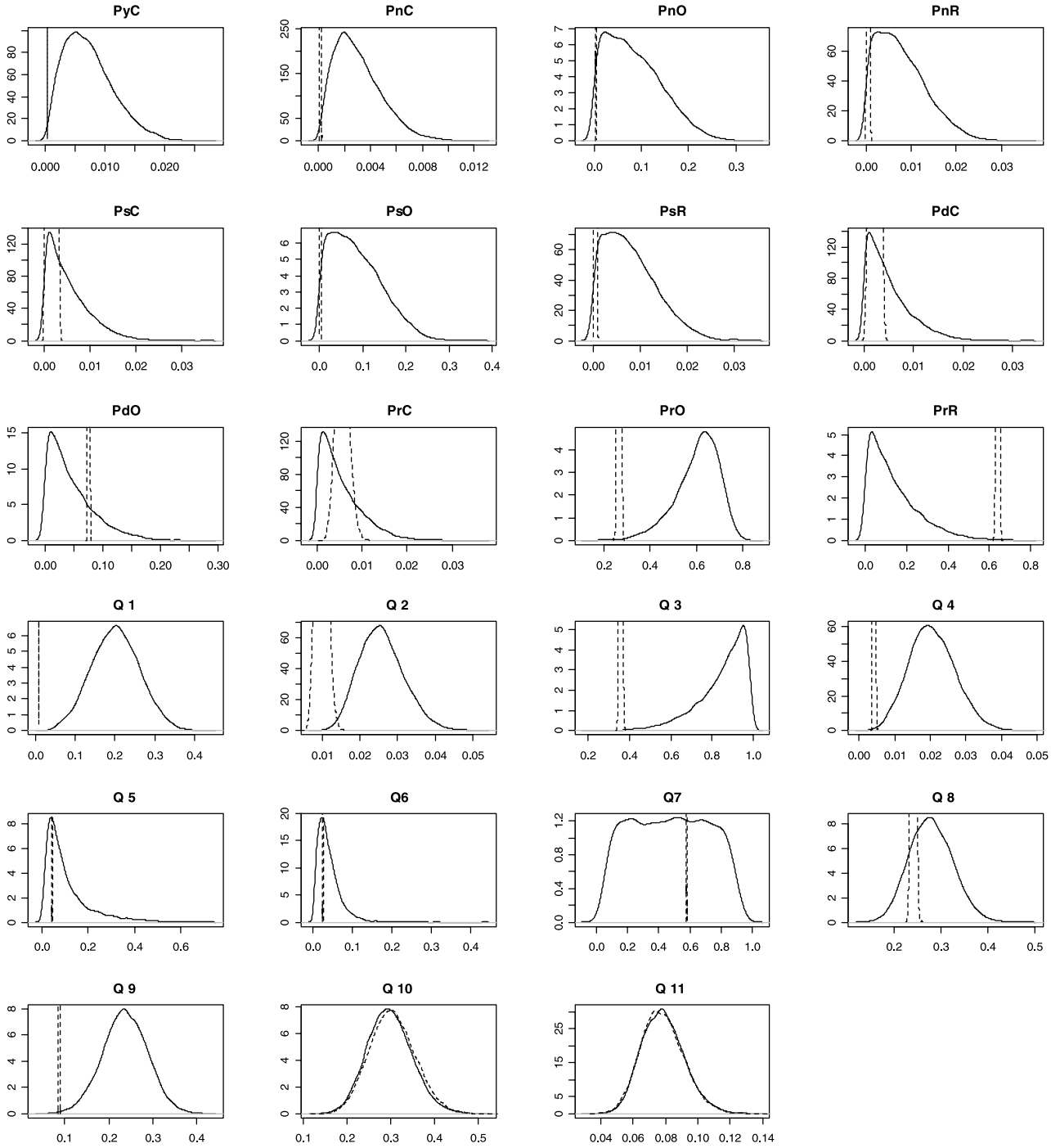

Fig. 2. Prior distributions of P and Q parameters (the dashed lines represent the posterior distributions for comparison). For each diagram, the x axis is associated with the possible values of the probability parameter and the y axis is associated with the density values.

posteriors, we conclude that the data provide a lot of information on the parameters with the exception of Q_{10} and Q_{11}, which seem to be informed only by the pseudo-data given by the experts (see Section 2.2.2). Some P posterior distributions are nearly multimodal because they cannot be informed from the data individually (Fig. 3 and Section 2.2.4). Fig. 4 gives the R posterior distributions and on the contrary we

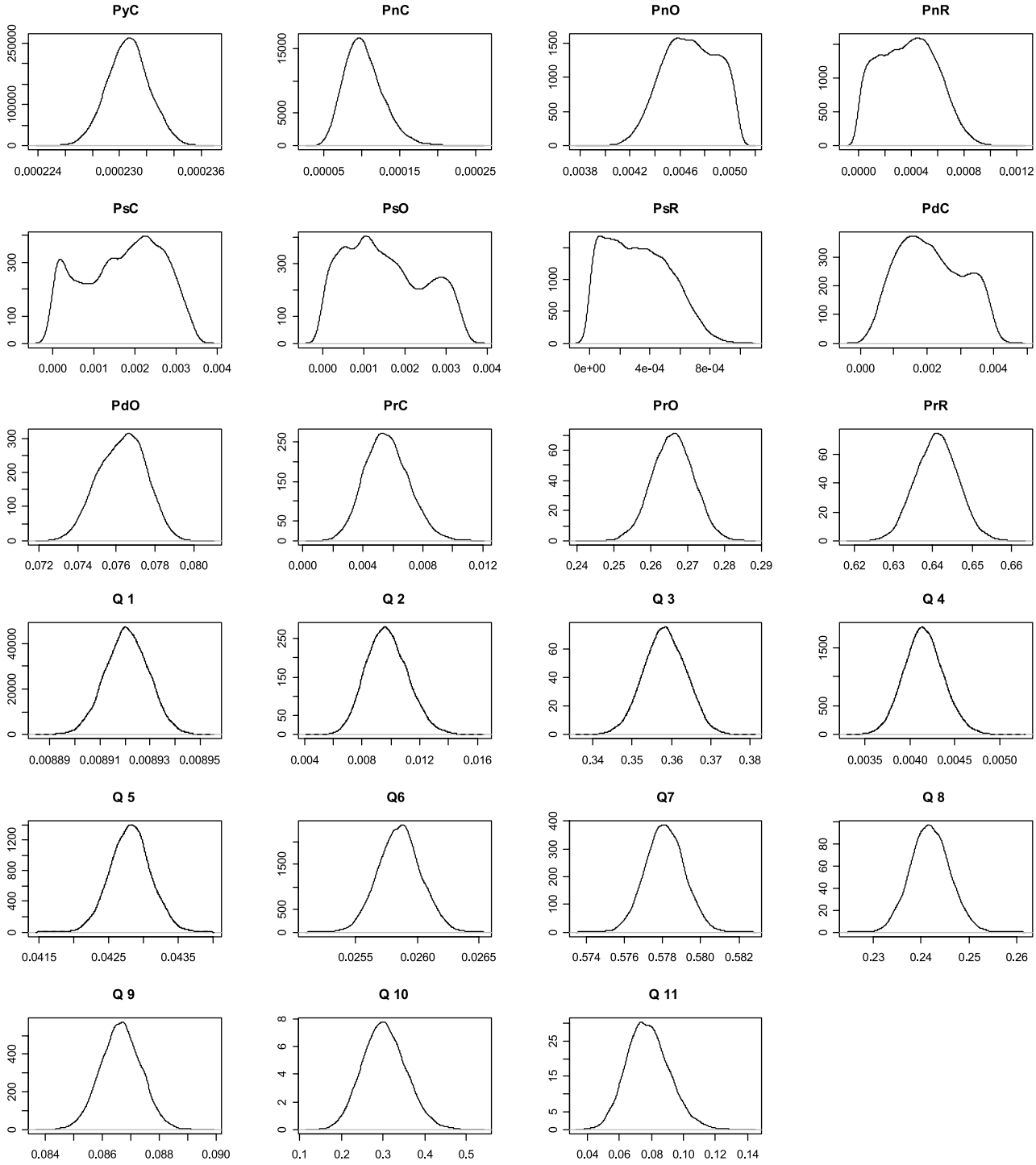

Fig. 3. Posterior distributions of the P and Q parameters (x-scales are magnified with respect to Fig. 2). For each diagram, the x axis is associated with the possible values of the probability parameter and the y axis is associated with the density values.

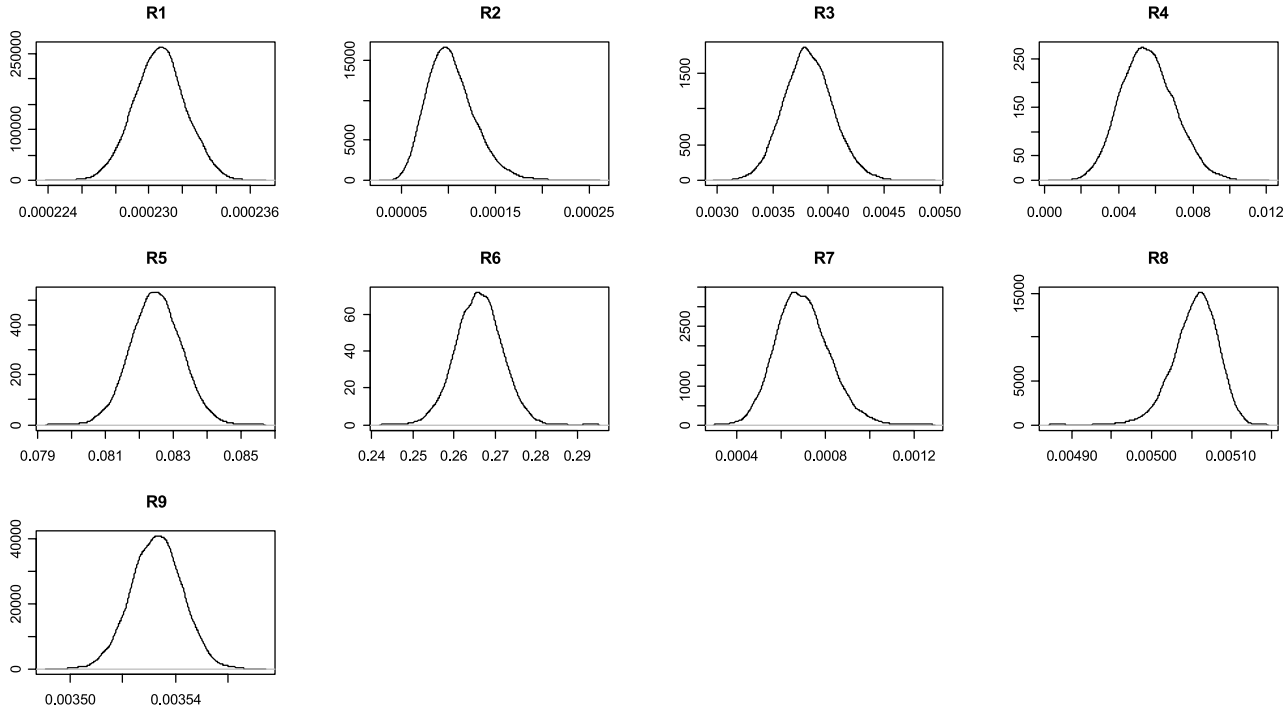

Fig. 4. Posterior distributions of the R parameters. For each diagram, the x axis is associated with the possible values of the probability parameter and the y axis is associated with the density values.

observed that they are all unimodal, well informed by the data sets. For the Q posterior distributions (Fig. 3), there are unimodal distributions summarized by 95% credible intervals in Table II.

Fig. 5 gives prior results from uninformative priors after excluding the pseudo-data versus the defined prior distributions. We observed that even if the experts give us little information, a lot of P distributions are wider; even the distribution of P_{rO} is inverted. Indeed, a Dirichlet with all parameters equal to one is the uniform distribution on the space of the probability parameters. For the Qs, we observed that the expert opinions bring a lot of information on the four probabilities they informed, Q_2, Q_8, Q_{10}, and Q_{11}; however, as all Qs are linked, the expert opinion also provides information on other Qs, for example, Q_4. Also, the marginal distribution of Q_7 is more uniform after the pseudo-data are added compared to the uninformative priors. This is an efficient way to discuss with the experts about the information they give, clarifying how it propagates in the model, especially into the burden-of-illness parameters, which they did not directly provide, but are modified by the mutual dependence of all parameters.

Fig. 6 gives posterior results from noninformative priors and data (no pseudo-data) versus posterior distributions of the Q and P parameters to observe the influence of expert opinions on the burden-of-illness estimates. Some parameter distributions are almost identical, but others are modified; among them, the distribution of Q_2, which is pulled to the right by the expert opinion. Fig. 7 focuses on distributions of Q_2 based on the information introduced for its estimation. The data introduced through the partition (excluding the English study that informs Q_2 directly) poorly inform Q_2 (note the comparison of the solid line and the dashed line). When only the English study is included, a binomial likelihood and a beta(1,1) prior on the Q_2 parameter are chosen. Nevertheless, other data pull the mode of Q_2 to the right using all the data but the pseudo-data. What is marked in this figure is that expert opinion—through pseudo-data—pulls the Q_2's distribution to the right because the expert opinion was a 95% variation interval for Q_2 between 1% and 3%. As expected, the posterior distribution we proposed offers a sensible compromise between expert opinion and the observed English data. The strength of the English data (large sample size, $n_2 = 4{,}026$, compared to

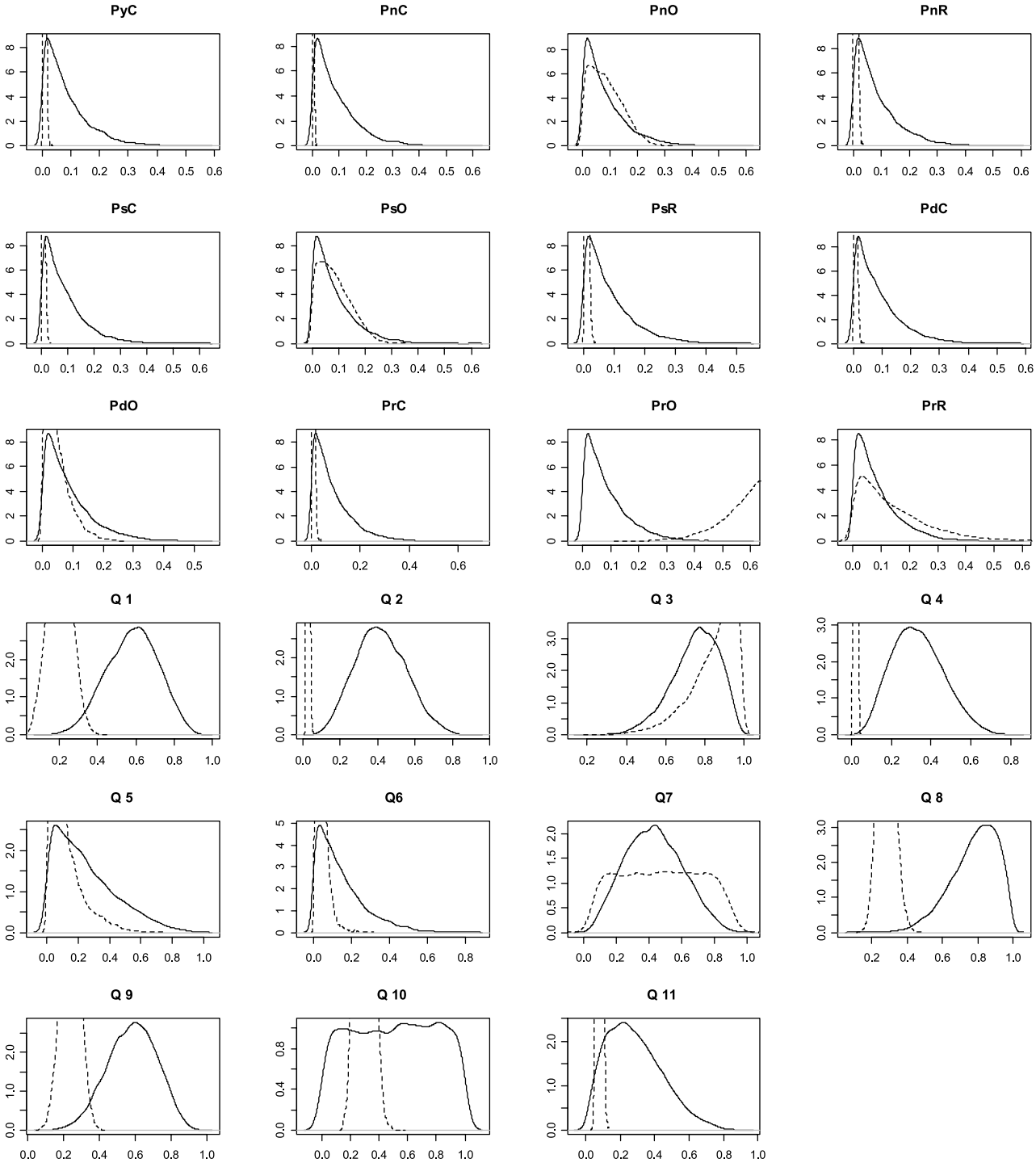

Fig. 5. Noninformative prior distributions of the P and Q parameters (the dashed lines represent the prior distributions with pseudo-data for comparison). For each diagram, the x axis is associated with the possible values of the probability parameter and the y axis is associated with the density values.

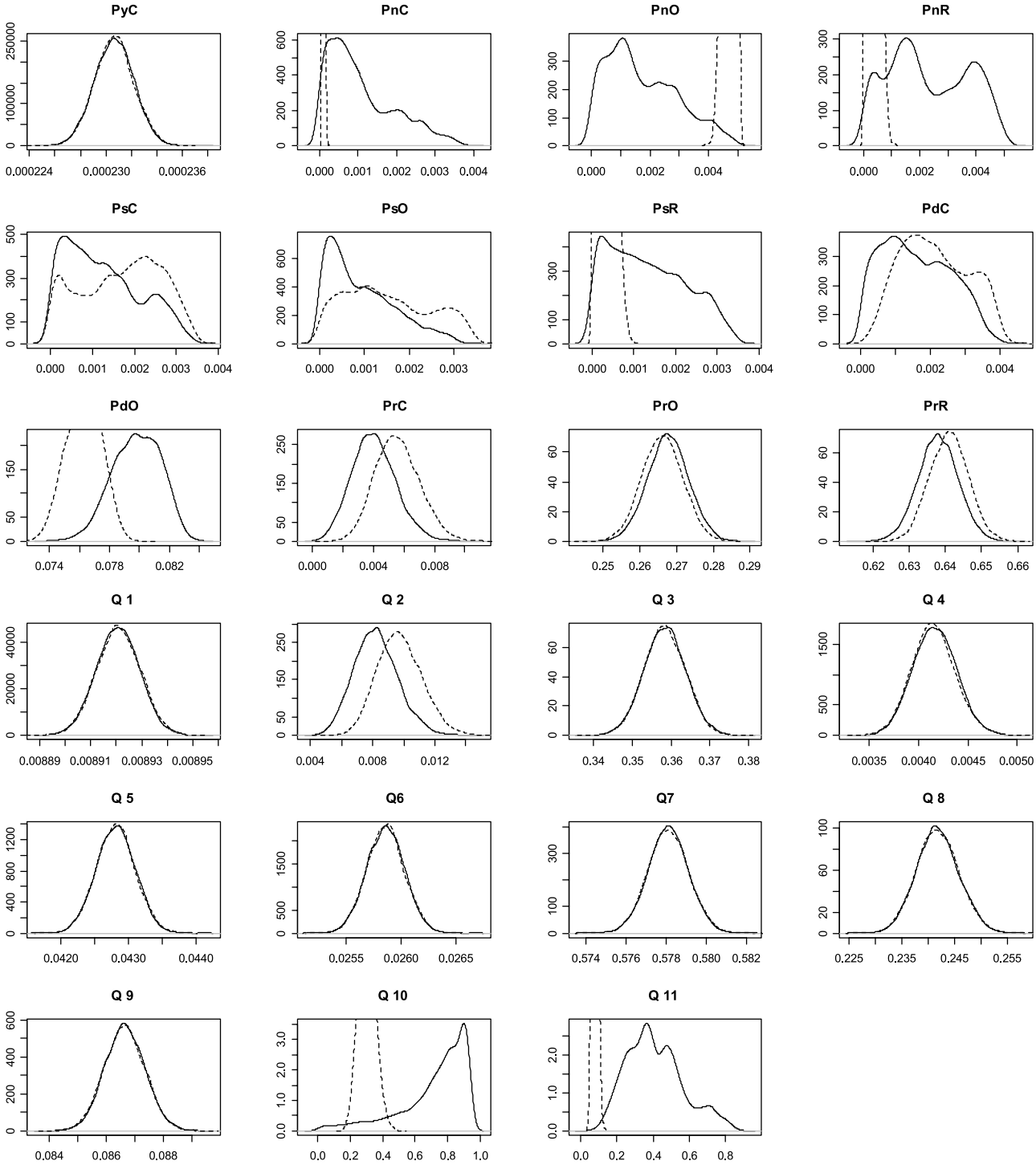

Fig. 6. Posterior distributions of the P and Q parameters without pseudo-data (the dashed lines represent the posterior distributions with pseudo-data for comparison). For each diagram, the x axis is associated with the possible values of the probability parameter and the y axis is associated with the density values.

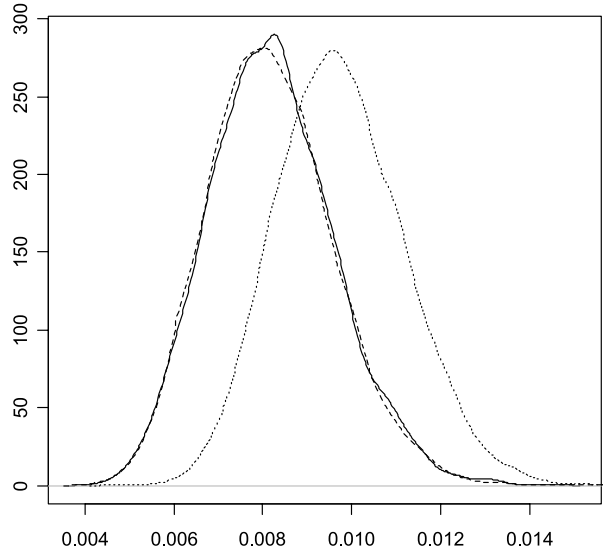

Fig. 7. Posterior distributions of Q_2 when only the English study is introduced (dashed line), when all data are introduced except pseudo-data (solid line), and when all data are introduced (dotted line). The x axis is associated with the possible values of the probability parameter and the y axis is associated with the density values.

pseudo-data, $n_{2,p} = 685$ [see Table II]) leads to a posterior 95% credible interval of [0.72%; 1.27%] from our posterior distribution (see Table II).

4. DISCUSSION

In this article, we synthesized complex published data to estimate campylobacteriosis prevalence and to simultaneously obtain estimates that guide the disease pathway. This synthesis is crude compared to what can be done in many supervised diseases, such as HIV infection.[7,8] However, it is important to gather heterogeneous data that examine the prevalence of campylobacteriosis in France. Our final estimate of the actual campylobacteriosis prevalence in France is between 0.72% and 1.27% (95% credible interval; see Table II). This result is in accordance with the English estimate (mean point estimate: 0.79%) but a little higher as shown in Fig. 7 due to the expectation of the experts for France (between 1% and 3%). Studies are at present ongoing on a large randomly selected population sample to study different parameters of interest such as consulting a physician, having a stool culture, and having a Campylobacter isolated. These data will allow to obtain more precise estimates of the different probabilities given in Table II.

Given the scarcity of data on this disease, we did not consider more complex solutions such as random hierarchical models taking temporal and spatial heterogeneity into consideration. As a consequence, the posterior distributions can appear to be too narrow compared to the probable heterogeneity of the data because some variability is neglected. Nevertheless, we propose a global estimate with accompanying uncertainty introducing expert opinions (Fig. 7).

Another possibility could have been to introduce bias parameters, for instance, to model the difference between the French and English situations as proposed in Albert *et al.*[2] But estimating even one of these parameters would be problematic given the available data.

One advantage of the proposed Bayesian approach is that the obtained operational evidence structure can be improved or corrected if new data are collected. Additional statistical analyses including future surveys on campylobacteriosis should be done to check and possibly to correct the present results. In this respect, data exclusion contributing to conflicts or extensions of modeling with more parameters could be considered, as in Presanis *et al.*[8] or in Turner *et al.*[23]

It has been checked with the expression of the R parameters (Table IV) that not all the P parameters were inferred from the available data used in Table I. Missing information can be completed focusing on one couple of P parameters. From it, some new collection of data can be proposed to obtain a complete inference. For example, if we choose the couple (P_{sC}, P_{sR}) as proposed in the first paragraph of Section 3, it seems convenient to suggest that among the individuals having a stool culture without Campylobacter research, the proportion of them not having acute gastroenteritis can be observed. This could give access to P_{sR}. Also as the knowledge of campylobacteriosis among acute gastroenteritis diseases is obtained from a Campylobacter research in stools, it seems that the second necessary survey would imply such researches among cases not yet investigated by the health system to obtain P_{sC}. Of course, *identifiability* is not sufficient and the precision of the estimation must also be considered. In the Bayesian framework, this can be done by the inspection of posterior distributions of the targeted parameters.

As shown in this study, the expert opinions provide a lot of information on some estimates. Expert opinions and data are assumed to be independent of each other, which may not be accurate. Nevertheless, the experts gave opinions on

campylobacteriosis. Therefore, in this context the Bayesian framework offers—via the posterior distribution—a synthesis of expert opinions and foreign data. In applications where data are sparse, it may be useful to acquire the opinions of many experts. This multiparameter evidence synthesis structure offers a convenient approach to discuss with experts. Given the multidependence between all the parameters (i.e., the Qs), the evidence structure shows the repercussions in the overall model introducing expert opinions in the marginal distribution of each parameter.

Additional data collection (notably the French data collection planned by the French Institute for Public Health Surveillance) is necessary to improve model fit. At the moment, however, the proposed approach makes a multiple data-source synthesis explicit, transparent, and open to sensitivity analysis; it also gives an indication of where new data must be collected.

REFERENCES

1. FAO/WHO. Hazard Characterization for Pathogens in Food and Water: Guidelines. Microbiological risk assessment series; no. 3. Geneva: WHO Library Cataloguing-in-Publication Data, 2003.
2. Albert I, Grenier E, Denis JB, Rousseau J. Quantitative risk assessment from farm to fork and beyond: A global Bayesian approach concerning food-borne diseases. Risk Analysis, 2008; 28(2):557–571.
3. Wheeler JG, Sethi D. Study of infectious intestinal disease in England: Rates in the community, presenting to general practice, and reported to national surveillance. British Medical Journal, 1999; 318:1046–1050.
4. Hardnett FP, Hoekstra RM, Kennedy M, Charles L, Angulo FJ. Epidemiologic issues in study and data analysis related to FoodNet activities. Clinical Infectious Diseases, 2004; 38(3):S121–S126.
5. Vaillant V, de Valk H, Baron E, Ancelle T, Colin P, Delmas M-C, Dufour B, Pouillot R, LeStrat Y, Weinberg P, Yougla E, Desenclos J-C. Foodborne infections in France. Foodborne Pathogens and Disease, 2005; 2(3):221–232.
6. Powell M, Ebel E, Schlosser W. Considering uncertainty in comparing the burden of illness due to foodborne microbial pathogens. International Journal of Food Microbiology, 2001; 69:209–215.
7. Goubar A, Ades AE, De Angelis D, McGarrigle CA, Mercer CH, Tookey PA, Fenton K, Gill ON. Estimates of human immunodeficiency virus prevalence and proportion diagnosed based on Bayesian multiparameter synthesis of surveillance data (with discussion). Journal of Royal Statistical Society A, 2008; 171:541–580.
8. Presanis AM, De Angelis D, Spiegelhalter DJ, Seaman S, Goubar A, Ades AE. Estimates of human immunodeficiency virus prevalence and proportion diagnosed based on Bayesian multiparameter synthesis of surveillance data (with discussion). Journal of Royal Statistical Society A, 2008; 171: 915–937.
9. Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. Methods for Meta-Analysis in Medical Research. Chichester UK: John Wiley & Sons, 2000.
10. Ades A, Cliffe S. Markov chain Monte Carlo estimation of a multiparameter decision model: Consistency of evidence and the accurate assessment of uncertainty. Medical Decision Making, 2002; 22:359–371.
11. Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian Approaches to Clinical Trials and Health-Care Evaluation. New York: Wiley, 2004.
12. Ades AE, Sutton AJ. Multiparameter evidence synthesis in epidemiology and medical decision-making: Current approaches. Journal of Royal Statistical Society A, 2006; 169:5–35.
13. Vaillant V, Baron E, de Valk H. Morbidité et mortalité dues aux maladies infectieuses d'origine alimentaire en France. Rapport InVS, France, 190p. [French], 2004.
14. Gallay A, Mégraud F. Etude de faisabilité d'une surveillance des infections à Campylobacter—Enquête auprès des laboratoires hospitaliers et privés 2000–2001. Rapport InVS, France, 2001.
15. Frosst GO, Majowciz SE, Edge VL. Factors associated with the use of over-the-counter medications in cases of acute gastroenteritis in Hamilton, Ontario. Canadian Journal of Public Health, 2006; 97: 489–493.
16. Anonymous. Acute Gastroenteritis in Ireland, North and South—A Telephone Survey, Safefood Report, Ireland, 2003.
17. Anonymous. National Gastroenteritis Survey 2001–2002, Australia, 2002.
18. Kuusi M, Aavitsland P, Gondrosen B, Kapperud G. Incidence of gastroenteritis in Norway—A population-based survey. Epidemiol Infect, 2003; 131: 591–597.
19. Carlin BP, Louis TA. Bayesian Methods for Data Analysis, Third Edition. Texts in Statistical Science Series. London: Chapman & Hall, 2008.
20. Gilks WR, Richardson S, Spiegelhalter DJ. Markov Chain Monte Carlo in Practice. London: Chapman & Hall, 486p., 1996.
21. Plummer M. Jags, 2008. Available at: http://www-fis.iarc.fr/m̃artyn/software/jags/.
22. Thomas A, O Hara B, Ligges U, Sturtz S. Making BUGS open. R News, 2006; 6:12–17.
23. Turner RM, Spiegelhalter DJ, Smith, GC, Thompson SG. Bias modelling in evidence synthesis. Journal of Royal Statistical Society A, 2009; 172:21–47.