Human salmonellosis: Estimation of dose-illness from outbreak data

Peer-reviewed author version

BOLLAERTS, Kaatje; AERTS, Marc; FAES, Christel; Grijspeerdt, Koen; Dewulf, Jeroen & Mintiens, Koen (2008) Human salmonellosis: Estimation of dose-illness from outbreak data. In: RISK ANALYSIS, 28(2). p. 427-440.

DOI: 10.1111/j.1539-6924.2008.01038.x
Handle: http://hdl.handle.net/1942/8256

Human Salmonellosis:

Estimation of Dose-illness from Outbreak Data

K. Bollaerts, M. Aerts, C. Faes, Grijspeerdt, K., Dewulf, J., Mintiens, K.

Abstract

The quantification of the relationship between the amount of microbial organisms ingested and a specific outcome such as infection, illness or mortality is a key aspect of quantitative risk assessment. A main problem in determining such dose-response models is the availability of appropriate data. Human feeding trials have been criticized because only young healthy volunteers are selected to participate and low doses, as often occurring in real life, are typically not considered. Epidemiological outbreak data are considered to be more valuable, but are more subject to data uncertainty. In this paper, we model the dose-illness relationship based on data of 20 *Salmonella* outbreaks, as discussed by the *World Health Organization*. In particular, we model the dose-illness relationship using Generalized Linear Mixed Models and fractional polynomials of dose. The fractional polynomial models are modified to satisfy the properties of different types of dose-illness models as proposed by Teunis et al [1]. Within these models, differences in host susceptibility (susceptible versus normal population) are modeled as fixed effects whereas differences in serovar type and food matrix are modeled as random effects. In addition, two bootstrap procedures are presented. A first procedure accounts for stochastic variability whereas a second procedure accounts for both stochastic variability and data uncertainty. The analyses indicate that the susceptible population has a higher probability of illness at low dose levels when the combination pathogen-food matrix is extremely virulent and at high dose levels when the combination is less virulent. Furthermore, the analyses suggest that immunity exists in the normal population but not in the susceptible population.

Keywords: Human Salmonellosis; Outbreak studies; Dose-illness; Fractional polynomials; Generalized Linear Mixed Models; Data uncertainty

1 Introduction

Salmonellosis, the illness from *Salmonella* infection, is one of the most frequently occurring foodborne diseases worldwide [2]. Global estimations vary between 14 and 120 per 100000 people [3]. The majority of the cases is due to *Salmonella* Enteritidis and *Salmonella* Typhimurium infections, which comprised almost 80% of the total number of *Salmonella* infections in Belgium in 2005 [4]. Salmonellosis is characterized by fever, stomach cramps, and diarrhea. Symptoms develop 8 hours to 3 days post-ingestion and last 4 to 7 days. Most cases are self-limiting. The degree to which a person becomes sick depends on his or her health status and the number and virulence of *Salmonella* spp. ingested. In general, the poorer the individuals health and the more *Salmonella* ingested, the greater the probability for serious illness and death. In the United States only, the yearly number of illnesses and hospitalizations due to foodborne Salmonellosis are estimated to be 1,3 million and 15,5 thousand, respectively. The number of deaths caused by foodborne Salmonellosis is estimated to be 553 per year, which is 30,6% of the total number of yearly deaths caused by known foodborne pathogens [2].

An important aspect of quantifying microbial risk is the assessment of the dose-response relationship, which is the relationship between the amount of microbial organisms ingested and a specific outcome, like infection, illness or even mortality. Different sources of heterogeneity in dose-response are known to exist. A first important source of heterogeneity are differences in host susceptibility [5]. Typically, newborns, young children, pregnant women, elderly and immunocompromised persons belong to the susceptible subpopulation being associated with increased risk. A second source of heterogeneity are differences in pathogen virulence [6]. For *Salmonella*, about $2,300$ serovars are known to exist, of which *S.* Enteritidis and *S.* Typhimurium are of the most prevalent. Among these serovars, large differences in virulence occur. In addition, differences in food matrix cause heterogeneity in dose-response relationships as well. D'Aoust [6] found that the infectivity of a pathogen is higher when a fatty food matrix is involved.

To model dose-response relationships, data from human feeding trials as well as from outbreak studies can be used. The most extensive human feeding trial of *Salmonella* was conducted by McCoullough & Eisele ([7], [8], [9]) and has often been used to develop dose-response models (e.g., [10], [11], [12]). However, human feeding trials have been seriously criticized [5]. One of the major shortcomings is the selection of young healthy volunteers. These volunteers are likely to have

high resistance to *Salmonella*, eventually resulting in substantive underestimation of the general public health risk. Second, low doses, although the most commonly occurring in real life, are not considered in such human feeding trials. As a consequence, estimation of the probability of the adverse outcome at low dose levels is problematic since it is bound to rely on extrapolation. Furthermore, in most human feeding trials, only one specific combination of pathogen, host and food matrix is considered whereas there is much more variability in these factors in real life. As an alternative to human feeding trial data, epidemiological data from outbreak studies can be used as well to model dose-response relationships. Clearly, epidemiological data refer to real-life situations, involving the whole population, different types of pathogens and food matrices and a wide range of possible doses, including low doses. However, compared to experimental data, epidemiological data are more subject to data uncertainty. For instance, estimates of dose are very uncertain because it is hard to accurately quantify the amount of contaminated food consumed. Similarly, it is very hard to exactly know the total exposed population. Nevertheless, outbreak data are considered to be the most valuable in order to model dose-response [10].

An example of modeling dose-response using outbreak data can be found in a report on risk assessment of *Salmonella* from the *World Health Organisation* (WHO) [10]. In this study, dose-illness is modeled using a beta-poisson model and the effect of data uncertainty is investigated by sampling from specific uncertainty distributions and refitting the beta-poisson model. Based on this study, it is concluded that 'there is insufficient evidence to conclude that some segments of the population have a higher probability of illness' [10]. However, beta-poisson models are developed to reflect the biological process of infection, not illness. Furthermore, heterogeneity is not accounted for. In the current study, we extend on the analysis of the WHO [10] and investigate different dose-illness models as proposed by Teunis et al [1]. We use Generalized Linear Mixed Models (GLMMs) ([13], [14]) and fractional polynomials of dose ([15]) to investigate dose-illness. In particular, we introduce modifications of fractional polynomials of dose to satisfy the properties of the different types of dose-illness models proposed by Teunis et al [1]. Within these models with modified fractional polynomial of dose, heterogeneity due to differences in host susceptibility, serovar type and food matrix is taken into account. Finally, since outbreak data are strongly subject to data uncertainty, bootstrap confidence intervals of the estimated dose-illness relationship are estimated while accounting for data uncertainty on top of the stochastic variability.

The remainder of this paper is organized as follows; in Section 2, a description of the epidemiological data on *Salmonella* is given. Section 3 contains a (methodological) discussion on dose-illness models. In particular, GLMMs and modified fractional polynomials of dose are introduced. Application of these models with the epidemiological data on *Salmonella* is given in Section 4. In Section 5, bootstrap procedures are applied to account for stochastic variability as well as data uncertainty. Finally, some concluding remarks are formulated in Section 6.

2 Data

Data from outbreak studies reported in the risk assessment on *Salmonella* by the WHO [10] are used to estimate the *Salmonella* dose-illness relationship. In this report, 33 outbreak studies found in literature are summarized in as much detail as possible. Of these outbreak studies, 20 are selected by the WHO (2003) to be used in their quantitative risk assessment because they are sufficiently well documented. We use the data of the same 20 outbreak studies in our analyses.

For these 20 outbreak studies, the WHO (2003) report contains information with respect to the ingested dose D ($\log_{10}$ CFU), the number of exposed persons N, the number of ill persons Y, serovar type t and food matrix m. In addition, distinction is made between persons belonging to the normal and the susceptible subpopulation (for definitions, see [10]). Finally, uncertainty distributions are given for D, N and Y. These distributions are derived from additional information available on the different outbreaks. In case no such information is available to characterize uncertainty, distributions are defined of which the lower limit (resp. upper limit) is the 25% underestimate (resp. 25% overestimate) of the reported values under consideration. For more detailed information, the reader is referred to the WHO (2003) report [10]. A summary of the data is given in Table **??**. The outbreak number Nr and the cluster number i (will be discussed later) are displayed as well. For D, N and Y, the expected values based on the uncertainty distributions as defined by the WHO (2003) are given. In Figure 1, a visual representation of the data is given by means of a bubble plot of the proportion of ill subjects $p = Y/N$ as a function of dose, with the area of the bubbles being proportional to the number of exposed subjects N. In this figure, observations on normal subjects are indicated using light gray colored bubbles with a dot indicating the middle point of the bubble whereas observations on susceptible subjects are indicated using bubbles with

the darkest gray color and a star indicating the middle point.

As can be derived from the table, the *Salmonella* serovars *S.* Typhimurium and *S.* Enteritidis are the most prevalent in the outbreak studies under consideration. All the other *Salmonella* serovars occur only once. Regarding food matrix, many different types of food matrices are involved but no big differences in occurence are recorded. Furthermore, there are much more normal subjects than susceptible subjects involved in these outbreak studies. Finally, large differences in the number of exposed subjects involved in the different outbreaks can be observed, with the number of exposed subjects ranging from 1 to 7572. As can be clearly seen in Figure 1, the number of exposed subjects is much smaller at high doses compared to low doses. As such, a proper statistical analysis should account for the differences in the number of exposed subjects as well as for heterogeneity caused by differences in host susceptibility, in serovar type and in food matrix. Other sources of heterogeneity (e.g., microflora of the food, differences in serovar strains) exist as well. Unfortunately, for the outbreak studies under consideration (WHO, 2003), no information is available regarding these sources and as such, these sources can not be taken into account in the current analyses.

[Figure 1 about here.]

[Table 1 about here.]

3 Methodology

3.1 Dose-Illness models

Dose-illness models are still not well determined. So far, most attention is given to dose-infection models with the most popular one being the beta-poisson model [16]. This model is developed to reflect the biological process of infection, which may result from survival of a single viable pathogen and yields monotonically increasing functions of dose bounded between zero and one. Teunis et al [1] advocate that the use of the beta-poisson model to model illness as a function of dose is questionable and introduce a multiple-stage model instead, which can be graphically represented as in Figure 2.

[Figure 2 about here.]

Following this model, exposure to pathogens might lead to infection and infection might lead to illness. However, note that exposure does not imply infection, nor does infection imply illness. This model is called a multiple-stage (or multiple-barrier) model since illness does not occur without infection. Hence, the probability of illness given dose equals

$$\pi(\text{ill}|\text{dose}) = \pi(\text{ill}|\text{infection,dose})\pi(\text{infection}|\text{dose}). \tag{1}$$

Whereas π(infection|dose) is typically assumed to be a monotonically increasing function of dose bounded between zero and one (e.g. beta-poisson model), some experimental evidence seems to indicate different relationships for π(ill|dose) [1]. To explain these differences, Teunis et al [1] explore three different alternatives for π(ill|infection,dose), namely (1) the increasing probability model for illness given infection, (2) the decreasing probability model and (3) the constant probability model. They assume that the probability of becoming ill depends on the duration of infection. The infection episode starts at time $t = 0$ when the pathogens start growing and ends at time $t = \tau$ when the pathogens have been succesfully removed by the host defense mechanisms. As such, the length of the infection period reflects the balance between pathogen growth and host defenses. Teunis et al [1] argue that the length of the infection period may be dose-dependent. Starting from the assumption that during infection the host has a certain probability of becoming ill and using a Gamma distribution for the duration of infection τ, Teunis et al [1] derive that the probability of illness given infection equals

$$\pi(\text{ill}|\text{infection, dose}) = 1 - (1 + \lambda)^{-r} \tag{2}$$

with $r > 0$ being the shape parameter of the Gamma distribution and with λ being the integral over duration time t of the probability function for illness in an infected person. Assuming that λ varies with dose, Teunis et al [1] explore three different alternatives:

1. λ increases linearly with dose D or $\lambda = \eta$D. This implies that π(ill|infection,dose) is a monotonically increasing function of dose bounded between zero and one. As such, given the common assumption that π(infection|dose) is a monotonically increasing function of dose bounded between zero and one, π(ill|dose) is also monotonically increasing as a function of dose with the same boundaries. Such a situation may arise when high initial doses of pathogens slow down the hosts defense mechanisms or, in other words, the higher the initial dose, the longer the episode of infection τ.

2. λ decreases with dose D or $\lambda = \eta/\mathrm{D}$. This implies that π(ill|infection,dose) is a monotonically decreasing function of dose bounded between zero and one. Given the common assumptions for π(infection|dose), it follows that π(ill|dose) is monotonically unconstrained with the probability of illness being zero for zero dose and infinitely large dose levels. Such a situation may arise when high initial doses of pathogens elicit strong defensive reactions of the host or, in other words, the higher the initial dose, the shorter the episode of infection τ.

3. λ is dose-independent or $\lambda = \eta$ such that π(ill|infection,dose) = π(ill|infection). Hence, given the common assumptions for π(infection|dose), it follows that π(ill|dose) is a monotonically increasing function of dose that is bounded by zero and reaches the asymptote of π(ill|infection) < 1 for infinitely large dose levels. Such a situation suggests no facilitating nor inhibiting effects of dose on hosts defense mechanisms.

Teunis et al [1] provide real-data examples illustrating the three different probability models of illness given infection (data from [7], [17], [18], respectively). In this paper, for illustrative purposes, fictive examples are given (Figure 3). Clearly, different types of dose-illness relationships are possible with the properties of these relationships depending on the effect of dose on the probability of illness given infection. To summarize, the increasing probability model of illness given infection ($\lambda = \eta\mathrm{D}$) gives rise to a monotonically increasing dose-illness model bounded between zero and one (DI-I), the decreasing probability model ($\lambda = \eta/\mathrm{D}$) to a monotonically unconstrained dose-illness model with the probability of illness being zero for zero dose and infinitely large dose levels (DI-II) and finally, the constant probability model ($\lambda = \eta$) gives rise to a monotonically increasing dose-illness model bounded between zero and π(ill|infection) being some constant $c < 1$ (DI-III). In the next section Generalized Linear Mixed Models (GLMMs) and fractional polynomials of dose are introduced. Different constraints on the parameters of the fractional polynomials are imposed in order to satisfy the different properties of the three types of dose-illness models.

[Figure 3 about here.]

3.2 GLMMs and fractional polynomial of dose

Most dose-response models fit within the framework of Generalized Linear Models (GLMs) [19]. GLMs represent a class of fixed effects regression models suited to model non-normal data (e.g.

counts, dichotomous data). However, fixed effects models, which assume that all observations are independent, are not appropriate to model clustered data. Clusters cause heterogeneity with, typically, observations belonging to the same cluster being more homogeneous compared to observations belonging to different clusters. For instance, outbreak studies in which the same food matrix is involved are more likely to be similar compared to outbreak studies in which different food matrices are involved. In case the observed clusters (e.g. clusters consisting of outbreak studies sharing the same food-matrix) are thought to represent a population of clusters, GLMs extended with cluster-specific random effects can be used. These extended models are called Generalized Linear Mixed Models (GLMMs) (see e.g. [13], [14]). For instance, GLMMs with food matrix-specific random effects are suited to model the outbreak data since the food matrices involved in the recorded outbreak studies are only a sample of all possible food matrices. More formally, an example of a GLMM for the dose-illness data under consideration is

$$Y_{ij} \sim Binomial(N_{ij}, \pi_{ij})$$

$$g(\pi_{ij}) = \beta_0 + \beta_1 D_{ij} + \beta_2 S_{ij} + b_i = \eta_{ij} \tag{3}$$

with Y_{ij} being the number of ill subjects from observation j belonging to cluster i. The Y_{ij} are assumed to follow a binomial distribution with parameters N_{ij}, the number of exposed subjects, and π_{ij}, the probability of illness. Then, the latter probality is modelled as a linear function of the known covariates dose (D_{ij}) and susceptibility of the population (S_{ij}) with corresponding fixed effects β_1 and β_2 and as a function of cluster-specific random intercepts $b_i \sim N(0, \sigma^2)$. The linear function, also called linear predictor η_{ij}, can take any value ranging from $-\infty$ to $+\infty$ whereas the probability π_{ij} is restricted to the interval $[0, 1]$. Therefore, the probability π_{ij} is transformed to the interval $[-\infty, +\infty]$ using the link function g. The logit link, the complementary log-log link and the probit link are the most commonly used link functions g to transform probabilities.

In most GLMMs, the linear predictor η contains only indicator variables for discrete covariates and conventional polynomials, mostly of linear or quadratic order, for continuous covariates (like in expression (3)). However, it is recognized that the conventional polynomials do not always provide a flexible and good fit to the data. To enhance flexibility, Royston and Altman [15] introduced fractional polynomials, which are a set of parametric models offering a wide range of functional forms including the conventional polynomials. Fractional polynomial models are already been

succesfully used in diverse application settings (e.g. [20], [**?**], [21], [22]). A fractional polynomial of degree m for a continuous covariate x subject to the constraint $x > 0$, is defined as

$$f(x; \boldsymbol{\beta}, \mathbf{p}, m) = \sum_{r=0}^{m} \beta_r H_r(x) = \eta \tag{4}$$

with $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_m)$ being a vector of coefficients and $\mathbf{p} = (p_0, p_1, p_2, \ldots, p_m)$ a vector of powers with $p_0 \equiv 0$ and $H_0(x) \equiv 1$ representing the intercept. The powers $p_1 \leq p_2 \leq \cdots \leq p_m$ can be positive or negative integers or fractional powers. $H_r(x)$ is a transformation on a continuous variable x defined as

$$H_r(x) = \begin{cases} x^{p_r} & \text{if } p_r \neq p_{r-1} \\ H_{r-1}(x) \times \ln(x) & \text{if } p_r = p_{r-1} \end{cases}$$

and with $x^0 \equiv \ln(x)$.

Some examples of fractional polynomials of degree $m = 2$ given in Table 2 will make this clear. As can be seen, fractional polynomials can take a wide range of functional forms. Example 1 of this table illustrates that conventional polynomials (in this case, of quadratic order) are special instances of fractional polynomials. Examples 3 and 5 illustrate what happens if both powers are equal whereas example 4 and 5 illustrate what happens if powers are zero. An example of a fractional polynomial of degree $m = 3$ is $\beta_0 + \beta_1 x^{-1} + \beta_2 x^{-1} \ln(x) + \beta_3 x^2$ having powers $\mathbf{p} = (-1, -1, 2)$. However, following Royston and Altman [15], models of degree $m = 2$ or lower with powers $\mathbf{p}$ selected from a fixed set $\mathcal{R}^m$ with $\mathcal{R} = \{-2, -1, -0.5, 0, 0.5, 1, 2, \ldots, \max(3, m)\}$ are sufficiently flexible to cover most practical cases adequately. Of course, other candidate powers can be considered as well. Then, given a degree m and a fixed set $\mathcal{R}^m$, all possible models are fitted using maximum likelihood estimation and the model with the lowest deviance is selected.

[Table 2 about here.]

However, fractional polynomial models do not inherently satisfy the properties of the dose-illness models described earlier. In case of a monotonically increasing dose-illness relationship bounded between zero and one (DI-I), a fractional polynomial of dose D of degree $m = 2$ can be easily modified to satisfy these properties as follows

$$f(D; \boldsymbol{\beta}, \mathbf{p}, 2) = \beta_0 + \beta_1 D^{p_1} + \beta_2 D^{p_2} \tag{5}$$

with $p_1 < 0$, $p_2 \geq 0$, $\beta_1 < 0$ and $\beta_2 > 0$. These constraints yield a monotonically increasing linear function $\beta_0 + \beta_1 D^{p_1} + \beta_2 D^{p_2}$ bounded between $-\infty$ and $+\infty$. As such, $g^{-1}(\beta_0 + \beta_1 D^{p_1} + \beta_2 D^{p_2})$, with g^{-1} being the inverse of the link function g, is monotonically increasing and bounded between zero and one, which are the properties of DI-I. In case of a monotonically unconstrained dose-illness relationship with the probability of illness being zero for zero dose and infinitely large dose levels (DI-II), fractional polynomials are defined as in (5) with $p_1 < 0$, $p_2 \geq 0$, $\beta_1 < 0$ and $\beta_2 < 0$. These constraints imply a monotonically unconstrained function being equal to $-\infty$ for zero dose and infinitely large dose levels. As such, $g^{-1}(\beta_0 + \beta_1 D^{p_1} + \beta_2 D^{p_2})$ displays the properties of DI-II. Finally, a fractional polynomial of degree $m = 1$ can be modified to satisfy the properties of a monotonically increasing dose-illness model bounded between zero and some constant $c < 1$ (DI-III) as follows

$$f(D; \beta, p, 1) = \beta_0 + \beta_1 D^{p_1} \tag{6}$$

with $p_1 < 0$ and $\beta_1 < 0$ and where $g^{-1}(\beta_0)$ is an estimate of c or $\pi(\text{ill}|\text{infection})$. Table 3 displays a summary of the different constraints on the fractional polynomials of dose in order to satisfy the properties of the three different dose-illness models that are proposed by Teunis et al [1]. As can be clearly seen, the three dose-illness models share the same constraints on p_1 and β_1. This reflects the obvious biological property that exposure is a prerequisite of illness. No illness without exposure and hence the probability of becoming ill is zero at zero dose, irrespective of the type of dose-illness model. However, the dose-illness models differ with respect to their behavior at infinitely large dose levels; the probability of illness equals 1 for DI-I, 0 for DI-II and $c < 1$ for DI-III.

[Table 3 about here.]

4 Application to *Salmonella* Outbreak Data

In this section, the introduced dose-illness models are applied to the epidemiological data on *Salmonella* while accounting for differences in host susceptibility, in food matrices and in serovar types. Differences in host susceptibility are modeled using fixed effects since the only information we have is whether persons belong to the susceptible or the normal subpopulation. Differences in food matrix are modelled using random effects since the recorded food matrices can be seen as a sample from the population of all possible food matrices. The same holds for differences in serovar types. However, there is perfect overlap between some food matrices and serovar types (see outbreak number 1, 16 and 17 in Table 1) yielding an overparametrized model when both food matrix-specific and serovar type-specific random effects are incorporated in the model. A sensible alternative for modeling differences in serovar type is by means of fixed effects using the categories *S.* Typhimurium, *S.* Enteritidis and others. Statistical analysis are conducted with fixed effects for susceptibility and serovar type and random effects for food matrix. However, these analyses yield non-significant effects for serovar type due to the strong nesting of food matrix within serovar types. Another possibility is to define outbreak-specific random effects. However, in that case, the number of clusters ($Nr = 1, ..., 20$) is almost the number of observations ($n = 23$). Because of data sparseness, we opt to specify random effects for the unique combinations of serovar type and food matrix. As such, the two sources of heterogeneity are taken into account but, however, they can not be separated. In total, 15 unique combinations of serovar type and food matrix ($i = 1, ..., 15$) are recorded (see also Table 1).

Hence, heterogeneity in dose-illness is accounted for using fixed effects for host susceptibility and random effects for the unique combinations of serovar type $\times$ food matrix. In particular, the following DI-I model is fitted

$$Y_{ij} \sim Binomial(N_{ij}, \pi_{ij})$$

$$g(\pi_{ij}) = \beta_0 + \beta_1 S_{ij} + b_i + \beta_2 D_{ij}^{p_1} + \beta_3 D_{ij}^{p_2} \times (1 - S_{ij}) + \beta_4 D_{ij}^{p_2'} \times S_{ij} \tag{7}$$

where D_{ij} is the ingested dose ($\log_{10}(\text{CFU}+1)$) for observation j for which the unique combination of serovar type $\times$ food matrix i has been recorded, S_{ij} is the indicator variable of host susceptibility ($S_{ij} = 1$ if population is susceptible, $S_{ij} = 0$ otherwise) and b_i are serovar type $\times$ food matrix-specific random intercepts with $b_i \sim N(0, d^2)$. A fractional polynomial of dose of degree $m = 2$

is fitted for each population with powers p_1 and p_2 for the normal and powers p_1' and p_2' for the susceptible population. However, computation time for fractional polynomial models increases exponentially with the number of powers. Therefore, the powers p_1 and p_1' are chosen to be common for both groups whereas the powers p_2 and p_2' are chosen to be group-specific. This way, differences in host susceptibility are modeled and computation time is still kept reasonable. Power p_1 is selected from $\mathcal{R}_1$ with $\mathcal{R}_1 = \{-2, -1.8, \ldots, -0.2, -0.1\}$ and powers p_2 and p_2' are selected from $\mathcal{R}_2^2$ with $\mathcal{R}_2 = \{0, 0.1, \ldots, 1.8, 2\}$. Constrained maximum likelihood is used to ensure the constraints on $\boldsymbol{\beta}$, i.e. $\beta_2 < 0$, $\beta_3 > 0$ and $\beta_4 > 0$. For DI-II, the same fractional polynomial model as in (7) is used except that different constraints on $\boldsymbol{\beta}$ are imposed, i.e. $\beta_2 < 0$, $\beta_3 < 0$ and $\beta_4 < 0$. Finally, for DI-III, the following model is fitted

$$Y_{ij} \sim Binomial(N_{ij}, \pi_{ij})$$

$$g(\pi_{ij}) = \beta_0 + \beta_1 S_{ij} + b_i + \beta_2 D_{ij}^{p_1} \times (1 - S_{ij}) + \beta_3 D_{ij}^{p_1'} \times S_{ij} \tag{8}$$

where D_{ij}, S_{ij} and b_i are defined as before and where the powers p_1 and p_1' are selected from $\mathcal{R}_1^2$. Again constrained maximum likelihood is used to ensure the constraints on $\boldsymbol{\beta}$, i.e. $\beta_2 < 0$ and $\beta_3 < 0$. For the three different types of dose-illness models, different link functions, i.e. the logit link, the probit link and the complementary log-log link, are considered.

The best fitting models within the three types of dose-illness models are summarized in Table 4. Goodness-of-fit is measured by means of Aikaike's Information Criterion with small sample size correction (AICC) with the smaller its value the better the fit [23]. As can be seen, DI-III with the complementary log-log link, $p_1 = -2.25$ and $p_1' = -0.25$ fits the data best. It is investigated whether this model can be simplified. First, the null hypothesis of zero-variance for the random intercepts d^2 is tested with a mixture of $\mathcal{X}^2$ distributions as reference distribution since the null hypothesis lies at the boundary of the parameter space [13]. This test yields a significant result $[P(\mathcal{X}_{0:1} > 2.14) = 0.02]$ meaning that the heterogeneity caused by differences in the combination food matrix $\times$ serovar type are substantial. Finally, since all fixed effects are significant as well, simplification of the model is not recommended.

[Table 4 about here.]

Table **??** gives parameter estimates and standard errors of the fixed effects $\boldsymbol{\beta}$ and the variance component d^2 of model (8). Empirical bayes estimates ([14]) of the serovar type $\times$ food matrix-

specific random intercepts b_i and the corresponding standard errors are given in Table **??**. In addition, this table contains serovar type $\times$ food matrix-specific estimates of π(ill|infection) for the normal and the susceptible population, calculated using $g^{-1}(\beta_0 + b_i)$ and $g^{-1}(\beta_0 + \beta_1 + b_i)$, respectively. Remind that π(ill|infection) is the asymptote for infinitely large dose levels. Corresponding standard errors are calculated using the delta method. As can be seen, π(ill|infection) for the susceptible population (although mathematically < 1) is estimated to be virtually one and hence, the standard errors are undefined. For the normal population, π(ill|infection) differs strongly depending on the combination of serovar type $\times$ food matrix with values ranging from 0.303 to 0.997. These findings suggest that immunity exists in the normal population but not in the susceptible population.

Finally, graphical representations of the marginal simulation-based dose-illness relationship [14] for the normal and susceptible population are given in Figure 4(a). However, random effects models have an interpretation conditional on the random effects implying that a marginal representation is less appropriate. Therefore, the (serovar type $\times$ food matrix)-specific dose-response relationships are calculated as well. The latter are displayed in Figure 4(b). These figures suggest a higher probability of illness for the susceptible population compared to the normal population. However, caution is needed since the variability in estimation of the dose-response relationship and the effect of data uncertainty are not yet investigated. This is addressed in the next section.

[Table 5 about here.]

[Table 6 about here.]

[Figure 4 about here.]

5 Assessing Uncertainty

Uncertainty regarding the estimated dose-illness model is expressed by means of confidence intervals. First, a 1-stage bootstrap procedure is used to estimate confidence intervals taking into account stochastic variability only. Second, a 2-stage bootstrap procedure is constructed to estimate confidence intervals that reflect both stochastic variability as well as data uncertainty.

Denote the original sample summarized in Table 1 as $\{(D_j, S_j, N_j, Y_j, t_j, m_j)\}_{j=1}^{23}$. In the 1-stage bootstrap procedure, a bootstrap sample $\{(D_j, S_j, N_j, Y_j^*, t_j, m_j)\}_{j=1}^{23}$ is generated by sampling binomial observations $Y_j^* \sim \text{Binomial}(N_j, \hat{\pi}_j)$ with $\hat{\pi}_j = Y_j/N_j$, for $j = 1, ..., 23$. On this bootstrap sample, DI-III with cloglog-link, $p_1 = -2.25$ and $p_1' = -0.25$ is refitted. This process is repeated B times, yielding B different estimated dose-illness curves $f^*(\widehat{\boldsymbol{\beta}}, \mathbf{X})$. In order to estimate the 90% pointwise confidence interval for $f(\widehat{\boldsymbol{\beta}}, \mathbf{X})$, percentile intervals are calculated conditional on $\mathbf{X}$.

The 2-stage bootstrap procedure, which incorporates both stochastic variability and data uncertainty, makes use of the uncertainty distributions regarding dose D, the total number of exposed subjects N and the number of ill subjects Y as given in the WHO report of 2003 [10]. According to these uncertainty distributions, a new 'pseudo-original' sample $\{(\tilde{D}_j, S_j, \tilde{N}_j, \tilde{Y}_j, t_j, m_j)\}_{j=1}^{23}$ is generated given the original sample $\{(D_j, S_j, N_j, Y_j, t_j, m_j)\}_{j=1}^{23}$ where $\tilde{D}_j$, $\tilde{N}_j$ and $\tilde{Y}_j$ are the values randomly sampled from the corresponding uncertainty distributions. This is the first stage in the 2-stage bootstrap procedure reflecting data uncertainty. Then, given this 'pseudo-original' sample, a bootstrap sample $\{(\tilde{D}_j, S_j, \tilde{N}_j, \tilde{Y}_j^*, t_j, m_j)\}_{j=1}^{23}$ is generated by sampling binomial observations $\tilde{Y}_j^* \sim \text{Binomial}(\tilde{N}_j, \tilde{\pi}_j)$ with $\tilde{\pi}_j = \tilde{Y}_j/\tilde{N}_j$, for $j = 1, ..., 23$. This is the second stage in the 2-stage bootstrap procedure reflecting stochastic variability. Note that second stage in the 2-stage bootstrap procedure is similar to the 1-stage bootstrap procedure. On this bootstrap sample, DI-III with cloglog-link, $p_1 = -2.25$ and $p_1' = -0.25$ is refitted. Again, this process is repeated B times, yielding B different estimated dose-illness curves $f^*(\widehat{\boldsymbol{\beta}}, \mathbf{X})$ based on which the 90% pointwise confidence intervals are calculated as before.

Figure 5 shows (serovar × food matrix)-specific dose-illness curves together with 90% pointwise confidence intervals based on the 1-stage bootstrap procedure with $B = 500$. Similarly, Figure **??** shows the results of the 2-stage bootstrap procedure with $B = 500$. These figures contain only a selection of the 15 (serovar × food matrix)-specific estimated dose-illness curves, namely the curves corresponding to serovar types occuring only once and two selected curves corresponding to the more prevalent serovar types, e.g. *S.* Typhimurium and *S.* Enteritidis. The thick full lines (thick dashed lines) correspond to the estimated dose-response relationships based on the original sample for the normal population (susceptible population) and the thin lines to the corresponding 90% bootstrap pointwise confidence intervals. Comparing both figures, it is clear that the 2-stage bootstrap procedure generates higher variability in estimated dose-response curves compared to

the 1-stage bootstrap procedure. This is as expected because in the 2-stage bootstrap procedure stochastic variability and data uncertainty are taken into account whereas in the 1-stage bootstrap procedure only stochastic variability is accounted for. Since outbreak data are heavily subject to data uncertainty, the 2-stage bootstrap procedure, of which the results are shown in Figure **??**, is more appropriate. Inspection of Figure **??**, reveals large differences in (serovar $\times$ food matrix)-specific estimated curves as well as in width of the confidence intervals. For most combinations of (serovar $\times$ food matrix) and at most dose levels the difference between normal and susceptible population in estimated dose-response relationship is not significant. However, in general, the steep dose-response curves indicate a significant difference at low dose levels (e.g. Figure **??**c, e and g) with a higher probability of illness for the susceptible population. In addition, flat dose-response curves generally indicate a significant difference at high doses (e.g. Figure **??**b and h) with, again, a higher probability of illness for the susceptible population compared to the normal population. Finally, when comparing the two *S.* Typhimurium-specific dose-response curves (e.g. Figure **??**h and i), it is confirmed that the infectivity of a pathogen is higher when a fatty food matrix is involved [6]. The same conclusion can be drawn when comparing the two *S.* Enteritidis-specific curves (e.g. Figure **??**b and c). However, in this case, it is harder to a priori judge the fattiness of the food matrices involved.

Finally, the standard errors of the (serovar $\times$ food matrix)-specific $\widehat{\pi}$(ill|infection) are estimated as well using both bootstrap procedures. The results are summarized in Table **??**. This table contains only the estimates for the normal population. For the susceptible population, the bootstrap estimates $\widehat{\pi}$(ill|infection) are virtually always equal to < 1.000, yielding uninformative standard errors. As can be seen in the table, the standard errors obtained by 2-stage bootstrap procedure are larger compared to the ones obtained by the 1-stage bootstrap procedure, which is as expected and in line with previous findings.

[Figure 5 about here.]

[Figure 6 about here.]

[Table 7 about here.]

6 Discussion

Outbreak studies are considered to be the most valuable to model dose-response [10] and are an important way to validate risk assessments. However, outbreak studies are strongly subject to data uncertainty. In this study, data on *Salmonella* outbreak studies is used to assess the *Salmonella* dose-illness relation. Sofar, dose-illness models have received little attention. An important counterexample can be found in Teunis et al [1]. They introduce three different types of dose-illness models (depending on the underlying probability model of illness given infection), namely the monotonically increasing dose-illness model bounded between zero and one (DI-I), the monotonically unconstrained dose-illness model with the probability of illness being zero for zero dose and infinitely large dose levels (DI-II) and the monotonically increasing dose-illness model bounded between zero and $\pi(\text{ill}|\text{infection})$ being some constant $c < 1$ (DI-III). Obviously, other probability models of illness given infection than the ones introduced by Teunis et al [1] could be (equally well) considered. These would imply different properties of the dose-illness relation. However, a large set of competing probability models of illness given infection may give rise to the problem of identifiability (i.e. the same properties of dose-illness may result from different underlying probability models of illness given infection). Ideally, the dose-illness relation is modeled using both dose-infection and dose-illness data. Only then conclusive results regarding the underlying probability model of illness given infection can be obtained. However, we could not find any outbreak data containing information on infection as well as illness.

In this paper, dose-illness is modeled using GLMMs and fractional polynomials of dose. The fractional polynomial models are modified in order to satisfy the properties of three different types of dose-illness models that are proposed by Teunis et al [1]. Within these models, heterogeneity due to differences in host susceptibility are modeled using fixed effects whereas heterogeneity due to differences in serovar type and food matrix are modeled using random effects that are defined for unique combinations of serovar type $\times$ food matrix. As such, the analysis account for heterogeneity due to differences in serovar type and food matrix but the two sources of heterogeneity can not be separated. Note that this heterogeneity is modeled using random intercepts only. This implies that the different serovar type $\times$ food matrix-specific curves can not cross ([13], [14]). Additional flexibility in modeling heterogeneity can be achieved by including random slopes in the model. Unfortunately, this is not feasible for the current analyses due to data sparseness. Finally,

differences in, amongst others, serovar strains and micorflora of the food are potential sources of heterogeneity as well. However, the available data do not contain any information on these sources and as such, they can not be taken into account in the analyses.

Nonetheless, the current application illustrates that random effects models are appropriate statistical tools to account for different sources of heterogeneity simultaneously. The analyses indicate that the *Salmonella* outbreak data are best described by a monotonically increasing dose-illness relationship bounded between 0 and some constant $c < 1$, supporting the constant probability model of illness given infection [1]. Based on confidence intervals that incorporate both stochastic variability and data uncertainty, it is concluded that the susceptible population has a higher probability of illness at low dose levels when the combination pathogen-food matrix is extremely virulent and at high dose levels when the combination is less virulent. Furthermore, the analyses suggest that immunity exists in the normal population but not in the susceptible population.

In addition, we show how bootstrapping can be used to assess the effect of data uncertainty on the estimated dose-illness relation. The 1-stage bootstrap procedure merely accounts for the usual stochastic variability. The 2-stage bootstrap procedure accounts for data uncertainty as well. Evidently, the latter bootstrap procedure yields wider confidence intervals. However, another type of uncertainty, namely model uncertainty, is not addressed in this paper. Model uncertainty arises from the fact that several competing models might provide comparably good fits to the data and/or are equally biologically plausible. Model averaging is a way to account for model uncertainty and an example of the latter approach can be found in [24] and [25]. However, model averaging is not the focus of this manuscript. Furthermore, in case of epidemiological data compared to experimental data, data uncertainty seems much more prominent compared to model uncertainty.

Finally, the estimated dose-illness models can be used to validate quantitative risk assessments (QRAs). In case, interest is in one specific (serovar type $\times$ food matrix)-combination i of which the corresponding dose-illness model is estimated then the probability of infection given dose can be calculated using the model formulation in (7) with the estimates of $\boldsymbol{\beta}$ and b_i given in Table 3 and 4, respectively. Then, amongst others, the delta method can be used to calculate the corresponding standard errors. As such, confidence intervals can be easily constructed to account for stochastic variability. For instance, the probability of infection for a normal person due to consumption of beef contaminated with *S.* Enteritidis can be calculated using $g^{-1}(0.323 + 0.486 - 8.462\mathrm{d}^{-2.25})$

with d being the ingested dose on scale $\log_{10}(\text{CFU} + 1)$. In particular, for d = 2, the predicted probability of infection equals 0.107. Using the delta method, the corresponding standard error is calculated to be 0.016, yielding a 95% confidence interval of $[0.073; 0.140]$. These calculations are then to be repeated for each simulated value of d, yielding eventually a distribution for the predicted probability, one for the lower limit of the confidence interval and one for the upper limit. However, it often happens that interest is in a whole range of food-matrices or that it is impossible to know in advance the serovar-type(s) involved. Furthermore, the cluster-specific dose-illness are estimated for a limited number of serovar type $\times$ food matrix combinations only. In these cases, one needs to randomly sample from the distribution of serovar type $\times$ food matrix-specific intercepts $b_i \sim N(0, 0.780)$ first (see Table 3).

Acknowledgments

This study has been carried out with the financial support of the Belgian Federal Public Service of Health, Food Chain Safety, and Environment research programme (R-04/003-Metzoon) 'Development of a methodology for quantitative assessment of zoonotic risks in Belgium applied to the 'Salmonella in pork' model'. The authors gratefully acknowledge the financial support from the IAP research Network P6/03 of the Belgian Government (Belgian Science Policy). Marc Aerts acknowledges support from FWO-Vlaanderen Research Project G.0151.05.

References

[1] P. F. M. Teunis, N. J. D. Nagelkerke, and C. N. Havelaar. Dose response models for infectious gastroenteritis. *Risk Analysis*, 19:1251–1260, 1999.

[2] P. S. Mead, L. Slutsker, V. Dietz, L. F. McCaig, J. S. Bresee, C. Shapiro, P. M. Griffin, and R. V. Tauxe. Food-related illness and death in the United States. emerging infectious diseases. *Emerging Infectious Diseases*, 5 (5):607–625, 1999.

[3] C. J. Thorns. Bacterial food-borne zoonoses. *Review scientific et technique (International Office of Epizootics)*, 19 (5):226–239, 2000.

[4] Nationaal Referentiecentrum voor *Salmonella* en *Shigella*. *Salmonella* en *Shigella* stammen afgezonderd in belgi in 2005. Wetenschappelijk Instituut voor de Volksgezondheid, Afdeling bacteriologie, Departement Microbiologie Report. Pages 50. 2005.

[5] M. J. Blaser and L. S. Newman. A review of human salmonellosis: I. infective dose. *Review of Infectious Disease*, 4:1096–1106, 1982.

[6] J. Y. D'Aoust. Pathogenicity of foodborne *Salmonella*. *International Journal of Food Microbiology*, 12:17–40, 1991.

[7] N. B. McCullough and C. W. Eisele. Experimental human salmonellosis: I. Pathogenicity of strains of *Salmonella meleagridis* and *Salmonella anatum* obtained from spray-dried whole egg. *Journal of Infectious Disease*, 88:278–279, 1951.

[8] N. B. McCullough and C. W. Eisele. Experimental human salmonellosis: III. Pathogenicity of strains of *Salmonella newport*, *Salmonella derby* and *Salmonella Bareilly* obtained from spray-dried whole egg. *Journal of Infectious Disease*, 89:209–213, 1951.

[9] N. B. McCullough and C. W. Eisele. Experimental human salmonellosis: IV. Pathogenicity of strains of *Salmonella pullorum* obtained from spray-dried whole egg. *Journal of Infectious Disease*, 89:259–265, 1951.

[10] World Health Organization. Risk assessments of *Salmonella* in eggs and broiler chickens. Microbial risk assessment series, nr. 2. World Health Organization (WHO), Geneva, Switzerland. 2003.

[11] H. K. Latimer, L. A. Jaykus, R. A. Morales, P. Cowen, and D. Crawford-Brown. A weighted composite dose-response model for human salmonellosis. *Risk Analysis*, 21:295–305, 2001.

[12] T. Oscar. Dose-response model for 13 strains of *Salmonella*. *Risk Analysis*, 24:41–49, 2004.

[13] G. Verbeke and G. Molenberghs. *Linear Mixed Models for Longitudinal Data*. Springer, New York, 2000.

[14] G. Molenberghs and G. Verbeke. *Models for Discrete Longitudinal Data*. Springer, New York, 2005.

[15] P. Royston and D. Altman. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion). *Applied Statistics*, 43:429–467, 1994.

[16] W. A. Furumoto and R. Mickey. A mathematical model for the infectivity-dilution curve of tobacco mosaic virus: Theoretical considerations. *Virology*, 32:216–223, 1967.

[17] R. E. Black, M. L. Levine, T. P. Clements, T. P. Hughes, and M. J. Blaser. Experimental *Campylobacter jejuni* infection in humans. *Journal of Infectious Diseases*, 157 (3):472–479, 1988.

[18] H. L. Dupont, C. L. Chappell, C. R. Sterling, J. B. Okhuysen, and W. Jakubowski. The infectivity of *Cryptosporidium parvum* in healthy volunteers. *The New England journal of medicine*, 332 (13):855–859, 1995.

[19] P. McCullagh and J. A. Nelder. *Generalized Linear Models, 2d edition*. Chapman and Hill, London, 1989.

[20] C. Faes, H. Geys, M. Aerts, and G. Molenberghs. On the use of fractional polynomial predictors for quantitative risk assessment in developmental toxicity studies. *Statistical Modelling*, 3:109–126, 2003.

[21] C. Faes, N. Hens, M. Aerts, Z. Shkedy, H. Geys, K. Mintiens, H. Laevens, and F. Boelaert. Estimating herd-specific force of infection by using random-effects model for clustered binary data using monotone fractional polynomials. *Journal of the Royal Statistical Society, Series C*, 55:595–613, 2006.

[22] N. Hens, C. Faes, M. Aerts, Z. Shkedy, K. Mintiens, H. Laevens, and F. Boelaert. Handling missingness when modelling the force of infection from clustered zeroprevalence data. *Journal of Agricultural, Biological and Environmental Statistics*, 00:000–00, 2007.

[23] C. M. Hurvich, J. S. Simonoff, and C.-L. Tsai. Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *Journal of Royal Statistical Society, Series B*, 60:271–293, 1998.

[24] C. Faes, M. Aerts, H. Geys, and G. Molenberghs. Model averaging using fractional polynomials to estimate a safe level of exposure. *Risk Analysis*, 27:00–00, 2007.

[25] H. Namata, M. Aerts, C. Faes, and P. F. M. Teunis. Model averaging in microbial risk assessment using fractional polynomials and generalized linear mixed models. *Submitted*, 00:00–00, 2007.

List of Figures

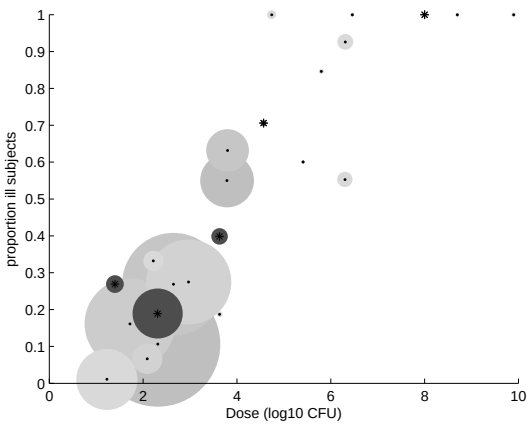
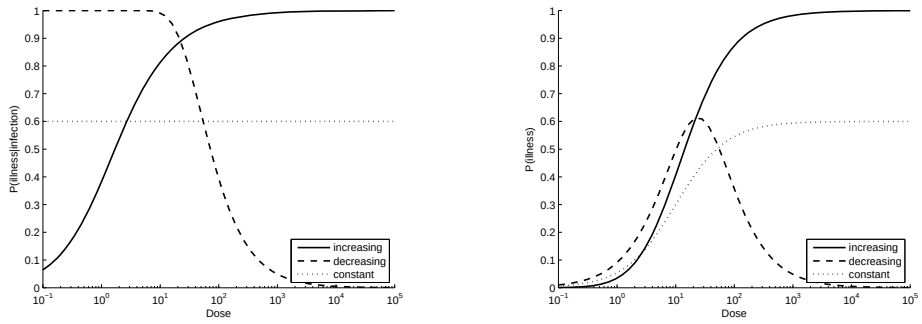


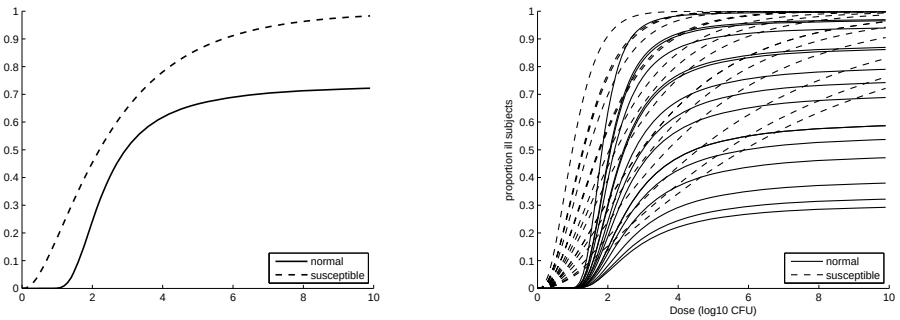
Figure 1: Bubble plot of proportion of ill subjects as function of dose, for 20 outbreak studies as reported in WHO, 2003. The area of the bubbles is proportional to the number of exposed subjects. Observations on normal subjects are indicated with a dot and light gray colored bubbles and observations on susceptible subjects are indicated with a star and dark gray colored bubbles.

Figure 2: Multiple-stage model for illness as proposed by Teunis et al [1].



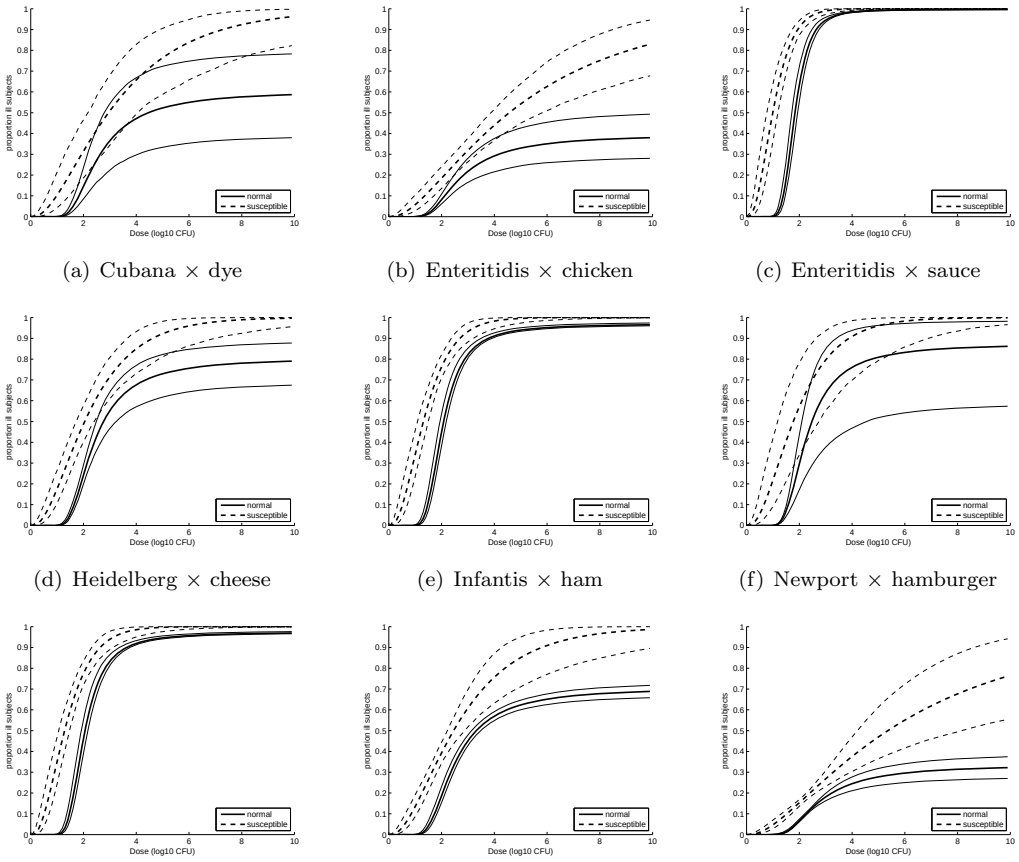
(a) π(ill|infection,dose) (b) π(ill|dose)

Figure 3: (a) Three different probability models of illness given infection (increasing, decreasing and constant probability) and (b) the corresponding dose-illness models.



(a) marginal (b) cluster-specific

Figure 4: Simulation based marginal dose-illness relationship based on GLMM with fractional polynomial of dose and corresponding (serovar × food matrix)-specific dose-illness relationships.



(a) Cubana × dye (b) Enteritidis × chicken (c) Enteritidis × sauce

(d) Heidelberg × cheese (e) Infantis × ham (f) Newport × hamburger

(g) Oranienburg × soup (h) Typhimurium × ice cream (i) Typhimurium × water

Figure 5: (Serovar × food matrix)-specific dose-illness curves based on GLMM with fractional polynomial of dose (thick lines) + 90% confidence intervals based on the 1-stage bootstrap procedure incorporating stochastic variability only (thin lines).