Estimating the seroprevalence of *Toxoplasma gondii* of the human population in the Netherlands

A Bayesian Approach

Author
Josanne Verheule

Supervised by
Dr. I. Kryven & Dr. A. Swart

Rijksinstituut voor Volksgezondheid
en Milieu
Ministerie van Volksgezondheid,
Welzijn en Sport

A Thesis submitted in fulfillment of requirements for the degree of
Master of Science in Mathematical Sciences

Department of Mathematics
Utrecht University
2024

Acknowledgement

I want to thank my supervisors, Dr. Ivan Kryven from Utrecht University and Dr. Arno Swart from the RIVM, for their endless help and guidance. I am also thankful to Marieke Opsteegh for her assistance and insights. Thanks to Titia for showing us around the lab. Special thanks to Oda for helping me navigate the data and find my way at the RIVM.

I am particularly grateful to Michiel for his endless patience and support during this process. Additionally, my appreciation goes out to Franc for taking the time to read through my work, even without any background in the subject. Your feedback was immensely helpful. Finally, the support of Margo was invaluable, I could not have done this without her.

Abstract

The dynamics of infections and their prevalence in populations are key elements for effective public health policies. In this study we investigated the change in prevalence and the force of infection over time within the Dutch population for *Toxoplasma gondii*, a widespread zoonotic foodborne parasite. We offer an innovative approach to modeling infection dynamics in disease surveillance, by applying Bayesian statistics and compartmental disease modeling. Using serological data from three independent studies conducted at 10-year intervals, a binary mixture model was implemented, characterizing the distribution of the measurements to estimate the prevalence without reliance on predetermined cut-off values for the classification of infected and non-infected subpopulations. Potential external covariates such as education level, pet ownership (specifically cats), and the consumption of food associated with higher risk of infection were explored to ascertain any changes in risk factors over time. A discrete-time and age-dependent SI(S) compartmental disease model was implemented to estimate the force of infection across years and ages.

This Master thesis project is conducted as part of an internship at the National Institute for Public Health and Environment (RIVM).

Keywords: ELISA; serology; epidemiology; mixture models; Bayesian statistics; HDI; ROPE; compartmental model; SIS.

Contents

1 Introduction

Toxoplasma gondii is a single-celled zoonotic foodborne parasite that can infect a wide variety of warm-blooded animals as intermediate hosts, including humans. When infecting a host, the parasite is believed to persist in the body indefinitely as latent tissue cysts [1–3], although the truth of this assumption has recently been questioned [4–6]. These cysts are frequently found in the brain, heart, and skeletal muscle. Infection is often asymptomatic, particularly in individuals with a healthy immune system. If symptoms do emerge, they are usually non-specific and may present themselves as flu-like, manifesting as swollen lymph nodes, muscle pain, and fatigue [7]. However, for certain groups infection poses a greater risk. If primary infection occurs during pregnancy, it may result in miscarriage or stillbirth or lead to congenital toxoplasmosis, which can cause major birth defects in the infant, such as ocular disease or neurological disorders [8–10]. For individuals with compromised immunity, such as those with HIV/AIDS or organ transplant recipients, a reawakened dormant infection can lead to a severe condition called toxoplasmic encephalitis, which can be fatal [8, 11].

Primary sources of infection include the consumption of tissue cysts in undercooked or raw meat from infected animals and the ingestion of oocysts through contaminated soil, food or water [12]. Vertical transmission, where the infection is passed from an infected mother to her fetus, is another significant source of infection, which was estimated at 1.3/1000 live born children in the Netherlands in 2017 [13]. Felids are the only known definite host, shedding oocysts in the environment that can survive in the environment for more than one year [14]. Some studies have found that cat ownership slightly increases the risk of infection [15, 16], while others have shown that direct contact with domestic cats is not likely to be a risk factor for infection, but rather contact with the environment contaminated by their faeces [17, 18]. It appears that the consumption of meat of infected animals is one of the most important sources of infection in Europe compared with other possible transmission routes. [16, 19]

The prevalence of *T. gondii* infection varies significantly by region, and is considered the second most important foodborne parasite for Europe and its regions, and ranked highest in Western Europe [20]. Studying this pathogen is crucial because of its potential to cause severe disease and has high presence in the environment.

The detection of infection relies primarily on serological assays [21]; upon infection with a pathogen, elevated antibodies against the parasite are detected in the blood. The level of this response varies among individuals. In 1995/1996, 2006/2007, and 2016/2017, the serological response for *T. gondii* was measured with specific enzyme-linked immunosorbent assays (ELISAs) in a random sample of the Dutch population in a nationwide survey conducted by the National Institute for Public Health and Environment. This data was analyzed in three earlier studies [13, 15, 22]. To estimate the prevalence of the entire population, individuals were classified as either infected or non-infected, based on a cut-off value derived from control measurements. Age-specific prevalence was estimated using logistic regression. With this approach, it was found that seroprevalence decreased between 1995/1996 from 40.5% to 26% in 2006/2007, and increased to 29.9% in 2016/2017.

In this thesis, the focus lies on understanding the change of the seroprevalence of *T. gondii* through 1995-2017. A Bayesian statistical framework is used to derive a distribution on the immune response measurements from the surveys among the Dutch population, so that the prevalence can be estimated without the use of a cut-off value. This improves the accuracy of the estimation, since serological response values of the sub-populations may overlap due to natural variation between individuals. We follow a similar approach used by Opsteegh et al. (2010) in [23] and Opsteegh et al. (2011) in [6], where the prevalence of *T. gondii* was estimated through a binary mixture of normal distributions on the serological response data

produced by ELISA for sheep and wild boar respectively. Research suggests that a mixture model outperforms an approach where a cut-off is used [5, 24, 25].

Differences in the contribution of risk factors between the three studies used in the analysis might explain differences in prevalence, by indicating behavior- or environmental changes through the years. These covariates are examined by supplying characteristics of the population as covariates in the Bayesian model. The Highest Density Interval (HDI) combined with the Region Of Practical Equivalence (ROPE) is used to provide insights into the relevance of these factors to the probability of infection. This differs from earlier, frequentist analyses [13, 15, 22] on the data, where their importance was determined using odds ratios.

The age- and time dependent dynamics of the infections was defined through a compartmental epidemiological model, characterizing the dynamics of the infections through a system of differential equations, similar to Opsteegh et al. (2011) in [6] and Dámek et al. (2023) in [5]. Our approach differs in the assumption that the population is in a steady state; in the context of this thesis, this assumption is not applicable since the age specific prevalence changed through the years. The age-dependent force of infection was studied only for the third survey by Van den Berg et. al (2023) in [13]. We fitted the age- and time dependent force of infection defined by the compartmental model directly through the estimated age-specific prevalence distribution from the binary mixture model, using the theory of age-structured population dynamics, such as constructed by Ianelli (1995) in [26] and Hetcote (2000) in [27]. This differs from the approach used in [6], where the mixture model was used to calculate the optimal cut-off value from the distributions which was then used to score each individual animal as positive or negative, and where transmission rates were estimated using least-squares fit. Different than in [5], where only constant force of infection was considered, a piecewise constant force of infection was assumed, which gives more flexibility and insight in the risk over time and over age.

The compartmental model also gives insight in the likeliness of the widely accepted assumption of lifelong persistence of the infection. This is done by comparing the model where only transitions from susceptible to infected are considered, to a model where infected individuals can become seronegative again.

The main objective of this thesis is to both describe and shed light on the infection spread patterns and risks associated with infection, as well as to enhance the accuracy and reliability of the prevalence estimates, thus contribute valuable insights to the field of mathematical epidemiology.

2 Material/Methodology

2.1 Study Population

In the Netherlands, the Peiling Immunisatie Effect Nederland Ter Evaluatie van het Rijksvaccinatieprogramma (PIENTER) project is established to assess the immunity of the population to infectious diseases. Through this project, the National Institute for Public Health and Environment (RIVM) conducts nationwide surveys under the direction of the Ministry of Health, Welfare and Sport every 10 years to estimate immunity levels and the spread of diseases such as measles, rubella, mumps, and polio. The focus is on monitoring the level of protection of the population against vaccine-preventable diseases and evaluating the effectiveness of vaccination programs. This is done with serological testing, identifying and measuring the levels of specific antibodies in blood samples collected from the participants in a cross-sectional study. In addition, a detailed questionnaire is administered on personal and demographic characteristics, history of illness, education, and activities that may increase the risk of exposure to infectious diseases, such as diet, profession and contact with animals. These studies play an important role in monitoring vaccination programs, preventing outbreaks and controlling disease spread in the Netherlands. At this time, three independent surveys have been completed within the project; the first study, PIENTER 1, was conducted in 1995 and 1996, PIENTER 2 in 2006 and 2007, and PIENTER 3 in 2016 and 2017. The participants who responded were between 0 and 79 years old in the first and second study, and between 0 and 89 years old for the third study, and were distributed across the Netherlands.

For the first study, PIENTER 1, a total of 7521 sera were tested for IgG antibodies against *Toxoplasma gondii*. The national seroprevalence was previously estimated by Kortbeek et al. [15] as 40.5% (95 % CI 37.5 - 43.4). The PIENTER 2 study included 5541 samples and a decrease was found of the overall seroprevalence to 26.0% (95% CI 24.0 - 28.0) by Hofhuis et al. [22]. In both the first and the second study, the maximum age was limited to 79. The third study contained 6414 sera, and analysis by Van den Berg et. al [13] estimated the seroprevalence at 30.9% (95% CI 29.4 - 32.4).

2.2 Serological Measurements

Enzyme-linked immunosorbent assay (ELISA) is a technique used to detect and quantify specific proteins or antibodies in a sample. It involves the use of an enzyme-linked antibody or antigen and a substrate that produces a measurable signal in response to the catalytic activity of the enzyme. The optical density (OD) of the reaction product is measured using a spectrophotometer. This is measured at a specific wavelength that corresponds to the absorption maximum of the reaction product. The intensity of the color produced is proportional to the amount of enzyme bound to the target protein or antibody, which in turn is proportional to the concentration in the sample.

Each plate contains multiple control serums that are kept identical across plates; negative and positive controls (NC and PC respectively), and a sample that determines the cut-off (QC) value for that plate. These controls are used to interpret the results and correct for background noise. The blank wells contain a substrate with no target protein/antibody present and are used to measure the background signal produced by non-specific binding and correct for other sources of noise. The positive controls are wells containing a serum that contains a known amount of the target protein/antibody and are used to verify that the assay is working correctly. For negative controls, three different serums that are known to contain no antibodies, were used. The configuration of the plates used in this study is shown in Figure 1. Each plate configuration was tested in duplicate.

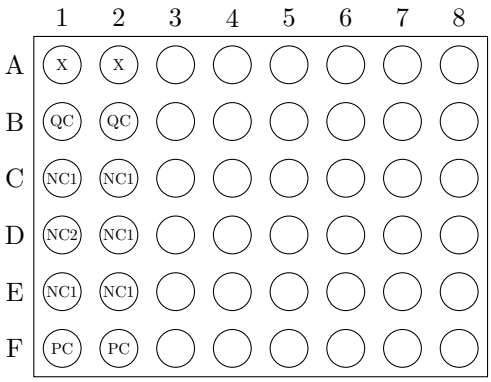
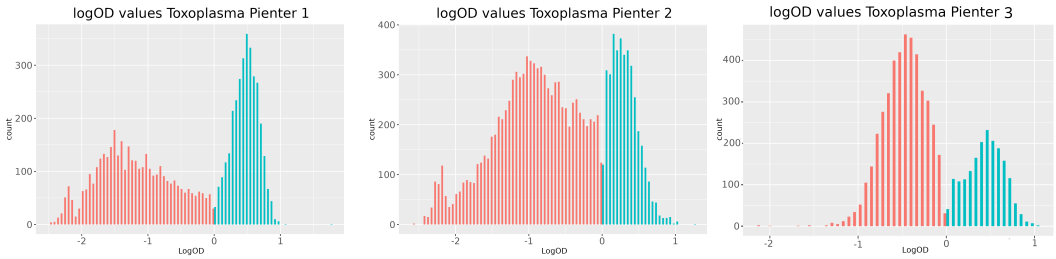


Figure 1: ELIZA plate configuration, where X denotes blanks, QC denotes the cut-off serum, the NC's denote the negative cut-off serums and PC denotes the positive control serum; each are performed in duplo.



(a) PIENTER 1 (1995/1996) **(b)** PIENTER 2 (2006/2007) **(c)** PIENTER 3 (2016/2017)

Figure 2: Histograms of log-transformed optical density (OD) values for each of the PIENTER studies, corrected with the blank measurements and weighed by cut-off measurements. The negative and positive classifications are indicated by red and blue respectively.

In earlier analyses of the data [13, 15, 22] For each sample, the average value of the blank measurements on the corresponding plate was subtracted to correct for background noise. The seroprevalence was then estimated by categorizing individuals on the basis of the mean of the cut-off (QC) measurements of the corresponding plates; if a ratio of ≥ 1 was observed, the sample was considered positive. The adjusted measurements and their characterization are shown in Figure 2.

2.3 Bayesian Inference

Bayesian inference systematically reallocates credibility across potential outcomes, based on the principles of conditional probability [28]. The model is specified by defining the joined distribution $P(Y, \theta)$ over observed variables Y and unobserved variables θ. Initial knowledge, supplied as prior distributions $P(\theta)$, of population parameters are combined with observed data Y to produce informed predictions about the true value of the parameters in question. These parameters are fixed unknown features of the model, where the uncertainty about the true value is embodied by the variation of the prior distributions. All unknown parameters can incorporate uncertainty, defined through a probability distribution. Priors should be established independently, before observing the data, based on systematic reviews, meta-analysis and previous studies on similar data. Updated estimates, represented by the posterior distribution, are then produced by weighing the priors with the likelihood function

of the observed data, through Bayes' rule

$$\overbrace{P(\theta|Y)}^{\text{Posterior}} = \frac{\overbrace{P(Y|\theta)}^{\text{Likelihood}}\overbrace{P(\theta)}^{\text{Prior}}}{P(Y)}, \tag{1}$$

where Y represents the data and θ the unknown model parameters.

Bayesian "significance" testing indices may be grouped into three main types: Bayes factors, posterior indices, and Region of Practical Equivalence (ROPE)-based indices [29]. Bayes factors compare the relative evidence of one model over another ("given the observed data, is the null hypothesis of an absence of an effect more, or less likely?"). Posterior indices analyze the posterior distribution's objective characteristics, such as the proportion of strictly positive values. ROPE-based indices redefine the null hypothesis to include a range of values that are considered too small to be of any practical relevance.

2.3.1 HDI + ROPE Decision rule

The Region of Practical Equivalence (ROPE) and Highest Density Interval (HDI) are techniques used in the context of hierarchical Bayesian models to assess the impact of hyperparameters on the model's predictions. The ROPE defines a range of values for a hyperparameter that are practically equivalent in their effect on the model. This is done by examining how changes in the hyperparameter influence the posterior distribution of the model parameters; if this distribution mostly falls within the ROPE, this implies the hyperparameter has minimal impact on the model's predictions. The HDI represents a range containing a specified proportion of the posterior probability density such that all points within the interval have a higher probability density than points outside the interval. It summarizes the range of most credible values of the measurements [30], which can be used to estimate the uncertainty of model parameter estimates, indicated by the width of the parameter distributions under different hyperparameter values.

The decision rule involving the HDI and the ROPE is straightforward. When assessing a parameter's null value, if the entire HDI lies within the ROPE, the null value is accepted as it indicates that the most credible values of the parameter are practically equivalent to the null. Conversely, if the entire HDI lies outside the ROPE, the null value is rejected, implying that none of the credible values are practically equivalent to the null. In cases where the HDI partially overlaps with the ROPE, a definitive decision cannot be made, as some credible values are equivalent to the null while others are not.

The choice of the range of the HDI as the Credible Interval and the limits for the ROPE highly influences the decision. The general choice for confidence intervals in the frequentist context of significance testing is the 95% interval. However, the percentage of 89% for the computation of the HDI's as suggested by McElrath (2016) [31] ("Because it is a prime") is more stable and is considered to be a better choice [32]. A range for the ROPE set at -0.1 to 0.1 is used by Krusche (2018) [30], which is, according to Cohen (1988) [33], a negligible effect size. For correlations, we follow conventions by using a value of 0.05, half the value of a negligible effect size [32].

2.3.2 Probability of direction

The probability of direction (*pd*) is an index of effect existence with a range between 50-100%, representing the probability that an effect is strictly positive or negative. It is defined as the proportion of the posterior distribution that is of the same sign as the median, given the data and the model [34]. High *pd*-values suggest the existence of an effect, but low values

do not give information on the certainty that no effect is present. The value does not require any additional information from the data or the model, so that it is an objective property of the posterior distribution. It is strongly related to the frequentist p-value, and is likewise not able to quantify evidence in favor of the null hypothesis [29].

2.4 Mixture Model

The population can be characterized by a mixture of two distinct components; positive and negative, also called susceptible and infected. By describing the data with a binary mixture of distributions, the prevalence of *Toxoplasma gondii* in the population can be directly estimated without a sharp cut-off on the values of the serological measurements, allowing for more flexibility in defining the components. This estimation takes the uncertainty associated with the data into account, providing more reliable results compared to a fixed cut-off value, by using the natural variation between individuals and noise introduced by the measuring technique. Moreover, the measurements of these distinct sub-populations may not be clearly divided but instead have overlapping distributions. Furthermore, binary mixture models enable the identification of covariates that might influence the presence of antibodies. By including additional variables such as age or gender in the model, we can explore how these factors affect the likelihood of belonging to the positive or negative group.

Denoting the distributions for the negative and the positive sub-populations by $f_-(x;\mu_-,\sigma_-)$ and $f_+(x;\mu_+,\sigma_+)$ respectively, the distribution describing the data can be obtained by conditioning on the possible outcomes. The expression for the binary mixture model then becomes the weighted sum of the two sub-populations, as

$$f(x) = f_-(x;\mu_-,\sigma_-)p(\text{neg}) + f_+(x;\mu_+,\sigma_+)p(\text{pos}).$$

The coefficient of the positive component, $p(\text{pos})$, is exactly the prevalence. This probability density function is fitted to the data by estimating $\mu_\pm$, $\sigma_\pm$ and $p(\text{pos})$.

The frequency distribution of the titres is usually log-normally distributed [35, 36]. From Figure 2, it can be clearly seen that for each of the studies, two clear components are present, each with a different mean and standard deviation, where the higher mean is associated with the seropositive population. Therefore, a binormal distribution is suitable for describing the log-normally transformed titre values.

The mixture model is not designed for the classification of individual samples. However, the probability of being seropositive can be calculated through the estimates, resulting in a more accurate classification. The probability that an individual with logOD value x is infected is given by [25]

$$\mathbb{P}(\text{pos}|x) = \frac{\mathbb{P}(x|\text{pos})\mathbb{P}(\text{pos})}{\mathbb{P}(x)} = \frac{f(x;\mu^+,\sigma^+)p(\text{pos})}{f(x)}$$

We can now, for example, define individual measurements x with $\mathbb{P}(\text{pos}|x) > 0.5$ as positive, from which we can derive a cut-off value c on the values from the data.

2.4.1 Addition of Covariates

Through the questionnaire, insight can be gained in the factors that may influence the risk of infection. These factors may include demographic classifications such as religion or education level, or behavioral and life-style characteristics such as pet ownership, time spent gardening or the consumption of raw meat. The influence of factor λ on the infection risk is examined through the conditional probability of the seroprevalence $p(\text{pos}|\lambda)$. The distributions of the negative and positive populations are independent of λ, since these characteristics do not

Table 1: Covariates

Name	Question
education	What is the highest level of education or training that you have completed?[b]
gardening	Have you worked with your bare hands in the soil in the garden/on land in the past 12 months?
agriculture	Have you kept farm animals in the past 5 years?
petcat	Have you kept a cat as a pet for the past 5 years?
contcat	Did you have contact with cats in the past 12 months?
rawmeat[a]	Have you consumed raw meat products in the past 12 months?
rawbeef[a]	Have you consumed raw beef in the past 12 months?
rawporkmeat[a]	Have you consumed raw pork meat in the past 12 months?
unwashedveget[a]	Do you consume unwashed raw vegetables?

[a] Unavailable for PIENTER 1;
[b] The educational levels are defined as follows: 'low' (primary school, lower vocational, or lower general secondary education), 'medium' (intermediate vocational, intermediate general secondary, and higher general secondary education), and 'high' (higher vocational secondary education or university-level education). For children under the age of 17, the level of their mother is taken.

typically influence the antibody response of individuals. Furthermore, age does not have a notable influence on the antibody response, and thus on the values of μ and σ. Therefore λ may also represent age (or age groups) to obtain the age-specific seroprevalence of the population. The conditional probability is estimated through the distribution of the data through

$$f(x|\lambda) = f_-(x;\mu_-,\sigma_-)p(\text{neg}|\lambda) + f_+(x;\mu_+,\sigma_+)p(\text{pos}|\lambda). \tag{2}$$

To identify key factors associated with increased risk of infection, we made a selection of variables that are often assumed of influence in the risk of infection from the questionnaires. The questionnaires used across the different PIENTER studies were not uniform. The selected variables are explained in Table 1.

In this thesis, to analyse the effect of the covariates on the prevalence, we used a three-step process. First, statistical significance is assessed through the Highest Density Interval (HDI), determining if an effect is present. Then practical significance is verified through the Region of Practical Equivalence (ROPE), assessing the relevance of these effects in a practical context, as explained in Section 2.3.1. Lastly, the probability of direction (*pd*) quantifies the certainty that the effect is in the observed direction as explained in Section 2.3.2.

2.5 Compartmental Epidemiological Model

Compartmental epidemiological models are used to understand the spread and dynamics of infectious diseases within populations, allowing us to track the progression of the disease and assess its impact on various subgroups. The concept of compartmental models can be traced back to the early 20th century when researchers like Kermack and McKendrick [37] developed the first mathematical models to describe the spread of diseases such as measles and smallpox. The fundamental idea behind compartmental models is to divide the population into distinct groups or compartments based on their disease status. The most commonly used compartments are susceptible (S), infected (I), and recovered (R). These compartments are connected by a set of equations that describe the flow of individuals between them. The parameters characterize the rates of at which individuals move between

the compartments, such as the transmission rate, or force of infection, which governs the rate at which susceptible individuals become infected, and the recovery rate, which determines how quickly infected individuals recover. Other factors, such as birth and death rates, can also be included in the model to account for population changes. The force of infection could depend only on age, on time, or on both age and time. In the first case, where the effect will be constant through time, a difference in prevalence between the different studies will present as a shift on the age-axis. A time-dependence will cause a shift in the y-axis (the prevalence axis).

In the case of *Toxoplasma gondii*, it is assumed that the parasite establishes a lifelong presence within its hosts after infection [2, 3, 5]. Individuals cannot recover in the conventional sense, and therefore the compartment consisting of the recovered individuals is not included in the models. The S compartment in this case indicates the compartment of seronegative individuals, and the I compartment encompasses the seropositive individuals. However, recent research suggests that the assumption that infection is lifelong might not be entirely correct [4, 6], suggesting that a flow of individuals from the infected to the susceptible compartment could be nonzero. To investigate this assumption, both the SI and the SIS model will be considered. To simplify the analysis, we disregard several factors: the impact of the disease on mortality, the influence of infection on fertility and we assume that newborns are not born infected. Furthermore, we assume that the rate at which individuals recover and become susceptible again, denoted by β, is independent of age and time.

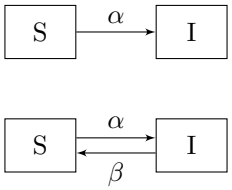


Figure 3: Infection models: SI and SIS.

2.5.1 SIS with age dependence

The age-structured SIS model for a closed population (there is no migration or movement of individuals out of the population), where individuals are assumed to become seronegative again, is described by the system of equations

$$\begin{aligned} \frac{\mathrm{d}S(a)}{\mathrm{d}a} &= -(\mu(a)+\alpha(a))S(a)+\beta I(a), \quad S(0)=N(0), \\ \frac{\mathrm{d}I(t)}{\mathrm{d}a} &= \alpha(a)S(a)-(\mu(a)+\beta)I(a), \qquad I(0)=0, \end{aligned} \tag{3}$$

where α is the force of infection, β is the recovery rate, μ the age dependent death rate and the total population at age a is $N(a) = S(a) + I(a)$. By adding the differential equations we find that

$$\frac{\mathrm{d}N(a)}{\mathrm{d}a} = -\mu(a)N(a).$$

Therefore the solution to the system (3) is found by solving

$$\frac{\mathrm{d}I(a)}{\mathrm{d}a} = \alpha(a)N(a)-(\mu(a)+\alpha(a)+\beta)I(a), \qquad I(0)=0.$$

The infected proportion of the population $i(a)$ is given by $I(a) = N(a)i(a)$, so that by the basic derivative rules,

$$\frac{\mathrm{d}i(a)}{\mathrm{d}a} = \frac{I'(a)N(a) - \mu(a)N(a)I(a)}{N^2(a)} = \frac{I'(a) - \mu(a)I(a)}{N(a)},$$

and the death rate drops out. This does not come as a surprise, since there are no interactions between the individuals and therefore the death rate does not influence the rate of change within the infected population. As a result, the system is reduced to the single equation

$$\frac{\mathrm{d}i(a)}{\mathrm{d}a} = \alpha(a) - (\alpha(a) + \beta)i(a), \qquad i(0) = 0. \tag{4}$$

We will denote $\phi(a) = \alpha(a) + \beta$, and $\Gamma(a) = \int_0^a \phi(s)\ \mathrm{d}s$. Note that

$$\frac{\mathrm{d}}{\mathrm{d}a}\left[e^{\Gamma(a)}\right] = \phi(a)e^{\Gamma(a)}. \tag{5}$$

By multiplying both sides of the equation (4) by $e^{\Gamma(a)}$, and then using the chain rule and (5), it is obtained that

$$\begin{aligned} \alpha(a)e^{\Gamma(a)} &= \frac{\mathrm{d}i(a)}{\mathrm{d}a}e^{\Gamma(a)} + i(a)\left[\phi(a)e^{\Gamma(a)}\right] \\ &= \frac{\mathrm{d}}{\mathrm{d}a}\left[i(a)e^{\Gamma(a)}\right]. \end{aligned} \tag{6}$$

Since $\alpha(a) = \phi(a) - \beta$, the left hand side of (6) can be rewritten with (5) to obtain

$$\frac{\mathrm{d}}{\mathrm{d}a}\left[i(a)e^{\Gamma(a)}\right] = \frac{d}{da}\left[e^{\Gamma(a)}\right] - \beta e^{\Gamma(a)}.$$

Integrating both sides with respect to a gives

$$i(a)e^{\Gamma(a)} = e^{\Gamma(a)} - \beta\int_0^a e^{\Gamma(s)}\ \mathrm{d}s + C.$$

From the initial condition $i(0) = 0$ follows that $C = -1$, so that the solution is given by

$$i(a) = 1 - e^{-\Gamma(a)}\left(1 + \beta\int_0^a e^{\Gamma(s)}\ \mathrm{d}s\right). \tag{7}$$

2.5.2 Piecewise constant infection rate dependent on age

We partition the age range into intervals $A_i = [a_{i-1}, a_i),\ i \in \{1, \ldots, m\}$, where $a_0 = 0$ and $a_m = a_{\max}$. Assuming that $\alpha(a)$ is constant on these intervals, $\alpha(a) = \alpha_k$ for $a \in A_k$, then $\alpha(a)$ is of the form

$$\alpha(a) = \sum_i \alpha_i \mathbb{1}\{a \in A_i\},$$

where $\mathbb{1}\{F\}$ denotes the indicator function of the event F. By integration, we find

$$\Gamma(a) = \beta a + \int_0^a \sum_i \alpha_i \mathbb{1}\{s \in A_i\}\ \mathrm{d}s = \beta a + \sum_i \alpha_i |A_i \cap [0, a]|$$

The solution to (7) is then given by

$$\begin{aligned} i(a) &= 1 - e^{-(\beta a + \sum_i \alpha_i |A_i \cap [0,a]|)} \left(1 + \beta \int_0^a e^{\beta s + \sum_i \alpha_i |A_i \cap [0,s]|} \, \mathrm{d}s\right) \\ &= 1 - e^{-(\beta a + \sum_i \alpha_i |A_i \cap [0,a]|)} \left(1 + \beta \sum_j |A_j \cap [0,a]| e^{\sum_{i \leq j} (\alpha_i + \beta) |A_i \cap [0,a]|}\right). \end{aligned}$$

If we take age intervals of unit length, then then the proportion of the population that is infected is given by

$$i(a) = 1 - \exp\left\{-\beta a - \sum_{k=1}^{a} \alpha_k\right\} \left(1 + \beta \sum_{k=1}^{a} \exp\left\{\beta k + \sum_{i=1}^{k} \alpha_j\right\}\right). \tag{8}$$

2.5.3 SIS with time- and age dependence

We examine again the closed population from Section 2.5.1, and expand the model by adding time dependence to the dynamics.

The age-structured, time dependent SIS model is then formulated by the following system of equations:

$$\begin{aligned} \left(\frac{\partial}{\partial a} + \frac{\partial}{\partial t}\right) S(t,a) &= -(\mu(a) + \alpha(t,a)) S(t,a) + \beta I(t,a), \quad S(t,0) = B(t), \quad S(0,a) = S_0(a), \\ \left(\frac{\partial}{\partial a} + \frac{\partial}{\partial t}\right) I(t,a) &= \alpha(t,a) S(t,a) - (\mu(a) + \beta) I(t,a), \qquad I(t,0) = 0, \qquad I(0,a) = I_0(a), \end{aligned} \tag{9}$$

where α is again the force of infection, β denotes the recovery rate which is independent of age and time, $\mu(a)$ denotes the age dependent death rate and the entire population at time t and age a is $N(t,a) = S(t,a) + I(t,a)$ where $B(t) = N(t,0)$ denotes the birth rate at time t.

By adding the PDE's, we know that the population $N(t,a)$ satisfies the conservation law for populations, also known as the von Foerster Equation [38],

$$\frac{\partial N}{\partial t} + \frac{\partial N}{\partial a} = -\mu N, \qquad N(0,a) = N_0(a) = S_0(a) + I_0(a).$$

Similar to the approach from section 2.5.1, we define new variables $S(t,a) = N(t,a)s(t,a)$, $I(t,a) = N(t,a)i(t,a)$, so that $s,\ i$ denote the susceptible and infected proportions of the population respectively. Then

$$\left(\frac{\partial}{\partial a} + \frac{\partial}{\partial t}\right) i(t,a) = \frac{I_t + I_a - \mu I}{N},$$

where $S_t,\ S_a$ denote the partial derivatives of S to t and a respectively, and again the death rate conveniently drops out. Since $s(t,a) + i(t,a) = 1$, we now only have to solve the transport equation

$$\left(\frac{\partial}{\partial a} + \frac{\partial}{\partial t}\right) i(t,a) = \alpha(t,a) - (\beta + \alpha(t,a)) i(t,a) \tag{10}$$

By integrating along the characteristic lines, the partial differential equation is reduced to an ordinary one. This method is called the method of characteristics. Since for every increase

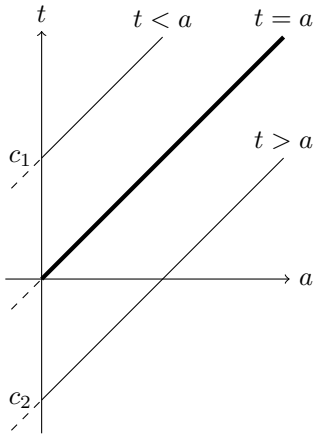


Figure 4: Characteristic lines $t(a) = a + c$.

in time we have an equal increase in age, this relation defines the characteristics as $\frac{dt}{da} = 1$, which are straight lines $t = a + c$ [39], illustrated in Figure 4. We will allow t to be positive as well as negative, so that the area that is considered includes the times of birth of each individual participating in the study. On these lines we have that

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}a} i(t(a), a) &= \frac{\partial i}{\partial t}\frac{dt}{da} + \frac{\partial i}{\partial a} = i_t + i_a, \\ &= \alpha(t(a), a) - (\beta + \alpha(t(a), a))i(t(a), a). \end{aligned} \tag{11}$$

Following the approach in section 2.5.1, we will denote $\phi(t(a), a) = \alpha(t(a), a) + \beta$ and $\Gamma(t(a), a) = \int_0^a \phi(t(s), s)\ \mathrm{d}s$ and use the corresponding version of (5) given by

$$\frac{\mathrm{d}}{\mathrm{d}a}\left[e^{\Gamma(t(a),a)}\right] = \phi(t(a), a)e^{\Gamma(t(a),a)} = (\alpha(t(a), a) + \beta)e^{\Gamma(t(a),a)}. \tag{12}$$

By multiplying both sides of the equation (11) by $e^{\Gamma(t(a),a)}$, and then using the chain rule and the first equality of equation (12), it is obtained that

$$\begin{aligned} \alpha(t(a), a)e^{\Gamma(t(a),a)} &= \frac{\mathrm{d}i(t(a), a)}{\mathrm{d}a}e^{\Gamma(t(a),a)} + i(t, a)\left[\phi(t(a), a)e^{\Gamma(t(a),a)}\right] \\ &= \frac{\mathrm{d}}{\mathrm{d}a}\left[i(t(a), a)e^{\Gamma(t(a),a)}\right]. \end{aligned} \tag{13}$$

Rewriting the left hand side of (13) with the second equality from Equation (12) results in

$$\frac{\mathrm{d}}{\mathrm{d}a}\left[i(t(a), a)e^{\Gamma(t(a),a)}\right] = \frac{d}{da}\left[e^{\Gamma(t(a),a)}\right] - \beta e^{\Gamma(t(a),a)}.$$

Integrating both sides with respect to a gives

$$i(t(a), a)e^{\Gamma(t(a),a)} = e^{\Gamma(t(a),a)} - \beta\int_0^a e^{\Gamma(t(s),s)}\ \mathrm{d}s + C.$$

From the initial condition $i(t(0), 0) = 0$ follows that $C = -1$, so that the solution is given by

$$i(t(a), a) = 1 - e^{-\Gamma(t(a),a)}\left(1 + \beta\int_0^a e^{\Gamma(t(s),s)}\ \mathrm{d}s\right). \tag{14}$$

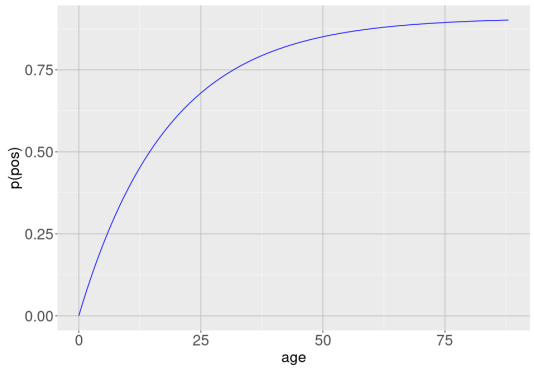


Figure 5: An example of the distribution of the prevalence for constant infection and recovery rates $\alpha = 0.06,\ \beta = 0.005$

2.5.4 Constant infection and recovering rates

In the most simple case, we assume that the force of infection $\alpha(a) = \alpha$ is constant and does not depend on age or time. This might be appropriate for diseases where the likelihood of infection is relatively uniform across the population. From (14), evaluating the integrals results in

$$i(t,a) = \frac{\alpha}{\alpha+\beta}\left(1 - e^{-(\alpha+\beta)a}\right).$$

An example of the distribution of the proportion of the population that is infected against age is shown in Figure 5.

2.5.5 Piecewise constant infection rate dependent on age and time

To capture the distinct effects of age and time on the infection rate and simplify the analysis, we assume that the effects of age and of time are independent. We assume that the age component of the force of infection is constant through time, and the time component is constant through age. We again partition the age range into intervals $A_i = [a_{i-1}, a_i),\ i \in \{1,\dots,m\}$ where $a_0 = 0$ and $a_m = a_{max}$ and the time range into intervals $T_j = [t_{j-1}, t_j),\ j \in \{1,\dots,n\}$ where t_0 denotes the point in time that is taken into account in the analysis, $-a_{\max}$, and $t_n = t_{P3}$, the time at which P3 was conducted. Now we let $\alpha(t,a) = \alpha_1(a)\alpha_2(t)$, with $\alpha_1(a)$ and $\alpha_2(t)$ constant on each of the intervals A_i and T_i respectively, so that

$$\alpha_1(a) = \sum_i \alpha_{1i}\mathbb{1}\{a \in A_i\}, \qquad \alpha_2(t) = \sum_j \alpha_{2j}\mathbb{1}\{t \in T_j\}.$$

Since $t(a) = a + c$, it follows that $\alpha(t(a), a)$ is given by

$$\begin{aligned}\alpha(t(a),a) &= \alpha_1(a)\alpha_2(a+c)\\ &= \sum_{i,j}\alpha_{1i}\alpha_{2j}\mathbb{1}\{a \in A_i \cap (T_j - c)\}.\end{aligned}$$

By integration, $\Gamma(a+c, a)$ is calculated as

$$\begin{aligned}\int_0^a \alpha_1(z)\alpha_2(z+c) + \beta\ \mathrm{d}z &= \beta a + \int_0^a \sum_{i,j}\alpha_{1i}\alpha_{2j}\mathbb{1}\{z \in A_i \cap (T_j - c)\}\ \mathrm{d}z\\ &= \beta a + \sum_{i,j}\alpha_{1i}\alpha_{2j}|A_i \cap (T_j - c) \cap [0,a]|\\ &= \beta a + \alpha_1^T A(c,a)\alpha_2, \qquad (15)\end{aligned}$$

where

$$A_{i,j}(c,a) = |A_i \cap (T_j - c) \cap [0,a]|.$$

Then it follows from (14) that

$$i(t,a) = 1 - \exp\{-(\alpha_1^T A(t-a,a)\alpha_2 + \beta a)\}\left(1 + \beta \int_0^a \exp\{\alpha_1^T A(t-a,z)\alpha_2 + \beta z\}\ \mathrm{d}z\right),$$

with

$$\int_0^a \exp\{\alpha_1^T A(c,z)\alpha_2 + \beta z\}\ \mathrm{d}z = \sum_{i,j} A_{i,j}(c,a) \exp\left\{\sum_{k\le i, l\le j} (\alpha_{1k}\alpha_{2l} + \beta) A_{k,l}(c,a)\right\}.$$

If we let the intervals A_i, T_j be of unit length (one year), then for $k \in \{A_i \cap (T_j - c)\}$, $k > 0$, we have $\alpha(k) = \alpha_1(k)\alpha_2(k+c)$ for given c, and so

$$\sum_{i,j} \alpha_{1i}\alpha_{2j} |A_i \cap (T_j - c) \cap [0,a]| = \sum_{k=1}^{a} \alpha_1(k)\alpha_2(c+k).$$

Therefore, for $a > 0$ we get

$$i(t(a),a) = 1 - \exp\left\{-\sum_{k=1}^{a} \alpha_1(k)\alpha_2(t(k))\right\}\left(1 + \sum_{k=1}^{a} \exp\left\{\sum_{j=1}^{k} \alpha_1(j)\alpha_2(t(j))\right\}\right). \tag{16}$$

2.6 Implementation

The models were implemented in the modelling language Stan [40], using the RStan interface [41] in R [42]. It operates with compiled C++; a parser translates a model expressed in the Stan language to C++ code, which is then compiled to an executable program. Stan uses the Hamiltonian Monte Carlo (HMC) method as sampling technique. This is a variation of the Metropolis algorithm, where a random walk through the parameter space is used to generate samples from a posterior distribution, that favors parameter values that have relatively high posterior probability. The proposal distribution in the HMC algorithm changes depending on the current position, by calculating the gradient of the posterior distribution. This improves sampling efficiency and reduces correlation between successive samples. [43] This is done through the potential function, which is the negative logarithm of the posterior density.

Each model parameter must have established convergence. There were several diagnostics used to determine the quality of the fit that is generated by Stan. These diagnostics are necessary but not sufficient. Stan utilizes the potential scale reduction statistic known as "R-hat" to assess convergence in Markov chains. The R-hat diagnostic is an estimation of the ratio between the overall variance and within-chain variance. Each chain is divided into two halves, ensuring agreement between the initial and latter halves. The variances are computed within each individual chain and collectively across all chains for comparison. As the Markov chains approach identical distributions, the R-hat statistic tends toward a value of 1, indicating convergence. If it is not approximately 1, then the Markov chains have not mixed well, and at least one of the chains is producing biased samples. The effective sample size n_eff is a measure of how informative the samples are. This size is usually less than the total number of sampling iterations due to correlation among the samples, but was made sure to be sufficiently large (n_eff / N $>$ 0.001). This correlation introduces redundancy, decreasing the precision of posterior estimates. The variance of Monte Carlo estimates is

conceptually equivalent to the variance expected from n_eff independent samples. Furthermore, the traces of each of the chains were examined to ensure good mixing and agreement among the chains. For each of the models, four chains were used. The number of iterations differed, with a minimum of 1000 iterations.

3 Results

3.1 Distribution of the Optical Density values

For each of the PIENTER studies, the mixture model was fitted to the log-transformed OD values for different transformations of the data to obtain the optimal prevalence estimations, depending on the availability; for PIENTER 1 and 3, the available data consisted of the optical density measurements from individuals (OD) and the corresponding cut-off (QC) measurements, both of which were corrected by the blank measurements on the corresponding plates by subtraction. For PIENTER 2, all of the original measurements from each plate were available.

Each of the models was fitted using 1000 iterations. An uninformed Beta$(1, 1)$ distribution was used as prior for the prevalence p. Normal distributions with a broad spread were used as weakly informative priors for the optical density means and variances for both the positive and negative populations.

3.1.1 PIENTER 1

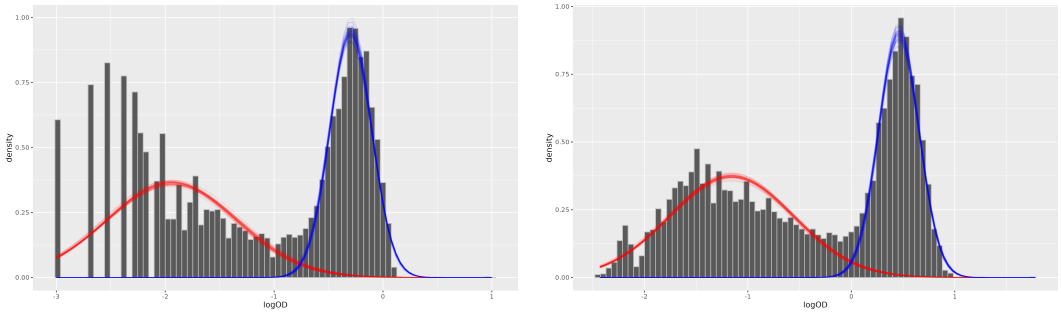
Both the corrected OD and OD/QC values were analyzed by fitting the mixture of two normal distributions. The priors for the means were chosen as $\mu_- \sim N(-2, 2)$, $\mu_+ \sim N(0, 1)$ in the first case, and $\mu_- \sim N(-1, 1), \mu_+ \sim N(0.5, 1)$ in the second case, and $\sigma_\pm \sim N(0, 1)$. The resulting distributions are shown Figure 6, the point estimates and 95% confidence interval of the parameters are given in Table 2. The lower range of the data in Figure 6a does not look Gaussian, since some of the measurement values in the data were extremely small or even negative, causing rounding errors when transforming to the log-scale. This could happen when the value is very close or lower than their corresponding blank control, and is likely due to measurement errors. Therefore, a better performance was achieved by estimation of the distributions for the OD/QC ratio; the scaling leads to less rounding errors, while the point estimations of the prevalence are similar. Both prevalence estimates are higher than the estimates found by Kortbeek et al. [15].

Table 2: Point estimates and 95% CI of means and prevalence for the different data corrections for PIENTER 1

data	μ_-	μ_+	p
OD - Blank	-1.94 (-1.96,-1.92)	-0.30 (-0.30,-0.29)	0.45 (0.44,0.47)
(OD - Blank) / (QC - Blank)	-1.16 (-1.18,-1.13)	0.46 (0.45,0.47)	0.44 (0.43,0.45)

3.1.2 PIENTER 2

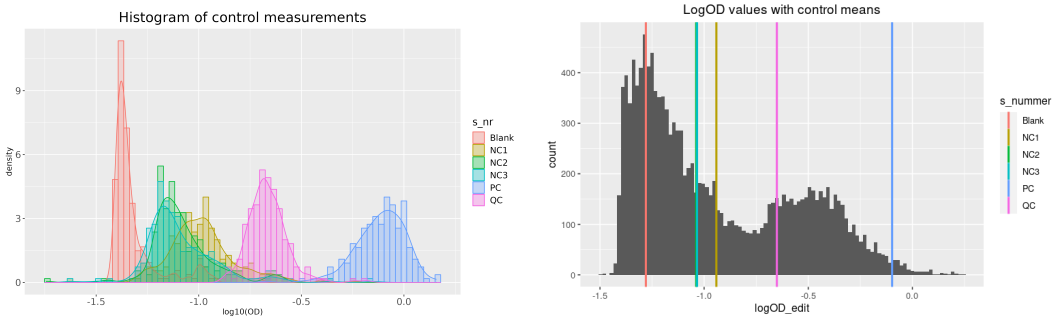
For the second study, PIENTER 2, the original plate data from the ELISA measurements, including the values for each of the control samples, was available. A histogram of the log-transformed OD values for the controls is depicted in Figure 7a. In the ideal situation, samples containing a serum do not have a lower value than the blank controls, so that the blank wells have the lowest value, and the sample used for determining the cut-off (QC) lies between the negative and the positive control values. However, we also see a lot of noise. The negative control values overlap with those of the cut-off samples, which would classify those measurements as positive. Unfortunately, these problems also arise in the measurement values of the participants in the study. A notable part of the measurements fall below the blanks as can be seen in Figure 7b, so that extracting blanks causes negative values. Low and negative values again cause issues when transformed to the log-scale. Without correction of the measurements with the blank values, the negative population is



(a) Log-transformed (OD - Blank) values

(b) Log-transformed (OD - Blank)/(QC - Blank) ratio values

Figure 6: Comparison of the distributions of the log-transformed OD values for PIENTER 1



(a) Log-transformed OD values of the plate-control samples for PIENTER 2.

(b) Histogram of log-transformed OD measurements with control means

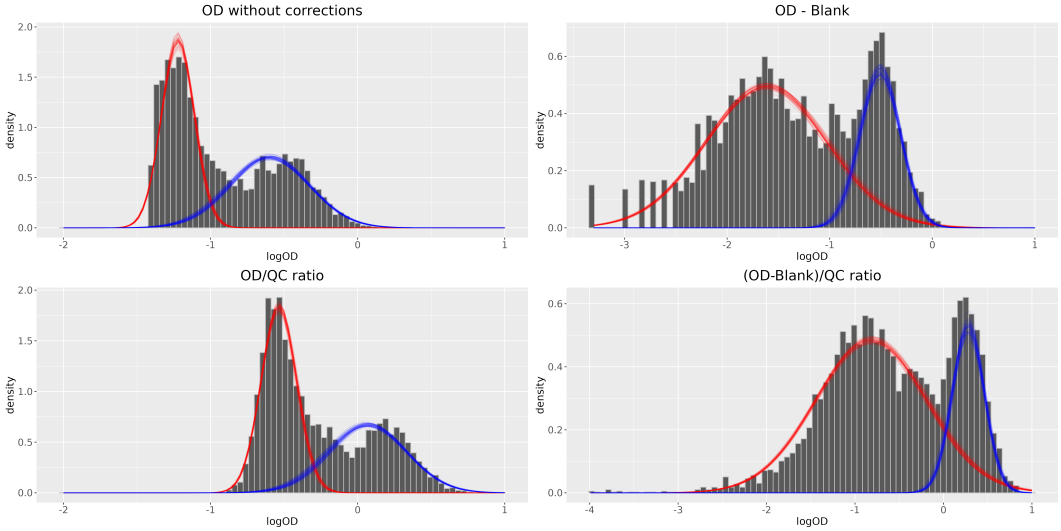
Figure 7: Histograms of the log-transformed OD values of the control samples, compared to the OD values from the population samples for PIENTER 2.

skewed. The distributions resulting from the different data transformations described are shown in Figure 8a. Point estimates and 95% confidence interval of prevalence and mean of the values of the positive and negative component of the distributions are shown in Table 3. The choice of these transformations highly influences the estimations of the prevalence. The estimation when using the ratio of the data corrected with blanks to the QC values is similar to the prevalence found by Hofhuis et al. [22], where the exact same transformation was used to determine the classification. However, the large overlap of the distributions makes the prevalence estimations unreliable and the model extremely sensitive to the choice of the priors for convergence.

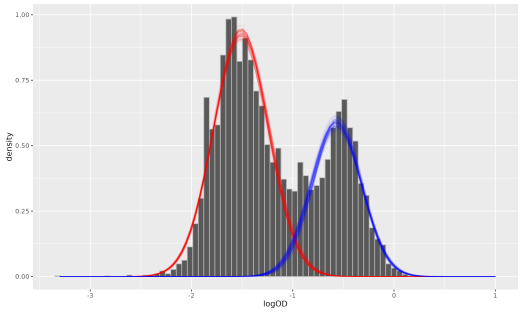
To avoid the issue of having negative values and to reduce skewness of the distribution for the negative population, we propose an alternative transformation on the data. Given that the data has been filtered to eliminate empty cells, it is reasonable to assume that all samples should produce positive measurement values. By adding a constant value to the OD measurements, we reduce skewness in the lower half of the data. This adjustment was determined empirically by examining the fit of the resulting distributions. The minimum value that eliminates negative OD values is 0.015, the fit does not improve by adding a higher value. The impact of this adjustment on the distributions are illustrated in Figures 8b and 8c.

Table 3: Point estimates and 95% CI of means and prevalence for the different data corrections for PIENTER 2

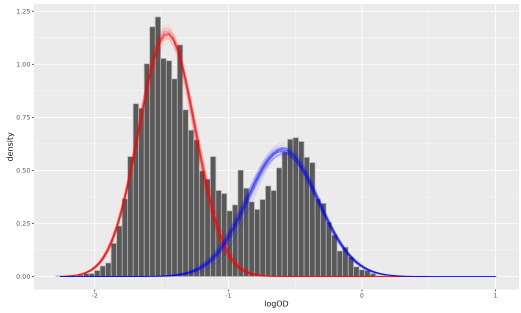
data	μ_-	μ_+	p
OD	-1.22 (-1.23,-1.22)	-0.60 (-0.62,-0.58)	0.48 (0.46,0.50)
OD - Blank	-1.62 (-1.65,-1.59)	-0.51 (-0.52,-0.49)	0.27 (0.25,0.29)
OD / QC	-0.54 (-0.54,-0.53)	0.07 (0.05,0.09)	0.45 (0.42,0.47)
(OD - Blank) / (QC - Blank)	-0.82 (-0.84,-0.79)	0.28 (0.27,0.29)	0.24 (0.22,0.26)
OD + 0.015	-1.51 (-1.52,-1.50)	-0.57 (-0.58,-0.55)	0.36 (0.35,0.38)



(a) Comparison of distributions of log-transformed measurements for multiple data transformations using the control samples



(b) Log-transformed OD values, shifted by 0.015



(c) Log-transformed OD values, shifted by 0.020

Figure 8: Comparison of distributions of the log-transformed OD values for PIENTER 2, with a transformation using the control samples or a shift on the measurements.

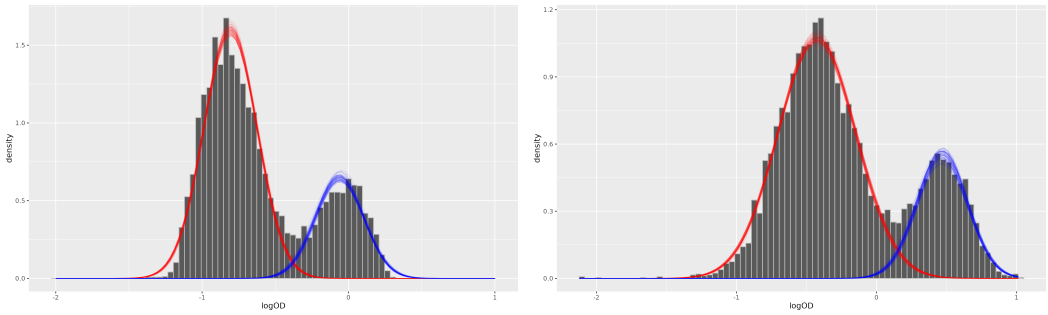
3.1.3 PIENTER 3

In the PIENTER 3 study, we had access to both OD values corrected for blanks and their ratios relative to QC values. The values of the blank measurements were not known. The distributions are shown in Figure 9. The estimations using the ratios performed better than those using only OD values, similar as for P1. The point estimates are given in Table 4. Both prevalence estimates are similar, but are lower than those estimated by Van den Berg

et al. [13].

Table 4: Point estimates and 95% CI of means and prevalence for the different data corrections for PIENTER 3

data	μ_-	μ_+	p
OD - Blank	-0.81 (-0.81,-0.80)	-0.06 (-0.07,-0.05)	0.27 (0.26,0.28)
(OD - Blank) / (QC - Blank)	-0.43 (-0.44,-0.42)	0.47 (0.46,0.48)	0.25 (0.24,0.27)



(a) Log-transformed OD values **(b)** Log-transformed OD/QC ratio values

Figure 9: Comparison of distributions of the log-transformed OD values for PIENTER 3

3.2 Prevalence within age categories

To study how the prevalence of the population changes through the ages, age categories are incorporated as a covariate in Equation (2) in the distributions on the log-transformed optical densities. The choice of the partition of the age range influences the uncertainty and accuracy of the estimations. Smaller ranges give higher uncertainty, while larger ranges give less information about the shape. For the interested reader, an illustration of these effects can be found in Appendix A, through a comparison of the distributions for different sizes of the intervals in Figure 18.

The estimated seroprevalence for each of the studies compared to the seroprevalences calculated with the cut-off method is shown in Figure 10, where for the age categories a width of 7 years is chosen[1]. The 95% and 50% quantile intervals are shown as ribbons. The line represents the prevalence as calculated from the cut-off method. A running average of width 2 is used for slight smoothing. The first study gives very similar results when estimated through the Bayesian method compared to the cut-off method; only for young individuals the estimations are significantly smaller. The estimations for the second study are higher than indicated by the cut-off method, while for the third study the estimations indicate a lower seroprevalence under the age of 42, but increased seroprevalence in higher age range.

3.3 Covariates

For each of the separate studies, the effect from the selection of factors given in Table 1 was analyzed in both the entire population and within age categories, by including those as covariates in the mixture model (2). The comparative analysis was done using the High-

[1]In the early 20$^{\text{th}}$ century, philosopher Rudolf Steiner developed a theory of human development based on 7-year cycles, corresponding to profound life changes. This is a coincidence.

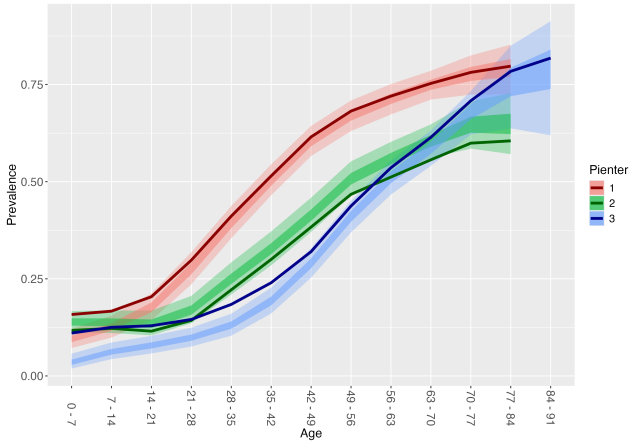


Figure 10: Estimation of seroprevalence, represented by the ribbons indicating 95% and 50% credible intervals, compared to the cut-off method, represented by lines, smoothed using a running average of width 2.

est Density Interval (HDI) and the Region of Practical Equivalence (ROPE) decision rule described in Section 2.3.1.

The distributions estimated on the entire population are fitted using 1000 iterations. The models where the prevalence is estimated on the age categories, resulting in smaller samples, is fitted using 2000 iterations. Dividing the population into different categories for each covariate decreases the sample size and thus increases the uncertainty. Therefore relatively large intervals of 10 year intervals are used.

3.3.1 PIENTER 1

The results from the ROPE + HDI analysis on the entire population are given in Table 5. When comparing the prevalence in the entire population, only education is identified as a risk factor; low education status is associated with higher risk of infection. For gardening and having a cat as pet, the method was indecisive; the *pd*-value is 1, indicating the existence of an effect, but it is too small to be of practical significance. Keeping farm animals (agriculture) was not associated with higher risk of infection when the entire population was considered.

We take a closer look at the effect of different risk factors on various age categories, as depicted in Figure 11. The interested reader can find the tables with the full analysis in Appendix B.1. The study finds that high edicational level is a risk factor for seropositivity in the age ranges of 20-30 and 60+. Agriculture was only identified as a risk factor in the 20-30 year old population, while no difference was found in children aged 0-10. Again, the method was indecisive for gardening; large *pd*-values indicate the existence of an effect between the ages of 40-60, however the HDI intervals are large. Having a cat as a pet was identified as risk factor from the age of 50 to 70 years old. In earlier analysis [15] factors were identified using odds ratios obtained through logistic regression. Our findings are in agreement on cat ownership and education level as a factor associated with seropositivity, but a definite relation with gardening and professional animal contact could not be identified through our method.

Table 5: PIENTER 1 HDI + ROPE analysis on the entire population. Here p_diff denotes the mean value of the distribution covar_1 - covar_2, the lower and upper columns give the lower and upper bounds of the 89% HDI, pd denotes the probability of direction.

covar	covar_1	covar_2	p_diff	lower	upper	pd	result
education	Low	High	0.16	0.14	0.19	1.00	Low higher
education	Middle	High	-0.02	-0.05	0.02	0.76	No difference
education	Middle	Low	-0.18	-0.21	-0.15	1.00	Low higher
gardening	No	Yes	-0.04	-0.06	-0.02	1.00	Undecided
agriculture	No	Yes	0.01	-0.01	0.05	0.79	No difference
petcat	No	Yes	0.04	0.02	0.06	1.00	Undecided

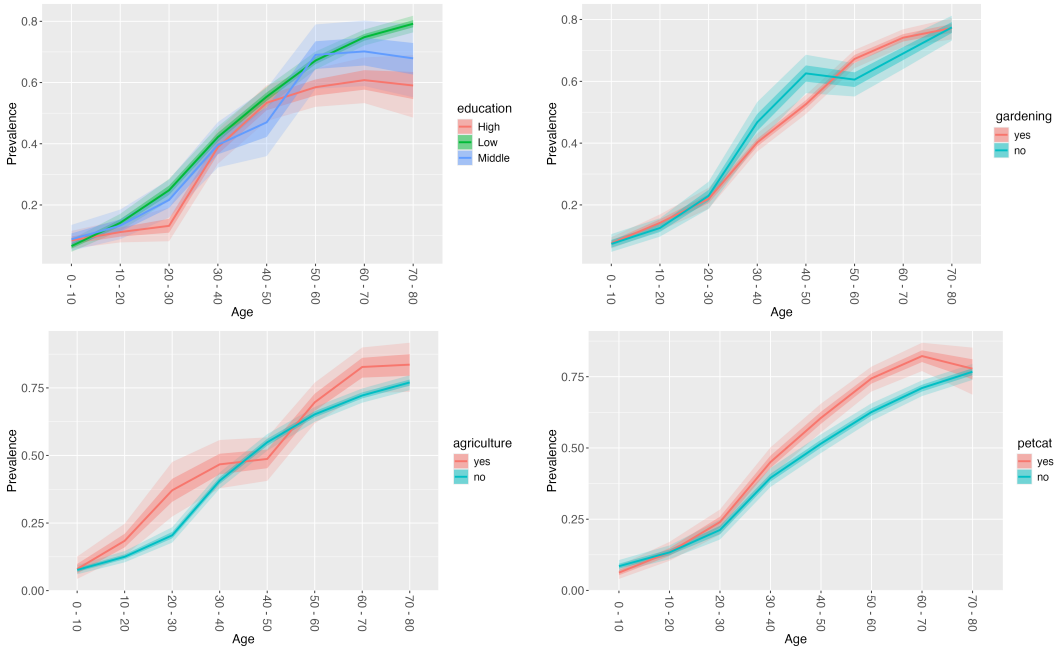


Figure 11: PIENTER 1 prevalence estimations by age groups and covariate: stratified analysis of education level, gardening, agriculture and cat ownership, where the 89% and 50% QI are represented by ribbons.

3.3.2 PIENTER 2

The results on the entire population from HDI+ROPE analysis are found in Table 6. Again, only having a low education level is a identified factor on population level, and gardening, keeping farm animals (agriculture), having a cat as pet (petcat) and eating raw, unwashed vegetables (unwashedveget) were not identified as risk factors.

We take a closer look at the effect of the different risk factors on various age categories, as illustrated in Figure 12. The interested reader can find the tables with the full analysis in Appendix B.2. Although the effect of having low education level is visible in the graph, the uncertainty is too high to indicate when dividing into age categories of 10 years. Gardening is clearly not a risk factor, as the graph shows almost identical lines and the 89% HDI contains 0 in almost all cases, with low *pd*-values. For agriculture, there is too much uncertainty and so no statements could be made. This is unlikely to be associated with prevalence,

Table 6: PIENTER 2 HDI + ROPE analysis on the entire population. Here p_diff denotes the mean value of the distribution covar_1 - covar_2, the lower and upper columns give the lower and upper bounds of the 89% HDI, pd denotes the probability of direction.

covar	covar_1	covar_2	p_diff	lower	upper	pd	result
education	Low	High	0.11	0.09	0.14	1.00	Low higher
education	Middle	High	-0.02	-0.04	0.01	0.82	No difference
education	Middle	Low	-0.13	-0.15	-0.10	1.00	Low higher
gardening	no	yes	-0.03	-0.05	-0.01	0.99	No difference
agriculture	no	yes	0.01	-0.03	0.04	0.59	No difference
petcat	no	yes	0.01	-0.01	0.03	0.69	No difference
contcat	no	yes	0.04	0.02	0.06	1.00	Undecided
vegetarian	no	yes	0.10	0.03	0.17	0.98	Undecided
rawmeat	no	yes	0.06	0.04	0.09	1.00	Undecided
rawporkmeat	no	yes	-0.07	-0.10	-0.04	1.00	Undecided
rawbeef	no	yes	-0.03	-0.05	-0.01	0.99	Undecided
unwashedveget	no	yes	0.01	-0.01	0.03	0.75	No difference

consistent with earlier analysis [22]. In the graph for having a cat as pet, the prevalence for having a cat as pet seems to be higher from around 40 years, but the effect is identified by the HDI + ROPE strategy from age 70. Also having contact with cats (reported by over half of the population) is a risk from a similar age, and the graph looks nearly identical to that of having a pet cat. Here the *pd*-value is already close to 1 from the age of 40, meaning that there is high proportion of the distribution for ”yes” that is higher than that of ”no”. Having a pet cat and having contact with cats were also identified as risk factors in earlier analysis [22]. Eating raw meat seems to be related to higher prevalence from the age of 50 but this is identified as risk factor with HDI + ROPE only for the age category of 60-70 years. Raw pork is identified as a risk factor from the age of 50 but raw beef seems to be less of a risk, and is identified as risk from 70 years. Unwashed vegetables seem to be associated with higher seroprevalence from older age, which is also indicated by high *pd*-value, but there is too much uncertainty for identification as a factor.

3.3.3 PIENTER 3

The results on the entire population from HDI + ROPE analysis are given in Table 7. Again, having a low education level is associated with higher prevalence. Having a cat as pet and eating unwashed vegetables seems negatively associated with higher prevalence. Gardening, having contact with cats, eating raw meat or pork are not associated with difference in prevalence.

A closer look at the effect for different age categories the findings is shown in Figure 13. The interested reader can find the tables with the full analysis in Appendix B.3. It seems that low education level has higher prevalence from the age of 20, however the HDI + ROPE decision rule could not give absolute conclusions. Also, for none of the other covariates a clear difference could be identified. For agriculture, too much uncertainty makes the intervals too large. Eating raw pork meat between the ages of 50 and 70 and eating raw beef or raw meat in general between the ages of 60 and 70 seem to be associated with higher prevalence; the *pd*-value is close to 1 but the HDI intervals are large.

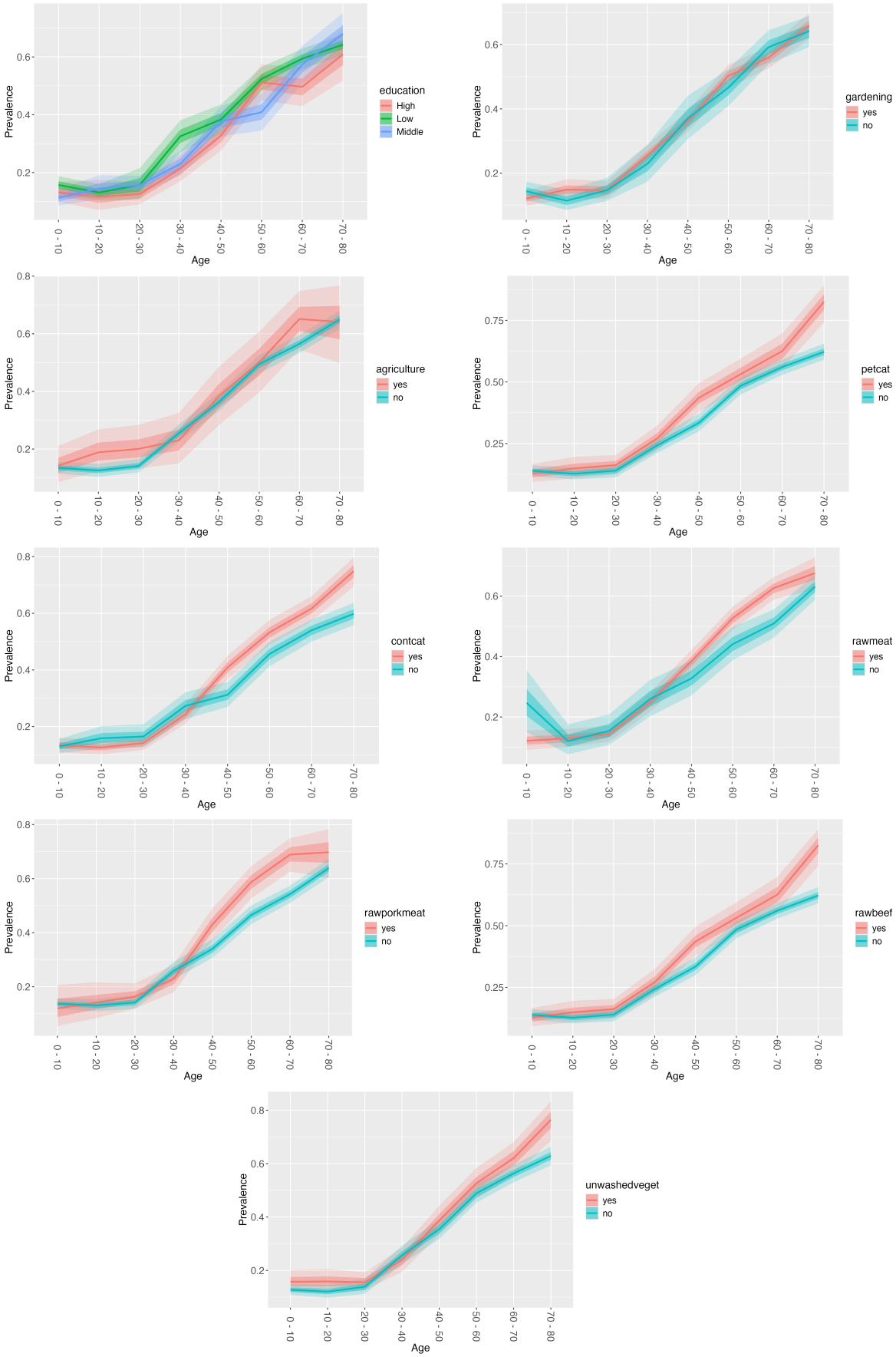


Figure 12: PIENTER 2 prevalence estimations by age groups and covariate: stratified analysis of education level, gardening, agriculture, cat ownership, contact with cats, eating raw meat, pork, beef and unwashed vegetables, where the 89% and 50% QI are represented by ribbons.

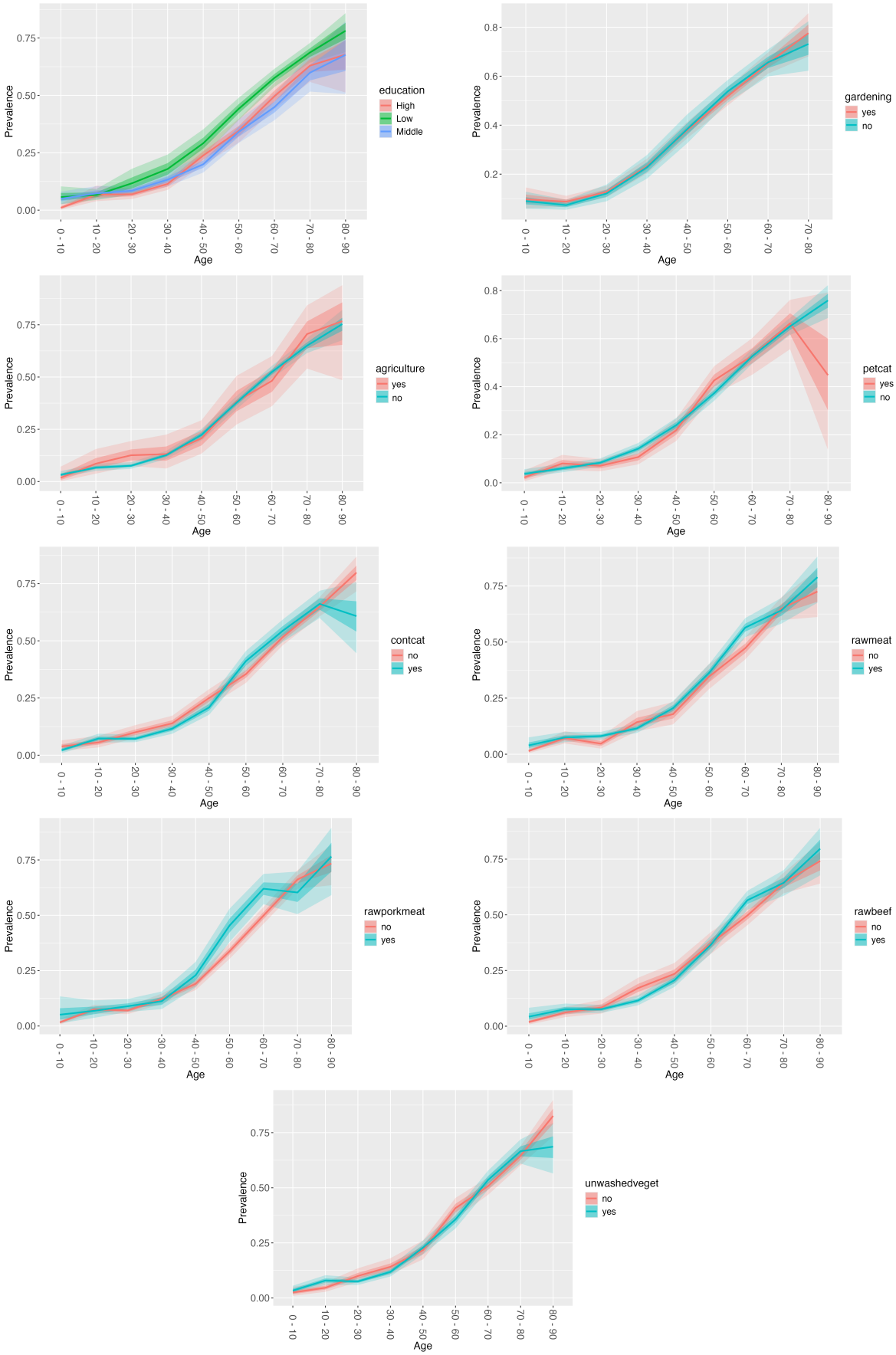


Figure 13: PIENTER 3 prevalence estimations by age groups and covariate: stratified analysis of education level, gardening, agriculture, cat ownership, contact with cats, eating raw meat, pork, beef and unwashed vegetables, where the 89% and 50% QI are represented by ribbons.

Table 7: PIENTER 3 ROPE+HDI analysis on the entire population. Here p_diff denotes the mean value of the distribution covar_1 - covar_2, the lower and upper columns give the lower and upper bounds of the 89% HDI, pd denotes the probability of direction.

covar	covar_1	covar_2	p_diff	lower	upper	pd	result
education	Low	High	0.17	0.14	0.19	1.00	Low higher
education	Middle	High	-0.03	-0.06	-0.01	0.99	Undecided
education	Middle	Low	-0.21	-0.23	-0.18	1.00	Low higher
gardening	no	yes	-0.02	-0.04	-0.00	0.95	No difference
agriculture	no	yes	0.04	0.00	0.07	0.94	Undecided
petcat	no	yes	0.08	0.06	0.10	1.00	No higher
contcat	no	yes	0.00	-0.04	0.05	0.53	No difference
rawmeat	yes	no	-0.02	-0.04	0.00	0.93	No difference
rawporkmeat	yes	no	0.02	-0.01	0.04	0.84	No difference
rawbeef	yes	no	-0.04	-0.06	-0.02	1.00	Undecided
unwashedveget	yes	no	-0.09	-0.11	-0.07	1.00	no higher

3.4 Force of infection

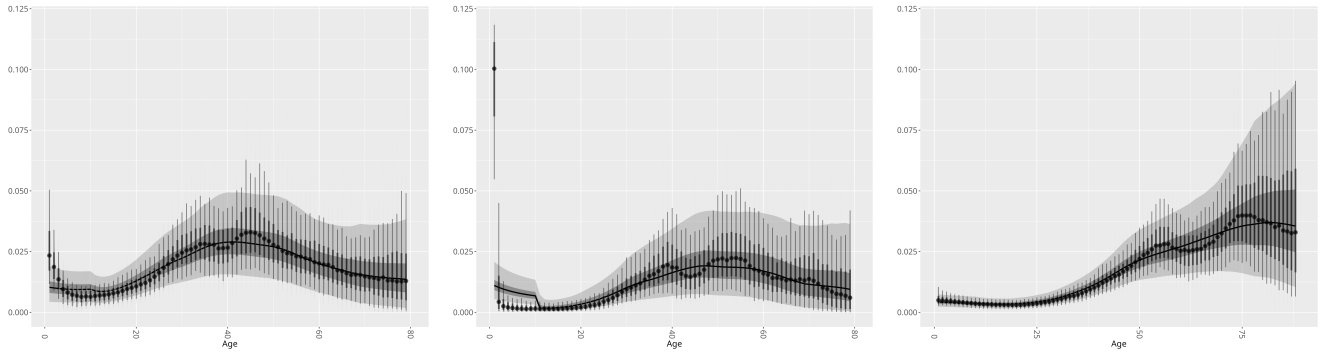
We now extend the model with the compartmental disease dynamical systems from Section 2.5. We first analyse the age-dependent force of infection $\alpha(a)$ for each PIENTER study independently, disregarding temporal dynamics in Section 3.4.1. The case where the rate β at which individuals become susceptible is zero is compared to the situation where β is non-zero. We then expand the model to include time dependency in the force of infection $\alpha(a,t)$ in Section 3.4.2. For each of the models, fitting was done using 1000 warm-up and 1000 sampling iterations.

3.4.1 Age-dependent force of infection

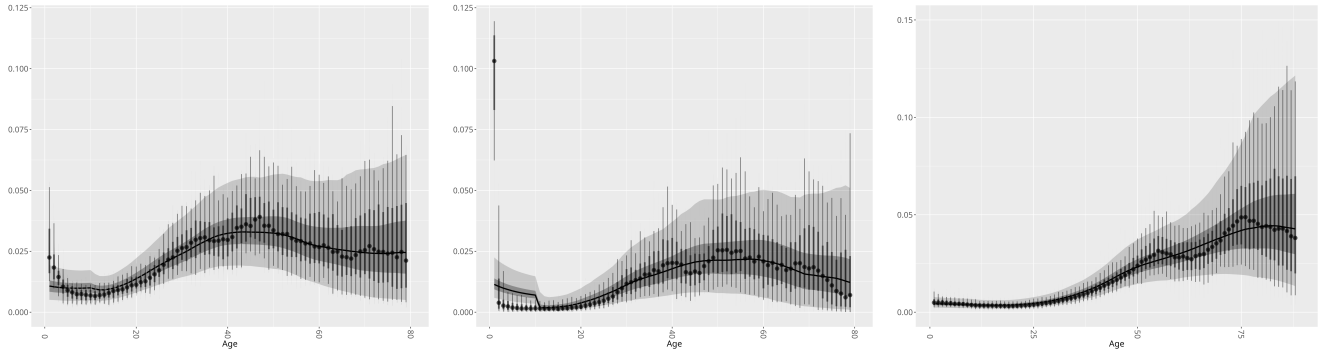
We take age intervals of unit length, so that the proportion of the population that is infected is characterized through Equation (8). In the scenario where the rate β is greater than 0, the value is expected to be relatively small. Therefore a normal prior $N(0, 0.005)$ is selected for β, centering the estimate around 0. To minimize over-fitting for α, subsequent values are assumed to lie within a small neighborhood of each other. The method is inspired by the idea of a random walk. We assume that the values have the relation $\alpha_{i+1} = \varepsilon_i \alpha_i$ for $1 \leq i < a_{\max}$, where $\varepsilon_i \sim N(1, \gamma)$, with γ having inverse gamma prior with shape parameter 1 and scale parameter 0.005, so that the distribution of ϵ will be centered around 1. Since individuals are born susceptible, the rate α_1 at which individuals become infected between birth and age one is assumed to be small and so a normal prior $N(0, 0.3)$ centered around 0 is chosen.

The resulting estimations of the age-dependent force of infection are shown in Figures 14a - 14f, with mean values indicated by dots, the 95% Quantile Intervals (QI) by bars, and the ribbons represent 50% and 95% QI, smoothed with a 20-year moving average. The specific findings for P3, as shown in Figure 14c mirror those reported by Van den Berg et al. [13]. For each study, the shape is consistent between the $\beta = 0$ and $\beta \neq 0$ scenarios, but the estimations are marginally higher and uncertainty is slightly increased in the case when $\beta \neq 0$, as can be seen from Figures 14g - 14i, picturing the 50% quantile intervals comparisons. Furthermore, the distributions of γ are similar across scenarios, as can be seen from Figures 14j - 14l The estimated values for β, as shown in Figures 14m - 14o, are minimal and tend to cluster around 0.

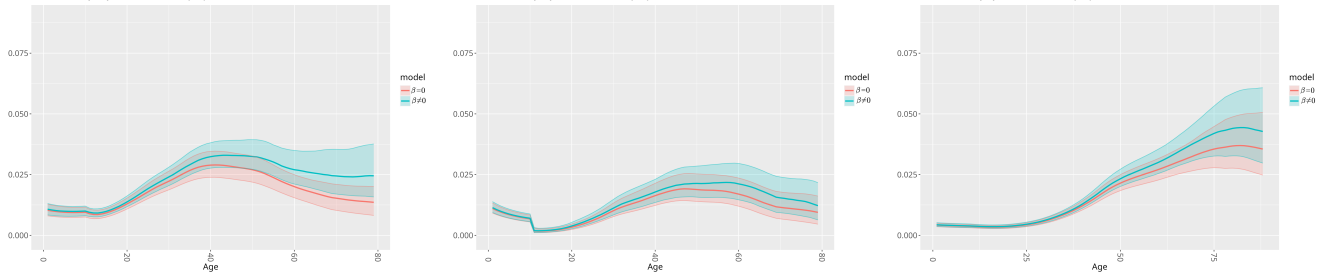
In the force of infection estimations of P1 and P2, there is a noticeable peak that shifts



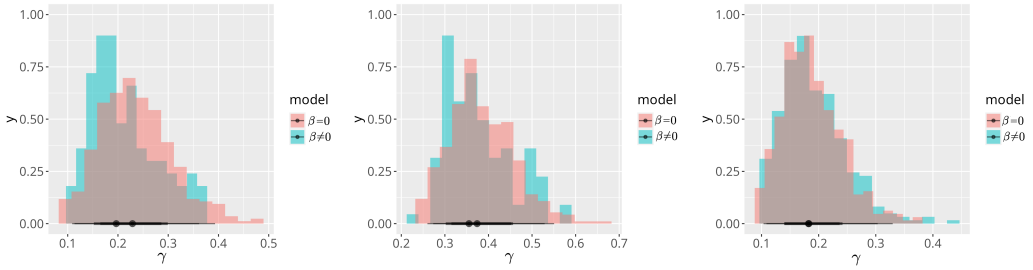
(a) P1: $\alpha(a)$, where $\beta = 0$ **(b)** P2: $\alpha(a)$, where $\beta = 0$ **(c)** P3: $\alpha(a)$, where $\beta = 0$



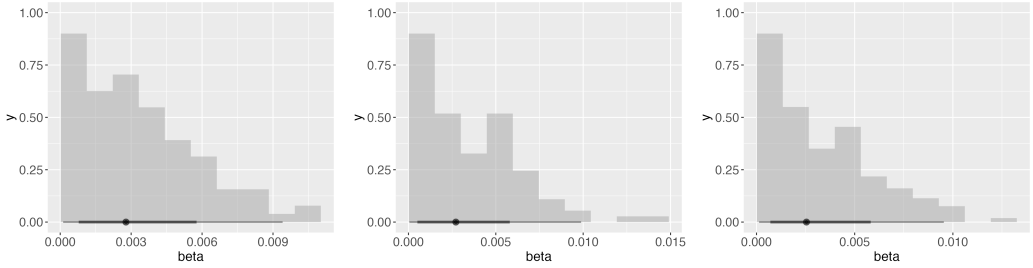
(d) P1: $\alpha(a)$, where $\beta \neq 0$ **(e)** P2: $\alpha(a)$, where $\beta \neq 0$ **(f)** P3: $\alpha(a)$, where $\beta \neq 0$



(g) P1: 50% QI for $\alpha(a)$ **(h)** P2: 50% QI for $\alpha(a)$ **(i)** P3: 50% QI for $\alpha(a)$

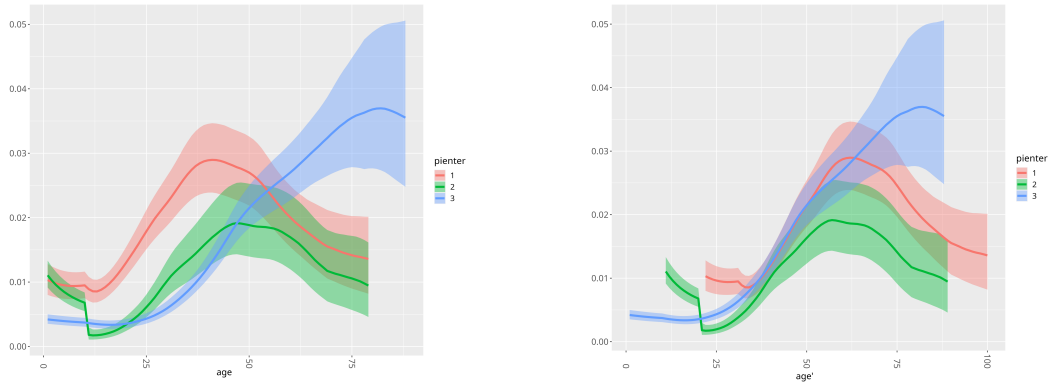


(j) P1: distributions of γ **(k)** P2: distributions of γ **(l)** P3: distributions of γ



(m) P1: distribution of β **(n)** P2: distribution of β **(o)** P3: distribution of β

Figure 14: Estimated values and distributions of the parameters for the age-dependent compartmental disease model, fitted for each PIENTER study independently, for both the case when $\beta = 0$ and $\beta > 0$. In (a)-(f), dots indicate the mean values, the bars represent the 95% QI and ribbons show 50% and 95% QI, smoothed with a 20-year running average. In (g)-(i), the smoothed 50% QI are depicted for both the case where $\beta = 0$ (red) and $\beta > 0$ (blue).



(a) Values of α for $\beta = 0$.

(b) Values of α, shifted so that individual ages align for $\beta = 0$.

Figure 15: Comparison of the estimated distributions of the age-dependent infection force $\alpha(a)$ for $\beta = 0$, fitted to each of the PIENTER studies independently.

towards older age groups in successive studies. This pattern seems to continue through P3, although truncated at the peak. Further examination, as illustrated in Figure 15a, shows the values of the force of infection for each of the PIENTER studies. This illustrates that in P2, the estimated force of infection is significantly reduced in younger age groups relative to P1. This trend does not continue in P3. Notably, P3 exhibits an increase in the force of infection from the age of approximately 45. Figure 15b aligns the age categories across the studies through time. Contrary to the decrease noted in P2 when compared to P1, P3 demonstrates an elevation in infection force post-70 years. This suggests a maintained low prevalence in the population for an extended duration, followed by an increase in older age. The shift in the peaks across the three studies may reflect changes in transmission dynamics over time.

3.4.2 Age- and time-dependent force of infection

We take age- and time intervals of unit length, so that the proportion of the population that is infected is given by Equation (16). From analysis on the age-scale, it may be conclude that β does not affect the estimations of the force of infection considerably. For practical considerations, to ensure convergence of the model, $\beta = 0$ is assumed.

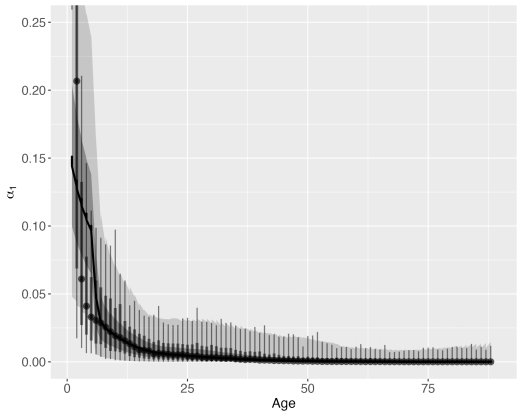
Similar to the approach in Section 3.4.1, to minimize potential overfitting of α, both the time- and the age component of α are informed through the relations

$$\begin{aligned} \alpha_1(a+1) &= \varepsilon_1(a)\alpha_1(a), & 1 \leq a < a_{\max} \\ \alpha_2(t+1) &= \varepsilon_2(t)\alpha_1(t), & t_1 \leq t < t_{\max}, \end{aligned}$$

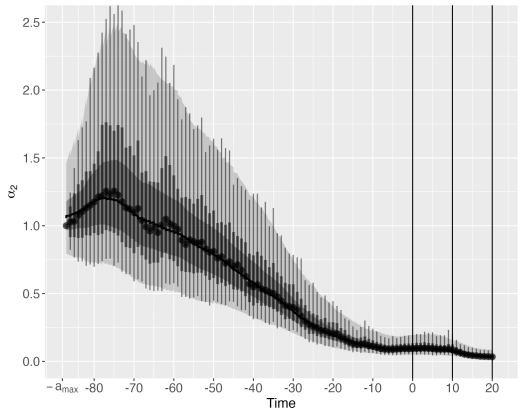
where $t_0 = -a_{\max}$ representing the initial timepoint, and $t_{\max} = 20$ corresponding to the timing of the P3 study. The parameters $\varepsilon_1(a)$, $\varepsilon_2(t)$ are modeled as $\varepsilon_i \sim N(1, \gamma)$, where γ_i is informed by an inverse gamma prior with shape parameter 1 and scale parameter 0.005, positioning the ϵ_i values close to 1. For the initial value for the age component $\alpha_1(1)$ again a normal prior $N(0, 0.3)$ is chosen, while the initial value for the time component $\alpha_2(t_1)$ is fixed at 1 to ensure identifyability.

The resulting estimations are shown in Figure 16, where for Figures 16a - 16d, mean estimates are represented by dots, the 95% Quantile Intervals (QI) by bars and the ribbons

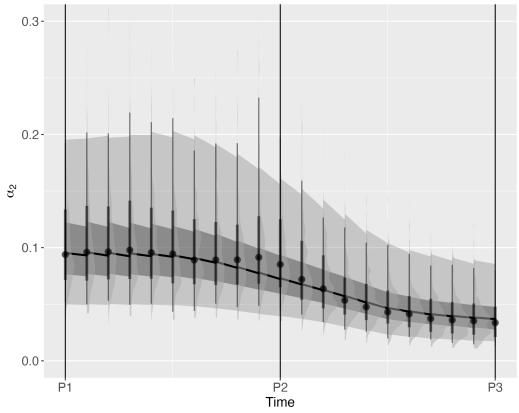
represent the 50% and 95% QI, smoothed with a 10-year moving average for clarity. Estimates of the age- and time components of the force of infection are shown in Figures 16a and 16b respectively. Surprisingly, the shape of the age-component differs from the age-dependent infection force estimated in Section 3.4.1. High uncertainty in the lower age-range suggests that the assumption that individuals are born susceptible is potentially incorrect, meaning that the boundary conditions in the model (Equation (9)) need to be revised. The time component $\alpha_2(t)$ shows a decreasing trend as time approaches the P3 study. The peak in the infection force seen in Section 3.4.1 that is absent in the age component $\alpha_1(a)$ could be determined by the time-component, but due to high uncertainty in the age-component, no definitive conclusions can be made. Concentrating on the period between P1 and P3, as illustrated in Figure 16c, the time component of the force of infection $\alpha_1(t)$ appears relatively stable, with only a subtle decrease towards the P3 study, although this decrease occurs mainly in the variability. This suggests a limited impact from temporal factors within this interval. This could suggest that any interventions or changes in population behavior have had a stable or sustained impact on the infection dynamics over the time span considered. The prevalence estimations are illustrated in Figure 16d, Those estimations are different from the estimations that are generated through the simpler model without infection dynamics, that were illustrated in Figure 10 in Section 3.2. A shift along the age axis is found rather than a change in overall prevalence levels. The estimations for the age specific random walk variance, γ_a, shown in Figure 16e are slightly higher than those found for the age specific variance found in Section 3.4.1, which could suggest more substantial fluctuations in infection rates with age. This could however also be influenced by the high uncertainty of $\alpha_1(a)$ in the lower age-range. The values for γ_t are even lower, hinting at less variability over time and possibly a stronger age effect than time effect in the dynamics of the infection force.



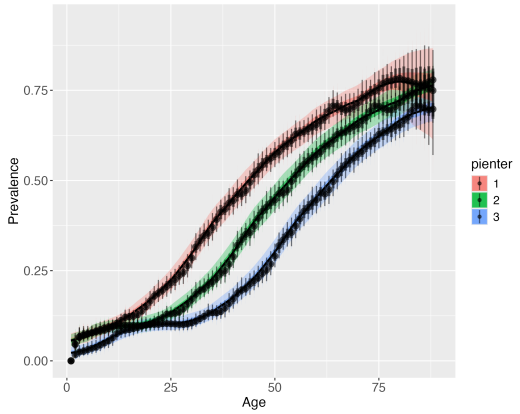
(a) Age-component, $\alpha_1(a)$



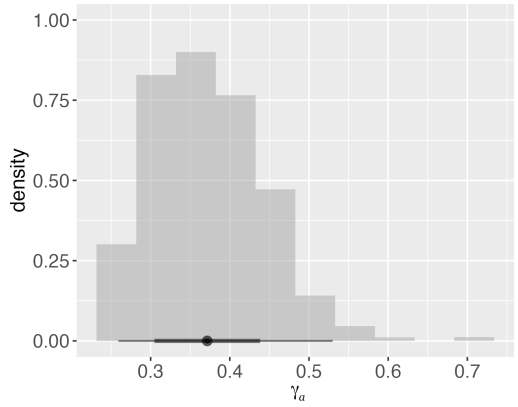
(b) Time component, $\alpha_2(t)$, with P1-P3 marked through lines



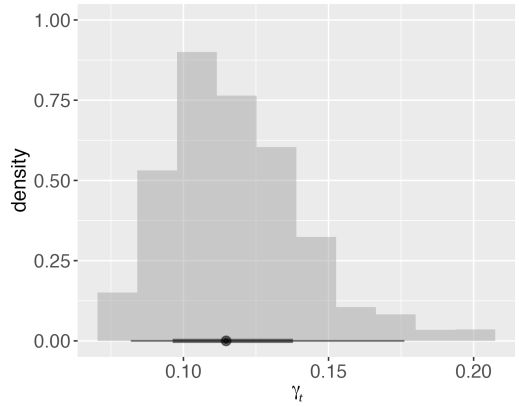
(c) $\alpha(t)$ for t between P1 and P3



(d) Estimation of prevalence p



(e) Random walk variance γ_a



(f) Random walk variance γ_t

Figure 16: Estimations of parameters of the age- and time dependent compartmental model. In (a)-(d), the dots represent mean estimates, the bars represent 95% QI, the ribbons represent 95% and 50% QI, smoothed with a 10-year moving average.

4 Discussion

4.1 Distribution of the data

In this analysis, it is assumed that the optical densities (OD) of the antibody titres are distributed as a mixture of two symmetric log-normal distributions. However, Jacobson [36] suggests that the distribution of the negative population often has a long right tail, which is strengthened by our findings in Section 3.1. Moreover, to obtain the OD values for each individual, both of their measurements are first corrected with the average value of the blank control measurements of the corresponding plate, and then the result is averaged. It is reasonable to assume that control measurements are log-normal as well. However, the sum of log-normal distributions is not log-normal itself, and does not follow a known probability distribution. In fact, even a shifted log-normal distribution does not follow a normal distribution on the log-scale, as demonstrated in Figure 17. This means that each of the transformation to calculate the OD values influences the log-normality of the distributions and possibly adds skewness. Previous studies [44, 45] have shown that a log-skew-normal distribution effectively approximates the sum of log-normal distributions, indicating that instead of a symmetric log-normal distribution, a skewed log-normal distribution may be more suitable for modeling the OD values.

In Section 3.1.2 we showed that the data that was available for P2 was affected by a lot of noise, and that the choice of value transformations highly influenced the resulting prevalence estimation. There are two types of errors that can be identified. Firstly, there was a high level of variability in the range of range of values, including differences in control measurements across the different plates, indicating a high plate-to-plate variation. Secondly, each type of control sample was tested twice on each plate, and blank wells containing no antibodies should produce the lowest values on the plate. Differences between repeated control measurements on the same plate discloses information on the on-plate variation and background noise. The model can be extended to reduce noise by adding distributions on these errors. wOn-plate variations are estimated through variation in repeated control measurements on the same plate, while plate-to-plate variations are estimated through the values of the control samples across different plates. These distributions could help account for the variability and noise in the data, allowing for more accurate results.

In our approach in estimating the age-specific seroprevalence, the age intervals have a predetermined, fixed width. The width was empirically established by comparing uncertainty

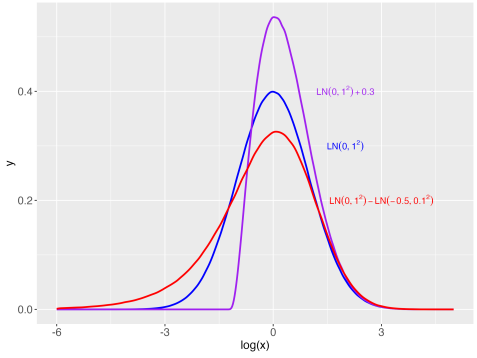


Figure 17: Density of a log-normal random variable (blue), compared to a shifted log-normal (ppurple) and a sum of log-normal random variables (red), resulting in skewed densities.

and level of detail in the resulting estimations. This strategy can be improved by extending the model to including the choice of the partition of the age range as a parameter. This could also be useful to improve the understanding of behavioral- and life-style changes that influence the risk of infection when examining the effect of covariates, and to assess the effectiveness of prevention methods and/or public awareness campaigns.

The analysis did not account for the representativeness of the sample population in relation to the full population in the Netherlands. This has been addressed in earlier analysis [13, 15, 22] by weighting the seroprevalence within municipalities for age, gender, ethnicity and urbanisation degree. The absence of such correction could have a small contribution to the difference in the prevalence estimations that we found in Section 3.2 from previous results.

4.2 Covariates

In the analysis of the effect of the covariates on the prevalence, the HDI + ROPE method combined with the *pd*-value was often not able to discern if an effect was present or absent. The ROPE metrics are sensitive to the sample size in the case where no effect is observed, and is therefore better suited for providing evidence for the absence of an effect [29]. On the other hand, the *pd*-value quantifies the amount of evidence that is observed for the presence of a true effect, in which case it is sensitive to the sample size. It is not able to reflect the amount of evidence of the absence of an effect; a high values suggest the existence of an effect, but low values do not give information on the certainty that no effect is present. Combining both indices should therefore result in an effective method to analyse risk factors. However, for both the ROPE and the *pd*-values, variability increases as noise increases, possibly indicating that small fluctuations in observed effects can significantly alter these measures [29]. Furthermore, the ROPE faces practical challenges in application due to the subjective nature of defining the non-significant effect range. This range, although arbitrary, imposes strict boundaries that may not accurately reflect the continuous nature of the data; a slight shift across the threshold changes the interpretation of results from negligible to significant. Additionally, ROPE indices are sensitive to changes in the scale of predictors; for example, another base for the logarithm would result in different ranges. To avoid these challenges, Bayes Factors offer an interesting alternative. Bayes Factors are able to provide clear evidence both for and against the presence of an effect. Their straightforward interpretation as odds favoring one hypothesis over another makes them effective for communication. It is the Bayesian equivalent to the approach used in the identification of risk factors in earlier analysis of the PIENTER studies [13, 15, 22]. However, Bayes Factors also face criticism for their sensitivity to the choice of priors, which can significantly influence their outcomes.

4.3 Disease dynamics

In Sections 2.5.1 and 2.5.3 we found the analytical solution to the differential equations, where α was assumed to be piecewise continuous, giving very few restrictions on the shape of α. Other types of functions for the force of infection are difficult to implement when using the Stan interface if the use of analytical solutions is desired. For example, if we study the simple case where α is a linear function, $\alpha(a) = k \cdot a + c$, then to calculate the age dependent

prevalence (7) we solve

$$\int_0^a \exp\left\{\frac{ks^2}{2} + (c+\beta)s\right\}\,\mathrm{d}s = \exp\left\{-\frac{(c+\beta)^2}{2k}\right\}\int_0^a \exp\left\{\left(\frac{ks+c+\beta}{\sqrt{2k}}\right)^2\right\}\,\mathrm{d}s$$
$$= \exp\left\{-\frac{(c+\beta)^2}{2k}\right\} \underbrace{\sqrt{\frac{\pi}{2k}}}_{\text{from: } \frac{\sqrt{\pi}}{2}\sqrt{\frac{2}{k}}} \operatorname{erfi}\left(\frac{ka+c+\beta}{\sqrt{2k}}\right)$$

where erfi is the imaginary error function

$$\operatorname{erfi}(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{x^2}\,\mathrm{d}x + C.$$

However, RStan does not support the use of imaginary arguments to the error function. When considering even more complex functions, an analytical solution does not always exist, or will become increasingly more difficult to calculate. A numerical solver could be implemented to investigate other suitable functions for α.

The age-dependent forces of infection observed in each of the independent PIENTER studies, discussed in Section 3.4.1, display a consistent shape with a peak that appears to shift rightward in each subsequent study. Adding a time component to this force might be expected to show a similar pattern. However, the age component of the force of infection in the model that combines the studies, analyzed in Section 3.4.2, is fundamentally different. This suggests that the assumption of a multiplicative relationship between age and time components of the infection force $\alpha(t,a) = \alpha_1(a)\alpha_2(t)$ may be incorrect. Additionally, the estimated prevalence p from this complex model deviates from that of the simpler model in Section 3.2 that only uses serological data to determine the distributions, indicating that enforcing this specific relationship on the infection force does not accurately reflect real-world conditions. Alternative relationships should be explored.

The age- and time dependent model was also highly sensitive to the choice of priors and initial values. Specification of initial values had a significant impact on the ability to obtain convergence of the outcomes, suggesting a complex and poorly fitting distribution landscape. The use of one-year age categories dramatically increases the number of variables that need to be estimated compared to the number of data points in the data set. The resulting high uncertainty further complicates model stability. Despite attempts, we were unable to achieve model convergence with age categories of larger width in the age- and time dependent model; for uniformity, we chose to use the same partitioning of the age range in both sections discussed. Convergence issues also persisted in scenarios where $\beta > 0$ in the complex model, though findings from Section 3.4.1 suggest that $\beta > 0$ is unlikely. Future research should consider using broader age categories to address these issues.

The estimated force of infection displayed unusual high values for newborns, as can be seen in Figures 14a - 14f and 16a. This suggests that the assumption that newborns are only born in the susceptible compartment, and thus have no detectable antibodies, is not suitable. Newborns may receive temporary protection through antibodies received from the mother during her pregnancy [46]. These newborns may be considered as an extra compartment with passive immunity, referred to as maternal immunity (M). The dynamics of the MSIS

model is then given by (adjusted from [47], eq 5.1)

$$\left(\frac{\partial}{\partial a}+\frac{\partial}{\partial t}\right)M(t,a) = -(\delta+\mu(a))M(t,a),$$
$$\left(\frac{\partial}{\partial a}+\frac{\partial}{\partial t}\right)S(t,a) = \delta M(t,a) - (\mu(a)+\alpha(a,t))S(t,a) + \beta I(t,a),$$
$$\left(\frac{\partial}{\partial a}+\frac{\partial}{\partial t}\right)I(t,a) = \alpha(a,t)S(t,a) - (\mu(a)+\beta)I(t,a),$$

where again $N = M + S + I$ is the total population. A numerical solver should be implemented to fit the dynamics to the data. Since the maternal immunity will fade after some time, this approach is better suited than using an initial value $I(t,0) > 0$.

5 Conclusion

We have studied the use of Bayesian statistical modelling in the prevalence estimation of the human population over age and over time for *Toxoplasma gondii*. We have made a mixture of two normal distributions to describe the distribution of the log-transformed optical density measurements from ELIZA. Furthermore, we described a dynamical system that estimates the force of infection over time and over age. There are still some challenges to overcome to improve the model, such as reduction of noise and improvements of the model and the identification in risk factors, but overall a Bayesian statistical approach performs very well in these type of situations. In fact, the observed distributions and noise in the data underline the Bayesian approach as a more effective tool than cut-off methods.

We have observed that the seroprevalence decreased from 44% (95% CI 43 - 45) in 1995/1996 to 36% (95% CI 35 - 38) in 2006/2007, and decreased further to 25% (95% CI 24 - 27) in 2016/2017, which is in contradiction to earlier results [13, 15, 22] where a decrease and then a slight increase was observed.

Risk factors were clearly identified in the first two studies as having low education level, having a cat as pet, eating raw pork meat and having contact with cats (both not reported in P1), but these effects did not seem to continue into the third study, indicating that policies or behavioral changes that reduce risk of infection through these behaviors were successful. This belief is strengthened by the findings for the time component of the force of infection, that is slightly declining towards P3.

Finally, we found that it is likely that infection is lifelong, and there is no real evidence to support the existence of a non-zero rate from the infected population to the susceptible compartment. This is inconsistent with findings in earlier studies [5, 6] that compared compartmental disease models for animal populations, but is consistent with current belief that infection is lifelong.

6 References

[1] Jr Sullivan, William J. and Victoria Jeffers. Mechanisms of *Toxoplasma gondii* persistence and latency. *FEMS Microbiology Reviews*, 36(3):717–733, 05 2012.

[2] C. G. K. Lüder and Uwe Groß. Apoptosis and its modulation during infection with *Toxoplasma gondii*: molecular mechanisms and role in pathogenesis. *Current topics in microbiology and immunology*, 289:219–37, 2005.

[3] Tatiane S. Lima and Melissa B. Lodoen. Mechanisms of human innate immune evasion by *Toxoplasma gondii*. *Frontiers in Cellular and Infection Microbiology*, 9, 2019.

[4] Solène Rougier, Jose G Montoya, and François Peyron. Lifelong persistence of *Toxoplasma* cysts: A questionable dogma? *Trends in Parasitology*, 33(2):93–101, 02 2017.

[5] Filip Dámek, Arno Swart, Helga Waap, Pikka Jokelainen, Delphine Le Roux, Gunita Deksne, Huifang Deng, Gereon Schares, Anna Lundén, Gema Álvarez García, Martha Betson, Rebecca K. Davidson, Adriana Györke, Daniela Antolová, Zuzana Hurníková, Henk J. Wisselink, Jacek Sroka, Joke W. B. van der Giessen, Radu Blaga, and Marieke Opsteegh. *Systematic Review and Modelling of Age-Dependent Prevalence of Toxoplasma gondii in Livestock, Wildlife and Felids in Europe*, 2023.

[6] Marieke Opsteegh, Arno Swart, Manoj Fonville, Leo Dekkers, and Joke van der Giessen. Age-related *Toxoplasma gondii* seroprevalence in dutch wild boar inconsistent with lifelong persistence of antibodies. *PLOS ONE*, 6(1):1–6, 01 2011.

[7] J.P. Dubey. *Toxoplasma Gondii*. University of Texas Medical Branch at Galveston, Galveston, TX, 4 edition, 1996.

[8] Louis M. Weiss and Jitender. P. Dubey. Toxoplasmosis: A history of clinical observations. *International Journal for Parasitology*, 39(8):895–901, 2009. Toxoplasma Centennial Issue.

[9] Justus G. Garweg, François Kieffer, Laurent Mandelbrot, François Peyron, and Martine Wallon. Long-term outcomes in children with congenital toxoplasmosis; a systematic review. *Pathogens*, 11(10), 2022.

[10] Sin-Yew Wong and Jack S. Remington. Toxoplasmosis in pregnancy. *Clinical Infectious Diseases*, 18(6):853–861, 1994.

[11] Sang-Bok Lee and Tae-Gyu Lee. Toxoplasmic encephalitis in patient with acquired immunodeficiency syndrome. *Brain Tumor Research and Treatment*, 5:34, 04 2017.

[12] D. Hill and J.P. Dubey. *Toxoplasma gondii*: transmission, diagnosis and prevention. *Clinical Microbiology and Infection*, 8(10):634–640, 2002.

[13] Oda E. van den Berg, Kamelia R. Stanoeva, Rens Zonneveld, Denise Hoek-van Deursen, Fiona R. van der Klis, Jan van de Kassteele, Eelco Franz, Marieke Opsteegh, Ingrid H. M. Friesema, and Laetitia M. Kortbeek. Seroprevalence of *Toxoplasma gondii* and associated risk factors for infection in the netherlands: third cross-sectional national study. *Epidemiology & Infection*, 151:e136, 2023.

[14] J. K. Frenkel, J. P. Dubey, and N. L. Miller. *Toxoplasma gondii* in cats: Fecal stages identified as coccidian oocysts. *Science*, 167(3919):893–896, 1970.

[15] LM Kortbeek, HE De Melker, IK Veldhuijzen, and MAE Conyn-Van Spaendonck. Population-based *Toxoplasma* seroprevalence study in the netherlands. *Epidemiology & Infection*, 132(5):839–845, 2004.

[16] Georg Kapperud, Pal A. Jenum, Babill Stray-Pedersen, Kjetil K. Melby, Anne Eskild, and Jan Eng. Risk Factors for *Toxoplasma gondii* Infection in Pregnancy: Results of a Prospective Case-Control Study in Norway. *American Journal of Epidemiology*, 144(4):405–412, 08 1996.

[17] Astrid M Tenter, Anja R Heckeroth, and Louis M Weiss. *Toxoplasma gondii*: from animals to humans. *International Journal for Parasitology*, 30(12):1217–1258, 2000. Thematic Issue: Emerging Parasite Zoonoses.

[18] J P Dubey. Sources of *Toxoplasma gondii* infection in pregnancy. *BMJ*, 321(7254):127–128, 2000.

[19] A J C Cook, Richard Holliman, R E Gilbert, W Buffolano, J Zufferey, E Petersen, P A Jenum, W Foulon, A E Semprini, and D T Dunn. Sources of *Toxoplasma* infection in pregnant women: European multicentre case-control studycommentary: Congenital toxoplasmosis—further thought for food. *BMJ*, 321(7254):142–147, 2000.

[20] Martijn Bouwknegt, Brecht Devleesschauwer, Heather Graham, Lucy J Robertson, Joke WB van der Giessen, and the Euro-FBP workshop participants. Prioritisation of food-borne parasites in europe, 2016. *Eurosurveillance*, 23(9), 2018.

[21] Florence Robert-Gangneux and Marie-Laure Dardé. Epidemiology of and diagnostic strategies for toxoplasmosis. *Clinical microbiology reviews*, 25(2):264–296, 2012.

[22] A. Hofhuis, W. van Pelt, Y. T. H. P. van Duynhoven, C. D. M. Nijhuis, L. Molema, F. R. M. van der Klis, A. H. Havelaar, and L. M. Kortbeek. Decreased prevalence and age-specific risk factors for *Toxoplasma gondii* igg antibodies in the netherlands between 1995/1996 and 2006/2007. *Epidemiology & Infection*, 139(4):530–538, 2011.

[23] Marieke Opsteegh, Peter Teunis, Marieke Mensink, Lothar Züchner, Adriana Titilincu, Merel Langelaar, and Joke van der Giessen. Evaluation of ELISA test characteristics and estimation of *Toxoplasma gondii* seroprevalence in dutch sheep using mixture models. *Preventive Veterinary Medicine*, 96(3):232–240, 2010.

[24] Judith A. Bouman, Julien Riou, Sebastian Bonhoeffer, and Roland R. Regoes. Estimating the cumulative incidence of SARS-CoV-2 with imperfect serological tests: Exploiting cutoff-free approaches. *PLOS Computational Biology*, 17(2):1–19, 02 2021.

[25] Arno Swart, Miriam Maas, Ankje de Vries, Tryntsje Cuperus, and Marieke Opsteegh. Bayesian binary mixture models as a flexible alternative to cut-off analysis of ELISA results, a case study of seoul orthohantavirus. *Viruses*, 13(6), 2021.

[26] M. Iannelli. *Mathematical Theory of Age-structured Population Dynamics*. Applied mathematics monographs. Giardini editori e stampatori, 1995.

[27] Herbert W. Hethcote. The mathematics of infectious diseases. *SIAM Review*, 42(4):599–653, 2000.

[28] Rens Schoot and Sarah Depaoli. Bayesian analyses: Where to start and what to report. *European Health Psychologist*, 16:75–84, 01 2014.

[29] Dominique Makowski, Mattan S. Ben-Shachar, S. H. Annabel Chen, and Daniel Lüdecke. Indices of effect existence and significance in the bayesian framework. *Frontiers in Psychology*, 10, 2019.

[30] John K. Kruschke. Rejecting or accepting parameter values in bayesian estimation. *Advances in Methods and Practices in Psychological Science*, 1(2):270–280, 2018.

[31] Richard McElreath. *Statistical rethinking: a Bayesian course with examples in R and Stan*. Number 122 in Chapman & Hall/CRC texts in statistical science series. CRC Press/Taylor & Francis Group, Boca Raton, 2016. largely / videos.

[32] J.K. Kruschke. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Academic Press. Academic Press, 2015.

[33] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences (2nd ed.)*. L. Erlbaum Associates, 1988.

[34] Dominique Makowski, Mattan S. Ben-Shachar, and Daniel Lüdecke. bayestestr: Describing effects and their uncertainty, existence and significance within the bayesian framework. *Journal of Open Source Software*, 4(40):1541, 2019.

[35] Michael Thrusfield. *Veterinary epidemiology*. John Wiley & Sons, 2018.

[36] R H Jacobson. Validation of serological assays for diagnosis of infectious diseases. *Rev. Sci. Tech.*, 17(2):469–526, 08 1998.

[37] William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

[38] A. G. M'Kendrick. Applications of mathematics to medical problems. *Proceedings of the Edinburgh Mathematical Society*, 44:98–130, 1925.

[39] James D. Murray. *Mathematical Biology I. An Introduction*, volume 17 of *Interdisciplinary Applied Mathematics*. Springer, New York, 3 edition, 2002.

[40] Stan Development Team. Stan Modeling Language User's Guide and Reference Manual, Version 2.33, 2023.

[41] Stan Development Team. RStan: the R interface to Stan, 2023. R package version 2.32.3.

[42] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023.

[43] Léo Grinsztajn, Elizaveta Semenova, Charles C. Margossian, and Julien Riou. Bayesian workflow for disease transmission modeling in stan. *Statistics in Medicine*, 40(27):6209–6234, 2021.

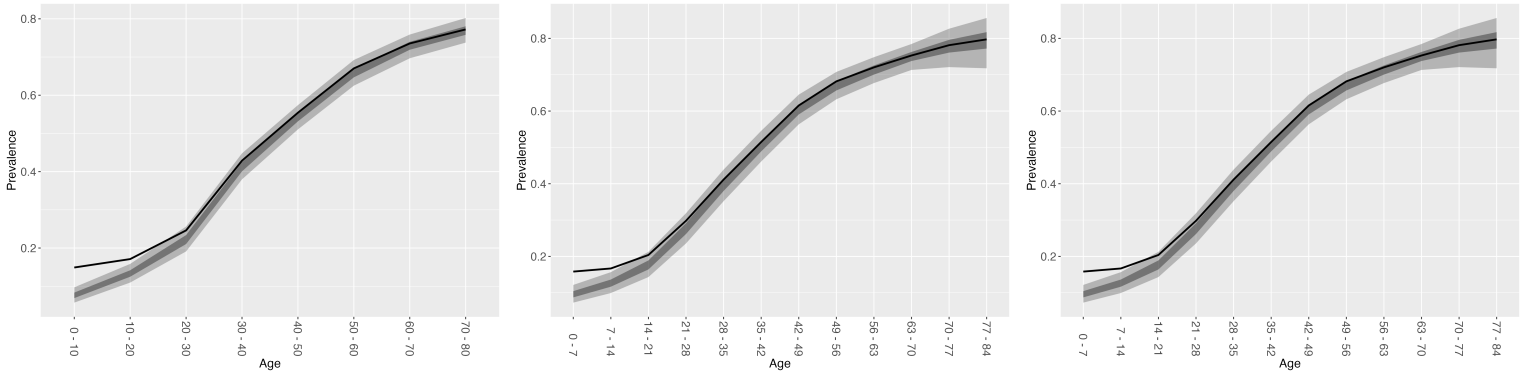
[44] Zhijin Wu, Xue Li, Robert Husnay, Vasu Chakravarthy, Bin Wang, and Zhiqiang Wu. A novel highly accurate log skew normal approximation method to lognormal sum distributions. In *2009 IEEE Wireless Communications and Networking Conference*, pages 1–6, 2009.

[45] Marwane Hcine and Ridha Bouallegue. On the approximation of the sum of lognormals by a log skew normal distribution. *International journal of Computer Networks & Communications*, 7:135–151, 01 2015.

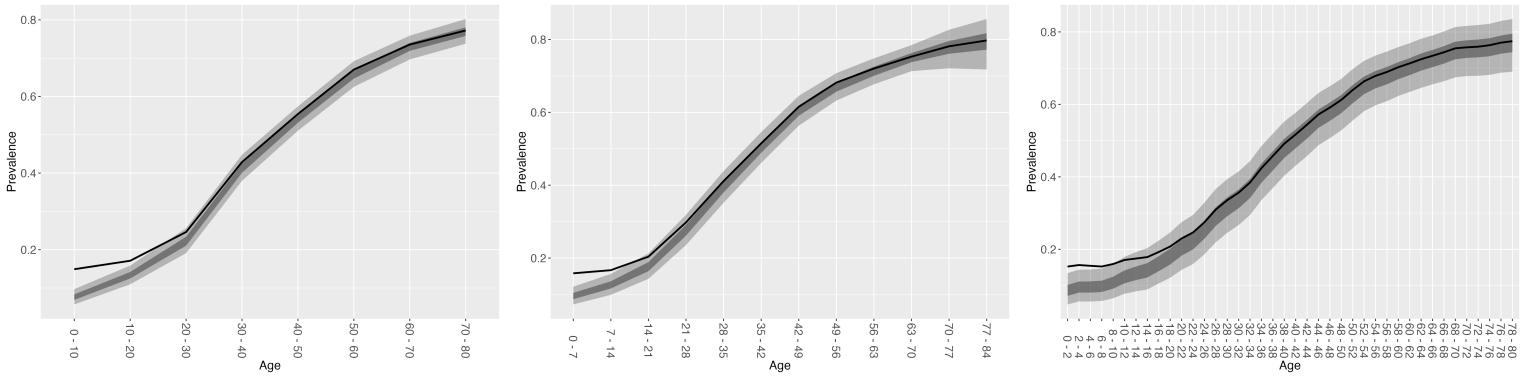
[46] F Gómez-Chávez, I Cañedo-Solares, LB Ortiz-Alegría, Y Flores-García, H Luna-Pastén, R Figueroa-Damián, JC Mora-González, and D Correa. Maternal immune response during pregnancy and vertical transmission in human toxoplasmosis. *Frontiers in Immunology*, 10:285, 2019.

[47] Herbert W. Hethcote. The mathematics of infectious diseases. *SIAM Review*, 42(4):599–653, 2000.

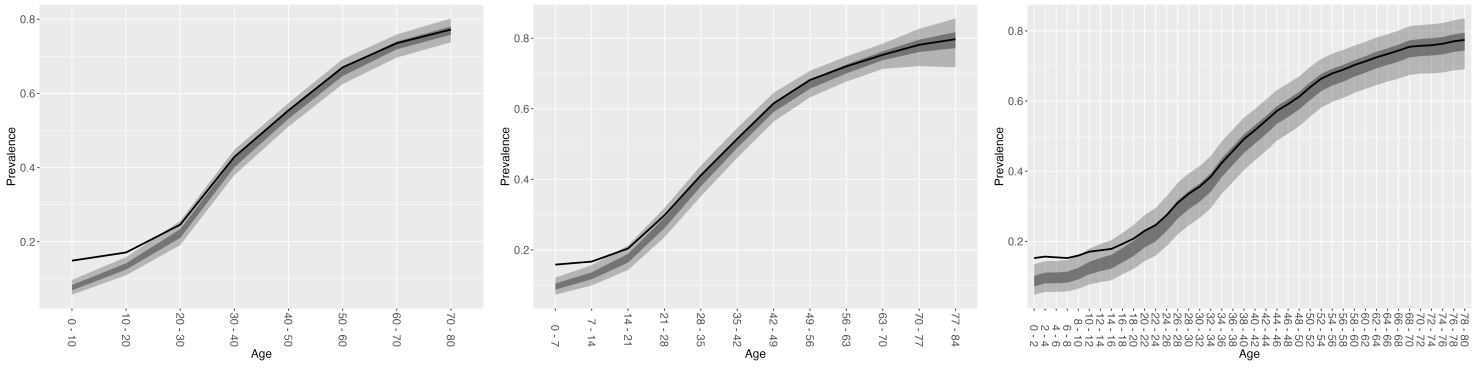
A Comparison of age categories



(a) PIENTER 1



(b) PIENTER 2



(c) PIENTER 3

Figure 18: For each PIENTER, the first figure (left) shows the prevalence estimated using 10 year intervals, the second (middle) using 7 year intervals where the estimates are smoothed with a running average of width 2 and the third (right) using 2 year intervals, smoothed with a running average of width 10. The 95% and 50% quantile intervals are shown as dark grey and light grey areas respectively and the black lines represent the prevalence from the cut-off method.

Increasing the width of the age-categories decreases the uncertainty of the estimates, since the sample size increases. Smaller width yields a more detailed image of the shape, at the expense of the certainty. Ideally, a width is chosen such that a detailed analysis can be made, while a minimum amount of certainty is lost. Smaller widths are more sensitive to outliers in the data, so that smoothing may be applied to reduce noise.

B Covariate analysis results

In each table, p_diff, p_lower and p_upper denote the median, lower bound and upper bound respectively of the 89% HDI, pd is the probability of direction of the distribution of covar1 - covar2. The result of the HDI with the ROPE interval $[-0.05, 0.05]$ decision rule is given in the result column.

B.1 Covariate analysis tables PIENTER 1

Table 8: HDI + Rope analysis for P1: "What is the highest level of education or training that you have completed?".

education1	education2	age_cat	p_diff	lower	upper	pd	result
Low	High	0 - 10	-0.02	-0.05	0.02	0.81	Undecided
Middle	High	0 - 10	0.00	-0.04	0.05	0.55	Undecided
Middle	Low	0 - 10	0.02	-0.02	0.07	0.79	Undecided
Low	High	10 - 20	0.03	-0.02	0.07	0.86	Undecided
Middle	High	10 - 20	0.02	-0.04	0.08	0.71	Undecided
Middle	Low	10 - 20	-0.01	-0.07	0.04	0.61	Undecided
Low	High	20 - 30	0.12	0.05	0.18	1.00	Low higher
Middle	High	20 - 30	0.08	0.01	0.17	0.95	Undecided
Middle	Low	20 - 30	-0.03	-0.11	0.04	0.75	Undecided
Low	High	30 - 40	0.03	-0.03	0.10	0.78	Undecided
Middle	High	30 - 40	0.01	-0.09	0.09	0.55	Undecided
Middle	Low	30 - 40	-0.03	-0.10	0.06	0.69	Undecided
Low	High	40 - 50	0.02	-0.05	0.08	0.68	Undecided
Middle	High	40 - 50	-0.06	-0.18	0.06	0.79	Undecided
Middle	Low	40 - 50	-0.08	-0.20	0.03	0.87	Undecided
Low	High	50 - 60	0.09	0.02	0.16	0.98	Undecided
Middle	High	50 - 60	0.11	-0.01	0.23	0.91	Undecided
Middle	Low	50 - 60	0.02	-0.09	0.13	0.61	Undecided
Low	High	60 - 70	0.14	0.06	0.22	1.00	Low higher
Middle	High	60 - 70	0.09	-0.04	0.22	0.87	Undecided
Middle	Low	60 - 70	-0.05	-0.15	0.06	0.75	Undecided
Low	High	70 - 80	0.20	0.09	0.30	1.00	Low higher
Middle	High	70 - 80	0.09	-0.08	0.23	0.81	Undecided
Middle	Low	70 - 80	-0.11	-0.24	0.01	0.94	Undecided

Table 9: HDI + Rope analysis for P1: "Have you kept farm animals in the past 5 years?".

agriculture1	agriculture2	age_cat	p_diff	lower	upper	pd	result
No	Yes	0 - 10	-0.00	-0.05	0.04	0.56	No difference
No	Yes	10 - 20	-0.06	-0.12	-0.00	0.95	Undecided
No	Yes	20 - 30	-0.17	-0.27	-0.06	1.00	Yes higher
No	Yes	30 - 40	-0.06	-0.15	0.03	0.85	Undecided
No	Yes	40 - 50	0.06	-0.03	0.14	0.87	Undecided
No	Yes	50 - 60	-0.04	-0.13	0.03	0.81	Undecided
No	Yes	60 - 70	-0.11	-0.19	-0.02	0.95	Undecided
No	Yes	70 - 80	-0.07	-0.15	0.04	0.84	Undecided

Table 10: HDI + Rope analysis for P1: "Have you worked with your bare hands in the soil in the garden/on land in the past 12 months?".

gardening1	gardening2	age_cat	p_diff	lower	upper	pd	result
No	Yes	0 - 10	-0.00	-0.03	0.03	0.53	No difference
No	Yes	10 - 20	-0.02	-0.06	0.02	0.75	Undecided
No	Yes	20 - 30	0.01	-0.05	0.06	0.61	Undecided
No	Yes	30 - 40	0.07	-0.01	0.14	0.92	Undecided
No	Yes	40 - 50	0.10	0.03	0.17	0.99	Undecided
No	Yes	50 - 60	-0.07	-0.13	-0.01	0.97	Undecided
No	Yes	60 - 70	-0.05	-0.11	0.01	0.93	Undecided
No	Yes	70 - 80	0.00	-0.05	0.05	0.53	Undecided

Table 11: HDI + Rope analysis for P1: "Have you kept a cat as a pet for the past 5 years?"

petcat1	petcat2	age_cat	p_diff	lower	upper	pd	result
No	Yes	0 - 10	0.02	-0.01	0.05	0.88	Undecided
No	Yes	10 - 20	-0.00	-0.04	0.04	0.53	No difference
No	Yes	20 - 30	-0.03	-0.08	0.03	0.79	Undecided
No	Yes	30 - 40	-0.05	-0.12	0.00	0.93	Undecided
No	Yes	40 - 50	-0.09	-0.15	-0.03	0.99	Undecided
No	Yes	50 - 60	-0.12	-0.17	-0.06	1.00	Yes higher
No	Yes	60 - 70	-0.11	-0.17	-0.05	1.00	Yes higher
No	Yes	70 - 80	-0.01	-0.09	0.08	0.57	Undecided

B.2 Covariate analysis tables PIENTER 2

Table 12: HDI + Rope analysis for P2: "What is the highest level of education or training that you have completed?"

education1	education2	age_cat	p_diff	lower	upper	pd	result
Low	High	0 - 10	0.03	-0.02	0.07	0.83	Undecided
Middle	High	0 - 10	-0.02	-0.06	0.02	0.77	Undecided
Middle	Low	0 - 10	-0.04	-0.09	-0.01	0.97	Undecided
Low	High	10 - 20	0.01	-0.04	0.07	0.66	Undecided
Middle	High	10 - 20	0.03	-0.04	0.10	0.76	Undecided
Middle	Low	10 - 20	0.01	-0.04	0.06	0.68	Undecided
Low	High	20 - 30	0.03	-0.04	0.09	0.79	Undecided
Middle	High	20 - 30	0.03	-0.02	0.08	0.84	Undecided
Middle	Low	20 - 30	-0.00	-0.06	0.06	0.51	Undecided
Low	High	30 - 40	0.11	0.04	0.18	1.00	Undecided
Middle	High	30 - 40	0.02	-0.04	0.08	0.69	Undecided
Middle	Low	30 - 40	-0.09	-0.16	-0.03	0.99	Undecided
Low	High	40 - 50	0.06	-0.02	0.13	0.89	Undecided
Middle	High	40 - 50	0.05	-0.03	0.12	0.84	Undecided
Middle	Low	40 - 50	-0.01	-0.09	0.06	0.56	Undecided
Low	High	50 - 60	0.01	-0.06	0.09	0.61	Undecided
Middle	High	50 - 60	-0.10	-0.19	-0.01	0.97	Undecided
Middle	Low	50 - 60	-0.11	-0.19	-0.03	0.99	Undecided
Low	High	60 - 70	0.10	0.02	0.18	0.97	Undecided
Middle	High	60 - 70	0.08	-0.01	0.17	0.90	Undecided
Middle	Low	60 - 70	-0.02	-0.09	0.06	0.67	Undecided
Low	High	70 - 80	0.03	-0.06	0.13	0.72	Undecided
Middle	High	70 - 80	0.07	-0.05	0.18	0.84	Undecided
Middle	Low	70 - 80	0.04	-0.04	0.12	0.76	Undecided

Table 13: HDI + Rope analysis for P2: "Have you worked with your bare hands in the soil in the garden/on land in the past 12 months?"

gardening1	gardening2	age_cat	p_diff	lower	upper	pd	result
No	Yes	0 - 10	0.02	-0.01	0.06	0.85	Undecided
No	Yes	10 - 20	-0.03	-0.08	0.00	0.90	Undecided
No	Yes	20 - 30	0.00	-0.05	0.04	0.50	No difference
No	Yes	30 - 40	-0.03	-0.09	0.04	0.74	Undecided
No	Yes	40 - 50	0.01	-0.07	0.08	0.58	Undecided
No	Yes	50 - 60	-0.04	-0.10	0.03	0.81	Undecided
No	Yes	60 - 70	0.03	-0.03	0.09	0.80	Undecided
No	Yes	70 - 80	-0.02	-0.08	0.04	0.68	Undecided

Table 14: HDI + Rope analysis for P2: "Have you kept farm animals in the past 5 years?"

agriculture1	agriculture2	age_cat	p_diff	lower	upper	pd	result
No	Yes	0 - 10	-0.01	-0.07	0.05	0.57	Undecided
No	Yes	10 - 20	-0.06	-0.14	0.01	0.92	Undecided
No	Yes	20 - 30	-0.06	-0.13	0.02	0.90	Undecided
No	Yes	30 - 40	0.02	-0.06	0.12	0.67	Undecided
No	Yes	40 - 50	-0.02	-0.12	0.09	0.63	Undecided
No	Yes	50 - 60	-0.01	-0.12	0.10	0.54	Undecided
No	Yes	60 - 70	-0.09	-0.19	0.02	0.90	Undecided
No	Yes	70 - 80	0.01	-0.13	0.15	0.54	Undecided

Table 15: PIENTER 2

Table 16: HDI + Rope analysis for P2: "Have you kept a cat as a pet for the past 5 years?"

petcat1	petcat2	age_cat	p_diff	lower	upper	pd	result
No	Yes	0 - 10	0.01	-0.03	0.05	0.67	Undecided
No	Yes	10 - 20	-0.02	-0.07	0.03	0.76	Undecided
No	Yes	20 - 30	-0.02	-0.07	0.03	0.77	Undecided
No	Yes	30 - 40	-0.03	-0.09	0.03	0.77	Undecided
No	Yes	40 - 50	-0.10	-0.17	-0.04	0.99	Undecided
No	Yes	50 - 60	-0.05	-0.12	0.02	0.86	Undecided
No	Yes	60 - 70	-0.07	-0.14	0.01	0.91	Undecided
No	Yes	70 - 80	-0.20	-0.28	-0.12	1.00	Yes higher

Table 17: HDI + Rope analysis for P2: "Did you have contact with cats in the past 12 months?"

contcat1	contcat2	age_cat	p_diff	lower	upper	pd	result
No	Yes	0 - 10	-0.00	-0.03	0.03	0.57	No difference
No	Yes	10 - 20	0.03	-0.01	0.08	0.87	Undecided
No	Yes	20 - 30	0.02	-0.03	0.07	0.78	Undecided
No	Yes	30 - 40	0.03	-0.03	0.09	0.80	Undecided
No	Yes	40 - 50	-0.10	-0.16	-0.04	0.99	Undecided
No	Yes	50 - 60	-0.08	-0.14	-0.02	0.98	Undecided
No	Yes	60 - 70	-0.08	-0.14	-0.02	0.99	Undecided
No	Yes	70 - 80	-0.15	-0.22	-0.09	1.00	Yes higher

Table 18: HDI + Rope analysis for P2: "Have you consumed raw meat products in the past 12 months?"

rawmeat1	rawmeat2	age_cat	p_diff	lower	upper	pd	result
No	Yes	0 - 10	0.12	0.02	0.23	0.98	Undecided
No	Yes	10 - 20	-0.01	-0.07	0.05	0.59	Undecided
No	Yes	20 - 30	0.01	-0.04	0.07	0.62	Undecided
No	Yes	30 - 40	0.01	-0.06	0.07	0.59	Undecided
No	Yes	40 - 50	-0.06	-0.12	0.02	0.90	Undecided
No	Yes	50 - 60	-0.08	-0.15	-0.02	0.98	Undecided
No	Yes	60 - 70	-0.12	-0.18	-0.06	1.00	Yes higher
No	Yes	70 - 80	-0.05	-0.12	0.02	0.86	Undecided

Table 19: HDI + Rope analysis for P2: "Have you consumed raw pork meat in the past 12 months?"

rawporkmeat1	rawporkmeat2	age_cat	p_diff	lower	upper	pd	result
no	yes	0 - 10	0.02	-0.07	0.09	0.64	Undecided
no	yes	10 - 20	-0.01	-0.08	0.05	0.59	Undecided
no	yes	20 - 30	-0.02	-0.07	0.03	0.74	Undecided
no	yes	30 - 40	0.03	-0.03	0.09	0.79	Undecided
no	yes	40 - 50	-0.09	-0.16	-0.02	0.98	Undecided
no	yes	50 - 60	-0.12	-0.19	-0.05	1.00	Yes higher
no	yes	60 - 70	-0.15	-0.22	-0.08	1.00	Yes higher
no	yes	70 - 80	-0.06	-0.15	0.05	0.81	Undecided

Table 20: PIENTER 2 HDI + ROPE for eating raw porkmeat per 10 years

Table 21: HDI + Rope analysis for P2: "Have you consumed raw beef in the past 12 months?"

rawbeef1	rawbeef2	age_cat	p_diff	lower	upper	pd	result
No	Yes	0 - 10	0.01	-0.03	0.05	0.67	Undecided
No	Yes	10 - 20	-0.02	-0.07	0.03	0.76	Undecided
No	Yes	20 - 30	-0.02	-0.07	0.03	0.77	Undecided
No	Yes	30 - 40	-0.03	-0.09	0.03	0.77	Undecided
No	Yes	40 - 50	-0.10	-0.17	-0.04	0.99	Undecided
No	Yes	50 - 60	-0.05	-0.12	0.02	0.86	Undecided
No	Yes	60 - 70	-0.07	-0.14	0.01	0.91	Undecided
No	Yes	70 - 80	-0.20	-0.28	-0.12	1.00	Yes higher

Table 22: HDI + Rope analysis for P2: "Do you consume unwashed raw vegetables?"

unwashedveget1	unwashedveget2	age_cat	p_diff	lower	upper	pd	result
No	Yes	0 - 10	-0.03	-0.07	0.01	0.87	Undecided
No	Yes	10 - 20	-0.04	-0.09	0.01	0.90	Undecided
No	Yes	20 - 30	-0.02	-0.06	0.02	0.74	Undecided
No	Yes	30 - 40	0.02	-0.04	0.07	0.69	Undecided
No	Yes	40 - 50	-0.03	-0.10	0.03	0.79	Undecided
No	Yes	50 - 60	-0.04	-0.10	0.03	0.81	Undecided
No	Yes	60 - 70	-0.06	-0.13	0.01	0.90	Undecided
No	Yes	70 - 80	-0.13	-0.21	-0.05	0.99	Undecided

B.3 Covariate analysis tables PIENTER 3

Table 23: HDI + Rope analysis for P3: "What is the highest level of education or training that you have completed?"

education1	education2	age_cat	p_diff	lower	upper	pd	result
Low	High	0 - 10	0.05	0.01	0.09	0.99	Undecided
Middle	High	0 - 10	0.04	0.01	0.07	0.99	Undecided
Middle	Low	0 - 10	-0.01	-0.06	0.03	0.61	Undecided
Low	High	10 - 20	-0.00	-0.04	0.04	0.52	No difference
Middle	High	10 - 20	0.01	-0.04	0.05	0.59	No difference
Middle	Low	10 - 20	0.01	-0.03	0.04	0.62	No difference
Low	High	20 - 30	0.05	-0.01	0.11	0.92	Undecided
Middle	High	20 - 30	0.02	-0.02	0.05	0.79	No difference
Middle	Low	20 - 30	-0.03	-0.09	0.03	0.82	Undecided
Low	High	30 - 40	0.07	0.00	0.13	0.96	Undecided
Middle	High	30 - 40	0.02	-0.02	0.06	0.77	Undecided
Middle	Low	30 - 40	-0.05	-0.11	0.02	0.87	Undecided
Low	High	40 - 50	0.05	-0.02	0.12	0.90	Undecided
Middle	High	40 - 50	-0.04	-0.09	0.02	0.85	Undecided
Middle	Low	40 - 50	-0.09	-0.15	-0.02	0.99	Undecided
Low	High	50 - 60	0.10	0.03	0.17	0.98	Undecided
Middle	High	50 - 60	-0.00	-0.07	0.08	0.53	Undecided
Middle	Low	50 - 60	-0.10	-0.17	-0.03	0.99	Undecided
Low	High	60 - 70	0.08	0.01	0.15	0.97	Undecided
Middle	High	60 - 70	-0.05	-0.13	0.03	0.83	Undecided
Middle	Low	60 - 70	-0.13	-0.20	-0.06	1.00	Low higher
Low	High	70 - 80	0.06	-0.02	0.14	0.87	Undecided
Middle	High	70 - 80	-0.03	-0.13	0.08	0.68	Undecided
Middle	Low	70 - 80	-0.09	-0.18	0.00	0.94	Undecided
Low	High	80 - 90	0.10	-0.07	0.27	0.84	Undecided
Middle	High	80 - 90	-0.00	-0.24	0.21	0.50	Undecided
Middle	Low	80 - 90	-0.10	-0.28	0.08	0.83	Undecided

Table 24: HDI + Rope analysis for P3: "Have you worked with your bare hands in the soil in the garden/on land in the past 12 months?"

gardening1	gardening2	age_cat	p_diff	lower	upper	pd	result
No	Yes	0 - 10	-0.01	-0.07	0.04	0.59	Undecided
No	Yes	10 - 20	-0.01	-0.04	0.02	0.77	No difference
No	Yes	20 - 30	-0.01	-0.05	0.04	0.60	No difference
No	Yes	30 - 40	-0.00	-0.06	0.05	0.55	Undecided
No	Yes	40 - 50	0.00	-0.06	0.07	0.54	Undecided
No	Yes	50 - 60	0.02	-0.04	0.08	0.72	Undecided
No	Yes	60 - 70	0.00	-0.06	0.07	0.53	Undecided
No	Yes	70 - 80	-0.04	-0.18	0.09	0.70	Undecided

Table 25: HDI + Rope analysis for P3: "Have you kept farm animals in the past 5 years?"

agriculture1	agriculture2	age_cat	p_diff	lower	upper	pd	result
No	Yes	0 - 10	0.01	-0.03	0.04	0.71	No difference
No	Yes	10 - 20	-0.02	-0.08	0.04	0.69	Undecided
No	Yes	20 - 30	-0.05	-0.12	0.01	0.92	Undecided
No	Yes	30 - 40	-0.00	-0.09	0.08	0.54	Undecided
No	Yes	40 - 50	0.02	-0.07	0.10	0.61	Undecided
No	Yes	50 - 60	-0.01	-0.13	0.11	0.53	Undecided
No	Yes	60 - 70	0.04	-0.08	0.17	0.70	Undecided
No	Yes	70 - 80	-0.06	-0.20	0.10	0.71	Undecided
No	Yes	80 - 90	-0.01	-0.23	0.24	0.53	Undecided

Table 26: HDI + Rope analysis for P3: "Have you kept a cat as a pet for the past 5 years?"

petcat1	petcat2	age_cat	p_diff	lower	upper	pd	result
No	Yes	0 - 10	0.01	-0.01	0.04	0.78	No difference
No	Yes	10 - 20	-0.02	-0.06	0.02	0.83	Undecided
No	Yes	20 - 30	0.01	-0.02	0.04	0.74	No difference
No	Yes	30 - 40	0.04	-0.00	0.08	0.92	Undecided
No	Yes	40 - 50	0.02	-0.03	0.08	0.73	Undecided
No	Yes	50 - 60	-0.05	-0.12	0.01	0.90	Undecided
No	Yes	60 - 70	-0.00	-0.09	0.07	0.51	Undecided
No	Yes	70 - 80	-0.01	-0.12	0.10	0.57	Undecided
No	Yes	80 - 90	0.31	-0.03	0.62	0.92	Undecided

Table 27: HDI + Rope analysis for P3: "Did you have contact with cats in the past 12 months?"

contcat1	contcat2	age_cat	p_diff	lower	upper	pd	result
yes	no	0 - 10	-0.02	-0.04	0.01	0.83	No difference
yes	no	10 - 20	0.02	-0.01	0.05	0.81	No difference
yes	no	20 - 30	-0.03	-0.06	0.00	0.92	Undecided
yes	no	30 - 40	-0.02	-0.06	0.02	0.83	Undecided
yes	no	40 - 50	-0.04	-0.09	0.01	0.90	Undecided
yes	no	50 - 60	0.06	-0.00	0.12	0.94	Undecided
yes	no	60 - 70	0.03	-0.04	0.09	0.76	Undecided
yes	no	70 - 80	0.01	-0.06	0.08	0.61	Undecided
yes	no	80 - 90	-0.19	-0.36	-0.02	0.97	Undecided

Table 28: HDI + Rope analysis for P3: "Have you consumed raw meat products in the past 12 months?"

rawmeat1	rawmeat2	age_cat	p_diff	lower	upper	pd	result
yes	no	0 - 10	0.02	-0.00	0.06	0.92	Undecided
yes	no	10 - 20	0.00	-0.03	0.04	0.54	No difference
yes	no	20 - 30	0.03	0.00	0.06	0.94	Undecided
yes	no	30 - 40	-0.03	-0.07	0.02	0.82	Undecided
yes	no	40 - 50	0.03	-0.03	0.08	0.77	Undecided
yes	no	50 - 60	0.02	-0.05	0.09	0.64	Undecided
yes	no	60 - 70	0.09	0.03	0.15	0.99	Undecided
yes	no	70 - 80	-0.01	-0.08	0.07	0.57	Undecided
yes	no	80 - 90	0.06	-0.07	0.22	0.76	Undecided

Table 29: HDI + Rope analysis for P3: "Have you consumed raw pork meat in the past 12 months?"

rawporkmeat1	rawporkmeat2	age_cat	p_diff	lower	upper	pd	result
yes	no	0 - 10	0.03	-0.02	0.10	0.89	Undecided
yes	no	10 - 20	-0.01	-0.05	0.04	0.58	No difference
yes	no	20 - 30	0.02	-0.01	0.05	0.82	Undecided
yes	no	30 - 40	-0.01	-0.05	0.04	0.63	Undecided
yes	no	40 - 50	0.04	-0.02	0.10	0.83	Undecided
yes	no	50 - 60	0.12	0.03	0.19	0.99	Undecided
yes	no	60 - 70	0.12	0.04	0.19	0.99	Undecided
yes	no	70 - 80	-0.06	-0.16	0.05	0.82	Undecided
yes	no	80 - 90	0.03	-0.16	0.19	0.61	Undecided

Table 30: HDI + Rope analysis for P3: "Have you consumed raw beef in the past 12 months?"

rawbeef1	rawbeef2	age_cat	p_diff	lower	upper	pd	result
yes	no	0 - 10	0.02	-0.01	0.06	0.89	Undecided
yes	no	10 - 20	0.01	-0.02	0.05	0.76	No difference
yes	no	20 - 30	-0.01	-0.04	0.03	0.65	No difference
yes	no	30 - 40	-0.06	-0.11	-0.01	0.98	Undecided
yes	no	40 - 50	-0.03	-0.08	0.03	0.80	Undecided
yes	no	50 - 60	-0.01	-0.07	0.06	0.56	Undecided
yes	no	60 - 70	0.07	0.01	0.13	0.97	Undecided
yes	no	70 - 80	0.00	-0.07	0.07	0.50	Undecided
yes	no	80 - 90	0.05	-0.09	0.20	0.73	Undecided

Table 31: HDI + Rope analysis for P3: "Do you consume unwashed raw vegetables?"

unwashedveget1	unwashedveget2	age_cat	p_diff	lower	upper	pd	result
yes	no	0 - 10	0.01	-0.02	0.04	0.70	No difference
yes	no	10 - 20	0.03	0.00	0.06	0.95	Undecided
yes	no	20 - 30	-0.03	-0.06	0.01	0.88	Undecided
yes	no	30 - 40	-0.02	-0.06	0.02	0.81	Undecided
yes	no	40 - 50	0.01	-0.04	0.06	0.63	Undecided
yes	no	50 - 60	-0.05	-0.11	0.02	0.91	Undecided
yes	no	60 - 70	0.03	-0.03	0.09	0.78	Undecided
yes	no	70 - 80	0.02	-0.04	0.09	0.67	Undecided
yes	no	80 - 90	-0.14	-0.28	0.00	0.94	Undecided