Compound processes as models for clumped parasite data

Heinzmann, D; Barbour, A D; Torgerson, P R

Postprint available at:
http://www.zora.uzh.ch

Posted at the Zurich Open Repository and Archive, University of Zurich.
http://www.zora.uzh.ch

Originally published at:
Compound processes as models for clumped parasite data

Abstract

Compound processes are proposed as models for the acquisition of hydatid cysts in sheep, caused by the parasite Echinococcus granulosus. The hypothesis of a clumped infection process against single ingestions is tested and it is shown that the clump-based approach provides a more accurate description of the two data sets investigated. Models with simple and mixed Poisson incidence processes and different clump size distributions are compared. A mixed Poisson incidence process with a zero-truncated negative binomial distribution for the clump sizes is shown to give an adequate description, suggesting that the acquisition of hydatid cysts in the sheep population is heterogeneous, and that the clump sizes are aggregated. The estimates of the parameters derived from the data take plausible values. The average infection rate and the clump size distribution are comparable in both data sets. Goodness-of-fit measures indicate that the model fits the data reasonably well.
Compound processes as models for clumped parasite data

Dominik Heinzmann\textsuperscript{1,2}, A.D. Barbour\textsuperscript{1}, and Paul R. Torgerson\textsuperscript{2,3}

\textsuperscript{1}Institute of Mathematics, University of Zurich, Switzerland
\textsuperscript{2}Institute of Parasitology, University of Zurich, Switzerland
\textsuperscript{3}School of Veterinary Medicine, Ross University, West Indies

Abstract

Compound processes are proposed as models for the acquisition of hydatid cysts in sheep, caused by the parasite \textit{Echinococcus granulosus}. The hypothesis of a clumped infection process against single ingestions is tested and it is shown that the clump-based approach provides a more accurate description of the two data sets investigated. Models with simple and mixed Poisson incidence processes and different clump size distributions are compared. A mixed Poisson incidence process with a zero-truncated negative binomial distribution for the clump sizes is shown to give an adequate description, suggesting that the acquisition of hydatid cysts in the sheep population is heterogeneous, and that the clump sizes are aggregated. The estimates of the parameters derived from the data take plausible values.

\textsuperscript{*}Corresponding author. Address: Winterthurerstrasse 190, 8057 Zurich. Tel. +41-44-635 58 92. Email address: dominik.heinzmann@math.uzh.ch
The average infection rate and the clump size distribution are comparable in both data sets. Goodness-of-fit measures indicate that the model fits the data reasonably well.

**Keywords:** Compound processes, clumped infection, mixed Poisson, parasite data, *Echinococcus*.

## 1 Introduction

Parasitic disease data often consist of counts of a parasite (or an intermediate stage) in an animal, together with the animal’s age. The data typically exhibit two well-known features, a substantial proportion of zeros and skewed positive counts \[1, 2, 3\], meaning that some hosts harbor many parasites while most have just a few. To analyze such aggregated parasite data, the fitting of the negative binomial distribution is a common method, as in \[4\] to model the abundance of the fluke *Diplostomum spathaceum* in fish, in \[5\] for *European red mite* on apple leaves, in \[6\] for the tapeworms *Echinococcus granulosus and multilocularis* in dogs, in \[7\] for the nematode *Trichinella spiralis* in rabbits and in \[8\] for the larval stage of the mites *Allothrombium pulvinum Ewing* in lice. However, these models do not take into account the age of the hosts, which is known to influence the parasite pattern \[9, 10, 11\]. To incorporate age, negative binomial regression can be used, as in modeling the age-dependent frequency of the nematode *Wuchereria bancrofti* in humans \[12\], or of the nematodes *Ostertagia gruehneri* and *Marshallagia marshalli* in reindeer \[13\]. The approaches in both studies allow one to model (exponentially) increasing or decreasing mean parasite burdens as a function of age, in the latter study with a rather complicated relation between the over-dispersion parameter and mean of the negative binomial distribution.
and the covariate age. However, they do not provide any biological reason as to why this should occur.

While the negative binomial model takes aggregation into account, it may not adequately deal with high numbers of parasite-free hosts. For that purpose, zero-inflated (ZI) models [14, 15, 16] and two-part conditional (TPC) models [17, 18] can be used. These have been shown to outperform the negative binomial regression [19] for applications with an excess of zeros. These models introduce a state $A$ in which the only counts are zeros, and a state $B$, in which the counts could be either zeros or positive values (ZI), or only positive values (TPC). The model parameters are $p_A$, the probability to be in state $A$, and the parameters of the conditional distribution given state $B$. The parameters (or combinations thereof) can be allowed to depend on covariates. In [20], a ZI negative binomial regression was applied to model egg counts of different gastrointestinal nematodes in fecal samples from young cattle by parametrizing $p_A$ and the mean of the negative binomial distribution as functions of age. A TPC was used in [21] for modeling the density of the nematode *Wuchereria bancrofti* in mosquitoes. They argued that a zero count of microfilariae in the blood sampled by a mosquito can arise either because the human bitten is uninfected or because the blood taken from an infected human happened to contain no microfilariae. They fitted a negative binomial TPC to the aggregated data, but did not attempt to fit the underlying age-dependent model that they envisaged, because of its prohibitive complexity.

Alternatively, mechanistic models are used to understand the mechanisms leading to aggregation in the parasite distribution in hosts. A vital source of such aggregation is the infection of hosts by parasite clumps rather than single parasite ingestions [22, 23]. [24] and [25] used infinite compartmentalisation of
hosts, according to their burdens of 0, 1, 2, 3, \ldots parasites per host, to model the transmission of *Schistosomiasis* between the definitive hosts, humans, and the intermediate hosts, water snails, by assuming clumped infections. The intermediate host is not explicitly modelled and they assume that there is no superinfection in humans. [26] used moment closure equations to describe the immunoepidemiology of trochostrongylid nematodes in wild ruminant populations. The infection of hosts is modelled by an (inhomogeneous) compound Poisson process to account for clumped infections, and they consider nonlinear effects such as immunity and parasite-induced host mortality. Their model contains many parameters; some were fixed based on values from other studies, and the remainder were estimated from the model. However, their model describes the mean parasite burden, but not the prevalence of infection in animals. [27] modelled the transmission dynamics between hosts and free-living larvae with an infinite system of differential equations based on clumped infections, allowing for superinfection. Then they assumed that parasites are distributed in hosts according to a negative binomial distribution, leading to a simplified four-dimensional system, whose qualitative behavior they discussed. However, it is difficult to estimate the model parameters for such diseases, as for example the rate at which larvae are produced by adult parasites, since appropriate data sets are in general not available. [22] used a model that allows several parasite stages, clumped infections and between-host heterogeneity, to describe macroparasitic transmissions involving a free-living parasite stage. As before, estimation of the parameters is difficult since this requires the knowledge of the distribution of the numbers of parasite larvae and mature parasites in hosts and of the life distribution since maturation.

In this paper, biologically interpretable mechanistic models for hydatid cysts
in sheep, caused by the parasite *Echinococcus granulosus* (E.g.) [1, 28], are discussed. E.g. causes echinococcosis, a (re-)emerging hydatid disease in many parts of the world and, in particular, in Eastern Europe and the former Soviet Union [29, 30, 31]. E.g. is also potentially dangerous for humans. For this disease, it can be assumed that the cysts survive their hosts but do not replicate, and that there is no parasite-induced mortality and no acquired immunity in sheep [1, 32]. This implies a simpler infection dynamics than for example that encountered by [22] and [26]. Compound processes ([33, p.49], [34, p.25], [35, p.22]) are used to investigate the biological hypotheses that clumped (super)infections and heterogeneity in the acquisition of infection in the host population can explain the substantial proportion of zeros and thus the prevalence of infection, and the skewed positive counts of E.g. cysts in sheep.

The processes explicitly describe the underlying infection process and thus allow a natural modeling of aggregation and excess of zeros of the parasite distribution in the hosts. The prevalence and intensity is described simultaneously. The parameters can be estimated based on (standard) field data containing age and cyst counts of sheep. Goodness-of-fit measures are introduced to assess the performance of the model.

Based on two data sets from Kazakhstan [3] and Jordan [2], it is shown that clumped acquisition of infection by biologically heterogeneous hosts, where the clump sizes are aggregated, provides a satisfactory fit. Heterogeneity of acquisition of clumped infections may result from behavioral differences of sheep on pasture, or from differences in the immune system of sheep. Aggregation of clump sizes are reasonable given the highly aggregated adult parasite distribution in the definitive host, the dog [1]. Fitting the models yields parameter estimates which take biologically reasonable values. Goodness-of-fit measures indicate the reason-
able performance of the model.

2 Data sets and models

2.1 Empirical data

The data sets used in this paper are from Kazakhstan [3] and Jordan [2]. The Kazakhstan sample contains 2505 individual reports of the variables age and hydatid cyst burden in sheep, caused by the parasite *Echinococcus granulosus* (E.g.) [32]. The Jordan sample counts 832 individual reports of the same variables.

Hydatid cysts develop conditional on ingestion of infective biomass by sheep (intermediate host) from contaminated environment. Contamination is caused by dogs (definitive host), which harbor adult E.g. worms in the intestine and release infective eggs in the feces. Hydatid cysts form in organs such as the liver (60 – 70%), lungs and brain and develop over a period of years in the sheep. Cysts do not proliferate inside their hosts, but protoscoleces are produced inside the cysts which play a role in the infection of the definitive host [32]. It can be assumed that cysts survive their hosts, that there is no parasite-induced mortality and no acquired immunity in sheep [1, 32].

The records were obtained at necropsy in abattoirs with examination of the viscera of the sheep, including the lungs and liver, for the presence of hydatid cysts. The ages of the sheep were estimated from the stage of dentition and by questioning the owners of the animals. Small immature cysts were not recorded, as resources were not available for the systematic slicing of organs. A more detailed discussion of the applied sampling frame can be found in [2] and [3].

In the Kazakhstan sample, the mean and median ages are 2.037 and 2 years respectively. The interquartile range is 1 – 3 years and the maximum age is 8
years. The prevalence in sheep is 0.363 (0.344, 0.382). Conditional on infection, a proportion of 0.774 (0.745, 0.800) harbors 1 – 10 cysts, 0.186 (0.161, 0.213) 11 – 30 cysts and the remaining 0.041 (0.029, 0.056) have more than 30 cysts. The maximal cyst burden is 64. In the Jordan sample, the mean and median ages are 2.267 years and 1 year respectively. The interquartile range is 0.5 – 4 years and the maximum age is 10 years. The prevalence is 0.293 (0.263, 0.325). Conditional on infection, a proportion of 0.672 (0.609, 0.730) have 1 – 10 cysts, 0.234 (0.183, 0.293) 11 – 30 and 0.094 (0.062, 0.140) harbor more than 30 cysts. The maximal burden is 80 cysts. The observations in both samples agree with other study areas in Central Asia [31].

2.2 Compound Poisson process

The positive cyst burdens of *Echinococcus granulosus* in sheep are in general in the range of 1 – 80 cysts per sheep [1, 3, 36]; the majority of cyst counts in sheep in both our data sets are rather low, with a large proportion of zeros. Since there is no acquired immunity in hosts [37, 38], and cysts survive for the lifetime of the sheep, the observations suggest a low infection rate and clumped ingestions of infective eggs. Sheep potentially make many random contacts with infective dog feces on pasture, but only a small proportion of the contacts lead to an infection. Thus the resulting infection process can be viewed as a thinning of the point process at which contacts with potential infective dog feces are made. A reasonable assumption for E.g. is that the transmission system of the parasite is in a steady state [2, 28, 36], so that the ingested clumps can be supposed to be identically distributed and the low incidence rate can be supposed to be constant. Additionally, we assume that clumps are independent since infected dogs spread their feces widely, so that consecutive infections of a sheep are likely to be due to
feces from different dogs. Possible clustering due to reinfection of a sheep with
the same feces can be neglected since clumps in the environment have a relatively
short survival time and the incidence rate is low.

The above assumptions make compound processes [33, 34, 35] a suitable choice
for modeling the cyst burdens in sheep. Let the random variable \( Y_t \) denote the
total number of cysts established in an individual up to time \( t \). Then

\[
Y_t = \sum_{j=1}^{N_t} S_j,
\]

where \( (N_t)_{t \geq 0} \) is a Poisson process with constant rate \( \mu \) describing the number of
clumps ingested by an individual sheep during the time interval \([0, t]\) and \( S_j \) (\( j = 1, 2, \ldots \)) are i.i.d. random variables with distribution \( Q \) on the positive integers
\( \mathbb{N} \), independent of \( N_t \), which describe the numbers of successfully established
cysts per ingested clump. The distribution of \( Y_t \) is given by

\[
\mathbb{P}_t = \sum_{k=0}^{\infty} \mathbb{P}(N_t = k)Q^{*k} = \sum_{k=0}^{\infty} \frac{e^{-\mu t}(\mu t)^k}{k!}Q^{*k},
\]

where \( Q^{*k} \) is the \( k \)th convolution of \( Q \). In particular,

\[
p_0(t) := \mathbb{P}(Y_t = 0) = e^{-\mu t}.
\]

The expectation and the variance of \( Y_t \) are

\[
\mathbb{E}(Y_t) = \mathbb{E}(N_t)\mathbb{E}(S_1) \quad \text{and} \quad \text{Var}(Y_t) = \mathbb{E}(N_t)(\text{Var}(S_1) + [\mathbb{E}(S_1)]^2).
\]
2.3 Compound mixed Poisson process

To account for possible heterogeneity in the rate of acquisition of clumped infections within the sheep population, for example caused by differential immune response between sheep, the Poisson process \((N_t)_{t \geq 0}\) with fixed rate \(\mu\) can be replaced by a mixed Poisson process \((\tilde{N}_t)_{t \geq 0}\), where the infection rate is a non-negative random variable \(M\).

It follows that

\[
\mathbb{P}(\tilde{N}_t = n) = \int_0^\infty \frac{e^{-\mu t} (\mu t)^n}{n!} dH(\mu),
\]

where \(H(\mu) = \mathbb{P}(M \leq \mu)\) and \(H(0) = 0\). The distribution function \(H\) of \(M\) is also referred to as the structure distribution of the mixed Poisson process [39]. A special case is the simple Poisson process where the random variable \(M\) is degenerate at some \(\mu > 0\). Mixed Poisson processes are particular examples of Cox processes or doubly stochastic Poisson processes [35, p.7].

An appropriate choice of \(H\) in (3) should provide a reasonably close approximation to the true distribution, should be easy to fit and should yield a useful interpretation of the parameters. The two-parameter gamma distributions offer a flexible and tractable family, with parameters conveniently identified as measures of skewness and scale. Let \(H\) be the distribution function of a gamma distributed random variable with shape and scale parameters \(\psi, \xi > 0\) such that

\[
dH(\mu) = \frac{1}{\xi^\psi \Gamma(\psi)} \mu^{\psi-1} e^{-\frac{\mu}{\xi}} d\mu,
\]

where \(\Gamma\) is the gamma function. Then

\[
\mathbb{P}(\tilde{N}_t = n) = \frac{t^n}{\xi^\psi \Gamma(\psi) n!} \int_0^\infty \mu^{\psi+n-1} e^{-\frac{\mu}{\xi}} d\mu
\]
and, since \( \int_0^\infty z^n e^{-az} \, dz = n!a^{-n-1}, \)

\[
P(\tilde{N}_t = n) = \frac{\Gamma(\psi + n)}{\Gamma(\psi)n!} \left( \frac{1}{t\xi + 1} \right)^\psi \left( \frac{t\xi}{t\xi + 1} \right)^n.
\]

Equation (5) describes a negative binomial distribution, with \( \text{Var}(\tilde{N}_t) > \mathbb{E}(\tilde{N}_t), \)

where

\[
\mathbb{E}(\tilde{N}_t) = \psi \xi t =: at \quad \text{and} \quad \text{Var}(\tilde{N}_t) = (\psi \xi t)(1 + \xi t) =: at + bt^2.
\]

Using (5) in (1), the distribution of \( Y_t \) becomes

\[
\tilde{P}_t = \sum_{k=0}^\infty \frac{\Gamma(\psi + k)}{\Gamma(\psi)k!} \left( \frac{1}{t\xi + 1} \right)^\psi \left( \frac{t\xi}{t\xi + 1} \right)^k Q^k.
\]

In particular,

\[
\tilde{p}_0(t) := \tilde{P}(Y_t = 0) = \left( \frac{1}{t\xi + 1} \right)^\psi,
\]

where \( \tilde{P} \) is the probability measure under \( \tilde{N}_t \) as counting process. Setting \( \xi = \mu/\psi, \) for fixed \( n, t \) and \( \mu, \) (5) becomes

\[
P(\tilde{N}_t = n) \rightarrow e^{-\mu t} (\mu t)^n / n!, \quad \psi \rightarrow \infty,
\]

where the exponential term in the limit is based on Euler’s formula \( \exp(x) = \lim_{N \rightarrow \infty} (1 + (x/N))^N, \) for any real \( x. \) The limit is thus a Poisson distribution.
3 Decompounding and estimation

Decompounding [40] defines the procedure of obtaining the base distribution \( Q \) and the Poisson rate parameter \( \mu \) based on a sample of the compound process \((P_t)_{t \geq 0}\). Given a parametric form of the discrete distribution \( Q \), the convolution \( Q^{*k} \) can easily be computed and (1) respectively (7) can be fitted to the data by the maximum likelihood estimation method. This approach is easy to implement and provides reasonable computational performance, since cyst burdens in sheep are mostly rather low, the maximal burdens being of magnitude 80. Since \( Q \) is defined on the positive integers, \( Q^{*k} \) needs only be computed for small \( k \)'s. In addition, simulation from the fitted model is computationally fast (we will use the fitted model in a subsequent paper).

A nonparametric alternative to estimate the distribution \( Q \) is presented in [40]. Using an empirical estimator for the distribution of \( Y_t \) for \( t \) fixed, an estimator for the distribution of the \( S_i \)'s is obtained by a suitable inversion of the Panjer recursions [41] of the distribution of \( Y_t \). As shown in [40], the procedure requires an accurate empirical estimation of the distribution of \( Y_t \) for each \( t \). Since the sheep in our sample are of many different ages and the loads are heavily skewed, it is difficult to obtain an appropriate empirical estimate of the distribution of \( Y_t \) for the nonparametric procedure.

Suppose that \( Q \) is the zero-truncated \( \text{Po}(\eta) \) distribution. Then the following result [42] is useful.

**Theorem 3.1.** Let \( S_j \ (1 \leq j \leq n) \) be i.i.d. zero-truncated \( \text{Po}(\eta) \) distributed
random variables, so that $\mathbb{P}(S_j = s) = \eta^s / (s!(e^\eta - 1))$ for $s \in \mathbb{N}$. Then

$$\mathbb{P}\left(\sum_{j=1}^{n} S_j = z\right) = \begin{cases} \frac{\eta^z}{z!(e^\eta - 1)^n} \sum_{k=0}^{n} (-1)^k (n-k)^{\binom{n}{k}} & \text{if } n \leq z \in \mathbb{N} \\ 0 & \text{else.} \end{cases}$$

To take into account aggregation of the clump size distribution, let $Q$ be the zero-truncated negative binomial distribution, so that for $s \in \mathbb{N},$

$$\mathbb{P}(S_j = s) = \frac{\Gamma(\theta + s)}{\Gamma(\theta) s!} \left(\frac{\zeta}{\zeta + 1}\right)^s \left(\frac{1}{\zeta + 1}\right)^{\theta n} \sum_{k=1}^{n} (-1)^{n-k} \binom{n}{k} \left(\theta^k + z - 1\right)^{\frac{\theta k + z - 1}{z}}$$

where $\theta$ is the shape and $\zeta$ is the scale parameter of the negative binomial distribution. Then the following results [43] applies.

**Theorem 3.2.** Let $S_j (1 \leq j \leq n)$ be i.i.d. zero-truncated negative binomial distributed random variables specified by (10). Then for $z \in \mathbb{N},$

$$\mathbb{P}\left(\sum_{j=1}^{n} S_j = z\right) = \begin{cases} \frac{1}{\left[1 - \left(\frac{1}{\zeta + 1}\right)^{\theta n}\right]} \left(\frac{\zeta}{\zeta + 1}\right)^z \left(\frac{1}{\zeta + 1}\right)^{\theta n} \sum_{k=1}^{n} (-1)^{n-k} \binom{n}{k} \left(\theta^k + z - 1\right)^{\frac{\theta k + z - 1}{z}} & \text{if } n \leq z \\ 0 & \text{else.} \end{cases}$$

Let $\mathbb{P}_\Omega$ be the probability measure corresponding to the compound Poisson process if $\Omega = \mu$ and to the compound mixed Poisson process if $\Omega = (\psi, \xi)$; let $N_t$ denote the corresponding incidence process. Then, $\mathbb{E}(Y_t|N_t = n) = n\mathbb{E}(S_1)$ and $\text{Var}(Y_t|N_t = n) = n\text{Var}(S_1)$. Hence for the a zero-truncated Poisson clump distribution,

$$\mathbb{E}(Y_t|N_t = n) = \frac{mn}{1 - e^{-\eta}} , \quad \text{Var}(Y_t|N_t = n) = \frac{mn}{1 - e^{-\eta}} \left(1 - \frac{\eta}{e^\eta - 1}\right).$$
and for a zero-truncated negative binomial clump distribution,

$$\mathbb{E}(Y_t | N_t = n) = \frac{n\theta \zeta}{1 - (1/(\zeta + 1))^{\theta}}$$

(11)

and

$$\text{Var}(Y_t | N_t = n) = n \left[ \frac{\theta \zeta (1 + \zeta + \theta \zeta)}{1 - (1/(\zeta + 1))^{\theta}} - \left( \frac{\theta \zeta}{1 - (1/(\zeta + 1))^{\theta}} \right)^2 \right].$$

(12)

Expressions (1) and (7) can be used with Theorems 3.1 and 3.2 to compute the unconditional distribution of $Y_t$,

$$P_{\Omega}(Y_t = j) = \begin{cases} 
  P_{\Omega}(N_t = 0) & \text{if } j = 0 \\
  \sum_{k=1}^j P_{\Omega}(N_t = k) P(\sum_{t=1}^k S_t = j) & \text{if } j \geq 1.
\end{cases}$$

(13)

Given independent realizations $y_i (1 \leq i \leq n)$ of $Y_t$ at time points $t_i$, the log-likelihood function is

$$l(\Omega, \eta) = \sum_{i=1}^n \left\{ I_{\{y_i=0\}} \ln P_{\Omega}(N_t = 0) + I_{\{y_i>0\}} \ln \left[ \sum_{k=1}^{y_i} P_{\Omega}(N_t = k) P(\sum_{t=1}^k S_t = y_i) \right] \right\},$$

(14)

where $I$ is the indicator function. The log-likelihood function for the case of a single ingestion mechanism, with clump size fixed to be 1, is thus

$$l_2(\Omega) = \sum_{i=1}^n \ln P_{\Omega}(N_t = y_i).$$

(15)

Let us introduce the following model notation for the rest of the paper. The single ingestion models with Poisson and mixed Poisson incidence process are
denoted by P/1 and MP/1 respectively. The compound process \((Y_t)_{t\geq 0}\) with \((N_t)_{t\geq 0}\) a Poisson process and with the clump size distribution \(Q\) specified to be the zero-truncated Poisson distribution is denoted by P/ztP, and if \((N_t)_{t\geq 0}\) is a mixed Poisson process, then the model is denoted by MP/ztP. Analogously, if the clump size distribution is specified to be the zero-truncated negative binomial distribution, we denote the resulting models by P/ztP and MP/ztP, depending on the incidence process.

4 Application

Parameter estimates for the models of interest are obtained from the two data sets of Kazakhstan and Jordan (Section 2.1). We test single against clumped infection, heterogeneity of the Poisson rate parameter of the incidence process, and aggregation of the clump size distribution. Then we compare the best fitting models for the two data sets and assess the goodness-of-fit.

4.1 Clumped infection

First, we compare the single ingestion models P/1 and MP/1 to the compound processes P/ztP and MP/ztP respectively using a standard likelihood ratio test based on (14) and (15) with 1 degree of freedom. The log-likelihood values are reported in Table 1. Testing the P/1 against the P/ztP results in p-values of < 0.001 for Kazakhstan and Jordan. Similarly, testing the MP/1 against the MP/ztP also results in p-values of < 0.001 for Kazakhstan and Jordan. Hence there is strong evidence for a clumped infection process in both samples.

[Table 1 about here.]
4.2 Heterogeneity in acquisition and aggregated clump sizes

In (9), we have seen that, if $\xi = \mu/\psi$ with $\mu$ fixed and $\psi \to \infty$, then the MP/ztP model converges to the P/ztP model. To test if the acquisition of hydatid cysts of sheep is heterogeneous, we have to test the null hypothesis $H_0 : \xi = 0$ against $\xi > 0$. Analogously, to test if the clump size distribution is aggregated, we note that if $\zeta = \eta/\theta$ with $\eta$ fixed and $\theta \to \infty$, then the P/ztP model converges to the P/ztP model, and thus we need to test $H_0 : \zeta = 0$ against $\zeta > 0$. Clearly, the MP/ztP and the P/ztPb models are also nested within the MP/ztP model, which allows heterogeneity in the acquisition of cysts together with an aggregated clump size distribution. For the tests with $H_0 : \xi = 0$ and $H_0 : \zeta = 0$, we test a parameter which is on the boundary of the parameter space under $H_0$. [44] showed that the asymptotic distribution of the likelihood ratio test statistic in the presence of a parameter that is on the boundary of the null hypothesis is $\frac{1}{2} \chi^2_0 + \frac{1}{2} \chi^2_1$, a 50 : 50 mixture of $\chi^2_0$ and $\chi^2_1$ distributions. Given the observed test statistic $\tilde{\chi}$, the p-value is given by $(\Pr(\chi^2_0 > \tilde{\chi}) + \Pr(\chi^2_1 > \tilde{\chi}))/2$.

Applying the likelihood ratio test with the above asymptotic $\chi^2$ mixture distribution to the reported log-likelihood values in Table 1 implies that the P/ztPb and the MP/ztP model both fit the Kazakhstan and Jordan samples significantly better than the P/ztP (all p-values smaller than 0.001). In addition, the MP/ztPb fits the two samples significantly better than the P/ztPb (p-values for Kazakhstan < 0.001 and Jordan 0.027) and the MP/ztP models (p-values for Kazakhstan and Jordan < 0.001).

To verify the asymptotic distribution of the test statistic under $H_0$, we apply a Monte Carlo method and simulate data under $H_0$ (simpler model), then fit both the simpler and more complex model to the generated data sets and compute the likelihood test statistic. For the generation of the data sets, starting with the
original ages $t_k$ ($1 \leq k \leq n$) of the $n$ sheep in the sample, a new cyst burden is attributed to each of them as a realization of the simpler model with $t = t_k$, with the model parameters fixed at their estimated values given in Table 2. Repeating this procedure 2000 times yields an approximating reference distribution of the test statistic under $H_0$. Testing the P/zt nb model against the MP/zt nb model for the Jordan sample implies a p-value of 0.035, which is slightly larger than the p-value of 0.027 obtained by using the asymptotic reference distribution. The other p-values computed with the simulated reference distribution also differ slightly from the ones obtained with the asymptotic reference distribution, however they are also smaller or equal to 0.002. It appears that our samples are too small to be able to rely completely on asymptotics. However, the test results with the simulated reference distribution also imply that the MP/zt nb model significantly better fits the data sets from Kazakhstan and Jordan than the other models. We conclude that there is evidence in the data that the acquisition of hydatid cysts of *Echinococcus granulosus* by sheep is heterogeneous, and that the clump size distribution is aggregated.

[Table 2 about here.]

Table 2 shows the estimates of the MP/zt nb model for the parameters $a = \psi \xi$ and $b = \psi \xi^2$ of the incidence process $N_t$ defined in (6) and for the mean $c := \mathbb{E}(Y_t|N_t = 1)$ and variance $d := \text{Var}(Y_t|N_t = 1)$ of the clump size distribution defined in (11) and (12). The parameter $a$ is not significantly different in the samples from Kazakhstan and Jordan, suggesting that a sheep gets infected on average every third year. The parameter $b$ is significant larger in the Kazakhstan sample, so that the variance of the infection rates $\text{Var}(N_t) = at + bt^2$ is larger for this sample. The difference of the variance of the infection rate in the two
samples is especially pronounced in older sheep since $\text{Var}(N_t) \sim bt^2$. The resulting gamma mixture distributions (4) of the infection rate for the two samples are plotted in Figure 1, indicating that in the Kazakhstan sample, the infection rates are more heterogeneous than in the Jordan sample. Table 2 also indicates that the estimated mean and variance for the clump size distribution are not significantly different in the two samples, suggesting that the number of successfully established cysts per infection is similar in the two samples. Thus on average, an infective clump leads to about $4 - 5$ established cysts in the sheep.

[Figure 1 about here.]

The fitted MP/ztnb model provides estimates for the prevalence of infection as well as for the probability mass function (pmf) of the positive loads. Figure 2 shows the estimated prevalence of infection for the MP/ztnb model for the Kazakhstan and Jordan samples together with the observed prevalences. In both samples, the estimated prevalence of the MP/ztnb explains the observations reasonably well.

[Figure 2 about here.]

The estimated pmf of the MP/ztnb model for the age classes reported in Figure 2 are displayed in Figure 3 for the Kazakhstan and in Figure 4 for the Jordan sample. Given an age class, the fitted pmf are computed as mixture of the pmf’s corresponding to the different ages within the class. The fitted pmf are reasonable in both samples, taking into account the small number of observed positive loads in some of the age classes, especially in the Jordan sample.

[Figure 3 about here.]

[Figure 4 about here.]
4.3 Goodness-of-fit

The goodness-of-fit of the MP/ztnb model is evaluated as follows. Divide the sheep into age classes, and treat the observations in the different classes as i.i.d. data. The classes are specified as in Figure 2. The observed and estimated distributions of cysts are then compared within each age class using an appropriate statistic. Note that, as before, the resulting pmf for an age class is a mixture of the pmf’s corresponding to the different ages within that class.

With the age classes as before, let $n_i \leq i \leq 6$ be the number of animals in age class $i$, and stratify them with respect to load into $c_i$ strata. Then two possible goodness-of-fit measures for the distribution of the numbers of cysts within any given age class $i$ are

\[ \chi^2 := \sum_{k=1}^{c_i} \frac{(m_{ik} - \mathbb{E}(M_{ik}))^2}{\mathbb{E}(M_{ik})} \]

and

\[ L := \sum_{k=1}^{c_i} \left| \mathbb{E} \left( \frac{M_{ik}}{n_i} \right) - \frac{m_{ik}}{n_i} \right| , \]

where $M_{ik}$ is a random variable describing the numbers of animals of age class $i$ having cyst counts in stratum $k$ ($1 \leq k \leq c_i$), and $m_{ik}$ is the (corresponding) observed count.

The number of strata $c_i$ for age class $i$ is chosen to be the maximal number such that the expected number of counts in each stratum is at least 10. The strata in the age classes are computed for the model with parameters fixed by their estimates in Table 2. To generate the reference distribution of $\chi^2$ and $L$, a Monte Carlo approach is used, where data sets are generated under the MP/ztnb model. Given the original ages $t_k$ ($1 \leq k \leq n$) of the $n$ sheep in the sample, a new cyst burden is attributed to each of them as a realization of the MP/ztnb model with $t = t_k$ and the parameters fixed by their estimates given in Table 2.
We then fit the MP/ztmb model to this new data set, and compute with the new estimates the test statistics for each of these sets. We use the same stratification of the age classes as before. The observed values of the two test statistics can then be compared to the reference distributions for each age class $i$.

Figures 5 and 6 display the results for the samples from Kazakhstan and Jordan for 1000 simulations. For the Kazakhstan sample, the observed values of the test statistics $\chi^2$ and $L$ (indicated by a solid line) although consistently large, are in reasonable agreement with the simulated distributions for all age strata. For the Jordan sample, the solid line lies well outside the simulated distribution in age class $(0, 1)$. This is for two reasons. First, the observed prevalence in that age class is overestimated by the model (see Figure 2). Secondly, there are only 12 positive loads in that class, which are not well described by the model. However, the results in the other age classes suggest that the model fit is reasonable.

The model seems to have some tendency to underestimate the zero load stratum and to overestimate the numbers of high cyst counts in the first age class. The opposite tendency can be observed in the age strata $4 - 6$. Since only 4 parameters are used in the model, to fit the distributions of prevalence and cyst burden observed in 6 different age classes, a perfect fit can hardly be expected.

[Figure 5 about here.]

[Figure 6 about here.]

5 Conclusion

In this paper, different mechanistic models are used to explain the acquisition of hydatid cysts in sheep, caused by the parasite *Echinococcus granulosus*. The
models allow one to test the biologically interesting hypotheses of clumped infections, host heterogeneity with respect to infection and aggregation of clump sizes. The experimentally supported assumptions of *Echinococcus granulosus* cyst infection in sheep such as life-long survival of cysts in the host, no replication inside the host, no parasite-induced mortality and no acquired immunity in sheep imply simpler infection dynamics than for example those encountered by [22] and [26], as discussed in the introduction to this paper. Hence our models are straightforward to fit to the most commonly available data sets, which only contain the ages and cyst burdens of the sheep. The models provide age-dependent estimates for the prevalence of infection and for the probability mass functions of positive cyst burdens in sheep.

The application of the models to two data sets from Kazakhstan and Jordan supports a clumped infection process, with a rate of acquisition of infection which is heterogeneous within the population, and with clump sizes which are aggregated. The infection process is described by a compound mixed Poisson process with a zero-truncated negative binomial distribution for the number of cysts per ingested clump. The goodness-of-fit measures indicate that the chosen model describes the given data reasonably well, but not perfectly. The estimates suggest a mean infection rate of about 0.315 infections per year and a mean clump size of about 4.5 cysts, suggesting that on average every third year, a sheep will ingest an infectious clump, each clump leading to approximately 4 – 5 established hydatid cysts in the sheep. The results indicate that the observed aggregation in the distribution of cysts among sheep may be the result both of differences between sheep and also of clumped infections.

Our model can be used to investigate how changes in the underlying parameters may affect the parasite distribution, and thus may be useful in assessing
control programs for *Echinococcus granulosus*. In particular, it can be used as sub-process for describing infections in the sheep population in a fully stochastic model for the complete life-cycle of *Echinococcus granulosus*.

**Acknowledgements** The authors gratefully acknowledge the comments and suggestions of two referees and the handling editor, that greatly improved the presentation. This work was supported by the Schweizerischer Nationalfonds (SNF), project no. 107726.

**References**


Figure 1: Estimated gamma density function (4) of the infection rate in the incidence process for the samples from Kazakhstan (solid line) and Jordan (dashed line).
Figure 2: Fitted prevalence curves $\hat{q}(t) = 1 - 1/(t\hat{\xi} + 1)^{\hat{\psi}}$ (with $\hat{\psi}$ and $\hat{\xi}$ given in Table 2) for the MP/ztnb model for the samples from Kazakhstan and Jordan, together with the observed prevalences and their 95% confidence intervals. The observed prevalences are computed for the age classes $(0, 1]$, $(1, 2]$, $(2, 3]$, $(3, 4]$, $(4, 5]$, 5+, where 5+ summarizes all sheep older than 5 years. For the age classes $1 - 4$, the majority of the observed ages coincide with the end points of the interval. The prevalences are plotted at the means of the ages of the animals in the corresponding classes.
Figure 3: Estimated probability mass functions of the MP/ztmb model for the positive loads of the Kazakhstan sample for the age classes (a) (0, 1], (b) (1, 2], (c) (2, 3], (d) (3, 4], (e) (4, 5], (f) 5+, together with a histogram of the corresponding observed quantities. The class sizes are 185, 315, 282, 84, 29 and 15. For a better presentation of the results, the following points are not plotted in the histograms: 64 (with corresponding mass 0.003) in age class (1, 2], 47 and 57 (mass 0.119 each) in age class (3, 4] and 56 (mass 0.034) in age class (4, 5].
Figure 4: Estimated probability mass functions of the MP/ztnb model for the positive loads of the Jordan sample for the age classes (a) (0, 1], (b) (1, 2], (c) (2, 3], (d) (3, 4], (e) (4, 5], (f) 5+, together with a histogram of the corresponding observed quantities. The class sizes are 12, 14, 23, 29, 47 and 119. For an better presentation of the results, the following points are not plotted in the histograms: 52 and 63 (mass 0.021 each) in age class (4, 5], and 57 (mass 0.008) and two loads of 80 (combined mass 0.016) in age class 5+. 
Figure 5: Goodness-of-fit of the MP/ztmb model in the Kazakhstan sample. The observed values of the test statistics (solid lines) $\chi^2$ ((a1)-(a8)) and $L$ ((b1)-(b8)) are plotted with the corresponding simulated distributions under the MP/ztmb model with parameters fixed with its estimates in Table 2 for the age classes (x1) $(0, 1]$, (x2) $(1, 2]$, (x3) $(2, 3]$ (x4) $(3, 4]$, (x5) $(4, 5]$ and (x6) $5+$, with x=a,b.
Figure 6: Goodness-of-fit of the MP/ztnb model in the Jordan sample, analogously to Figure 5.
Tables

Table 1: Log-likelihood values for the models fitted to the Kazakhstan and Jordan samples, together with the number of parameters in the models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Kazakhstan</th>
<th>Jordan</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>P/I</td>
<td>−10648.570</td>
<td>−2643.109</td>
<td>1</td>
</tr>
<tr>
<td>MP/1</td>
<td>−4230.321</td>
<td>−1133.255</td>
<td>2</td>
</tr>
<tr>
<td>P/ztP</td>
<td>−4647.557</td>
<td>−1161.142</td>
<td>2</td>
</tr>
<tr>
<td>P/zt nb</td>
<td>−4179.769</td>
<td>−1018.524</td>
<td>3</td>
</tr>
<tr>
<td>MP/ztP</td>
<td>−4180.413</td>
<td>−1079.412</td>
<td>3</td>
</tr>
<tr>
<td>MP/zt nb</td>
<td>−4160.347</td>
<td>−1016.665</td>
<td>4</td>
</tr>
</tbody>
</table>
Table 2: Maximum likelihood estimates for the parameters and key quantities of the MP/ztnb model for the Kazakhstan and Jordan samples, together with 95% confidence intervals computed by the bootstrap percentile method. The parameters $a = \psi \xi$ and $b = \psi \xi^2$ of the incidence process $N_t$ are defined in (6), so that $\mathbb{E}(N_t) = at$ and $\text{Var}(N_t) = at + bt^2$. The mean $c := \mathbb{E}(Y_t | N_t = 1)$ and variance $d := \text{Var}(Y_t | N_t = 1)$ of the clump size distribution are defined in (11) and (12) respectively.

<table>
<thead>
<tr>
<th></th>
<th>Kazakhstan</th>
<th>Jordan</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\psi$</td>
<td>0.941 (0.629, 1.260)</td>
<td>5.154 (2.579, 8.061)</td>
</tr>
<tr>
<td>$\xi$</td>
<td>0.343 (0.225, 0.741)</td>
<td>0.060 (0.029, 0.173)</td>
</tr>
<tr>
<td>$\theta$</td>
<td>0.351 (0.139, 0.617)</td>
<td>0.212 (0.126, 0.442)</td>
</tr>
<tr>
<td>$\zeta$</td>
<td>5.859 (3.215, 9.763)</td>
<td>7.861 (5.394, 10.565)</td>
</tr>
<tr>
<td>$\hat{a}$</td>
<td>0.323 (0.237, 0.499)</td>
<td>0.309 (0.195, 0.521)</td>
</tr>
<tr>
<td>$\hat{b}$</td>
<td>0.111 (0.064, 0.198)</td>
<td>0.019 (0.008, 0.061)</td>
</tr>
<tr>
<td>$\hat{c}$</td>
<td>4.186 (2.343, 6.276)</td>
<td>4.500 (2.724, 7.022)</td>
</tr>
<tr>
<td>$d$</td>
<td>19.798 (10.177, 29.828)</td>
<td>27.125 (14.411, 35.917)</td>
</tr>
</tbody>
</table>