QMRA for Drinking Water: 2. The Effect of Pathogen Clustering in Single-Hit Dose-Response Models

Vegard Nilsen* and John Wyller

Spatial and/or temporal clustering of pathogens will invalidate the commonly used assumption of Poisson-distributed pathogen counts (doses) in quantitative microbial risk assessment. In this work, the theoretically predicted effect of *spatial* clustering in conventional "single-hit" dose-response models is investigated by employing the *stuttering Poisson* distribution, a very general family of count distributions that naturally models pathogen clustering and contains the Poisson and negative binomial distributions as special cases. The analysis is facilitated by formulating the dose-response models in terms of probability generating functions. It is shown formally that the theoretical single-hit risk obtained with a stuttering Poisson distribution is lower than that obtained with a Poisson distribution, assuming identical mean doses. A similar result holds for *mixed Poisson* distributions. Numerical examples indicate that the theoretical single-hit risk is fairly insensitive to moderate clustering, though the effect tends to be more pronounced for low mean doses. Furthermore, using Jensen's inequality, an upper bound on risk is derived that tends to better approximate the exact theoretical single-hit risk for highly overdispersed dose distributions. The bound holds with any dose distribution (characterized by its mean and *zero inflation index*) and any *conditional* dose-response model that is *concave* in the dose variable. Its application is exemplified with published data from Norovirus feeding trials, for which some of the administered doses were prepared from an inoculum of aggregated viruses. The potential implications of clustering for dose-response *assessment* as well as practical *risk characterization* are discussed.

KEY WORDS: Aggregation; clustering; dose-response; overdispersion; QMRA; stuttering Poisson

1. INTRODUCTION

In both natural and engineered systems, waterborne microbial pathogens such as viruses, bacteria, and protozoan parasites may, in principle, exist in aqueous suspensions as completely dispersed single pathogens or they may instead be spatially associated to some extent, in aggregates/clusters/clumps.[1–3] The extent and strength of the association will depend on the pathogen concentration, the processes that resulted in aggregation, the mechanisms by which pathogens are associated, and the physicochemical properties of the water. Some processes may introduce pathogens in the water in a clumped form, e.g., if a host sheds pathogens that are aggregated, if solids with accumulated pathogens detach from filter media, or if parts of biofilms separate. In the latter two cases, spatially associated pathogens are likely to be part of a large, complex particle that may not be easily dissociated. In other cases, it may be primarily electrostatic forces that hold pathogens together, and such interaction is likely to be more sensitive to changes in the environment of the pathogens.

Department of Mathematical Sciences and Technology, Norwegian University of Life Sciences, N-1432 Aas, Norway.

*Address correspondence to Vegard Nilsen, Department of Mathematical Sciences and Technology, Norwegian University of Life Sciences, P.O. Box 5003, N-1432 Aas, Norway; vegard.nilsen@nmbu.no; vgnils@gmail.com.

0272-4332/16/0100-0163$22.00/1 © 2016 Society for Risk Analysis

Within the field of quantitative microbial risk assessment (QMRA)[4] for drinking water, a baseline assumption is that pathogen numbers in water samples are Poisson distributed. In particular, this is a common assumption in the development and application of conventional semi-mechanistic *single-hit* dose-response models[4–9] that provide the probabilistic link between pathogen exposure levels (dose) and the resulting health consequences (response) for exposed individuals. The Poisson assumption is appropriate when pathogens are completely and randomly dispersed in the water source throughout the time period of interest.

However, in practice, it is most commonly observed that the variance in pathogen counts is larger than what can be accommodated by the one-parameter Poisson distribution.[10–13] This *overdispersion* will result if pathogens accumulate in space or time in excess of that which could occur by chance in a completely dispersed suspension. The phenomenon of *temporal* variation[13] in pathogen concentrations is well known, documented, and attempts are often made to account for it in applications. It can, e.g., be caused by relatively slow variation in raw water quality due to seasonal effects or could be the result of sudden changes such as treatment plant failures. *Spatial* accumulation of pathogens in the form of physical clusters, i.e., two or more pathogens sticking together or to the same suspended particle, is more difficult to document experimentally, and information on the pathogen clustering state is practically never available in applications.

Conventional water treatment (both drinking water and waste water) involving coagulation/flocculation processes is designed to promote particle aggregation in order to enhance downstream particle separation processes. This treatment is likely to affect pathogens (that are particles) to some extent as well, although difficult to verify and quantify experimentally. On the other hand, the generally low concentration of pathogens in drinking water implies that the average distance between pathogens is much larger than the pathogens themselves, reducing the chance of pathogens colliding and sticking together. Furthermore, colloid stability theory[1,14] predicts increasing dispersion of microorganisms at low ionic strength and pH-values away from their isoelectric points (typically less than neutral pH), which coincide with common conditions in drinking water.

Nevertheless, some empirical indications of clustering do exist. Gale and co-workers showed[10,11] that the variation between replicate counts of bacterial spores in water samples increased significantly after water treatment. Clustering would indeed produce such overdispersion, but independent confirmation of physical clustering is needed to fundamentally distinguish it from temporal variation in mean spore concentrations and/or variation in analytical recovery between samples. In another case of possible clustering,[15] polio virus plaques grown from sewage samples were shown to contain two different types of polio viruses, clashing with the standard assumption that each plaque arises from a single virus particle. Among several possible explanations, the authors found aggregation of viruses to be the more plausible. Clustering was also observed during electron microscopy in a protein-rich laboratory stock suspension of Norovirus that was used in human feeding trials for dose-response assessment.[16] The latter has motivated efforts to represent clustering in single-hit dose-response models.[16–18]

In general, at least four aspects of QMRA may be identified, in which the clustering state of pathogens may impact the analysis:

1. Clustering may obscure interpretation of microorganism counts from laboratory methods. First, it could possibly affect the recovery of concentration procedures. Second, some methods typically return results that relate to the total number of organisms, such as quantitative real-time polymerase chain reaction (qPCR) that measures the number of genome copies present. Other methods will tend to return results that relate more to the total number of clusters, such as plaque/colony counting methods where it is difficult to assess whether a macroscopic plaque/colony stems from a single organism or a cluster of organisms.[15] A dispersion step (e.g., using Tween[10]) may be added to the laboratory protocol of the latter methods to obtain the total numbers of organisms instead.
2. Clustering may play a role in the exposure assessment in a broad sense, since the transport properties of pathogens in nature and their removal and inactivation during water treatment and distribution may depend on the extent of clustering. For example, settling and filtration processes are size-sensitive, as well as disinfection processes such as chlorination (see, for example, Thurston-Enriquez *et al.*[19]) and ultraviolet radiation (where clustering of pathogens/particles may shield pathogens from radiation).

3. Clustering could affect pathogen infectivity upon entering a human host. That is, for a given number of pathogens ingested, is it relevant for the host-pathogen interaction whether they occur as single particles or are part of a cluster of a certain size? Any such dependence would induce a correlation between the dose and the infectivity of a single pathogen (since the dose and occurrence of certain cluster sizes would be correlated), which is inconsistent with traditional single-hit models (Section 2.1). If such effects exist and are important, dose-response models would require modification to account for them, which would complicate the modeling process. Designing an experiment that can detect and quantify such effects, if they are present, appears challenging.
4. Clustering affects the dose distribution. Even if the host is insensitive to the pathogen clustering state, clustering of pathogens will affect the probability distribution for the total number of pathogens included in a water sample (the *dose*), whether the "sample" is for human consumption or for laboratory analysis. The choice of a dose distribution (usually Poisson) is an integral part of the development of classical single-hit dose-response models, as well as in designing Monte Carlo simulations for practical risk characterization.

This article focuses on item 4 above; i.e., the effect of clustering on the dose distribution (total pathogen count) as it applies to single-hit dose-response models. Regarding item 3, it will be assumed that the host/pathogen interaction is insensitive to pathogen clustering state. This is potentially unrealistic, but has nevertheless been the assumption (tacitly or explicitly) in published work on Norovirus dose response[16–18] and it seems difficult to relax in a simple way. A primitive generalization of single-hit models to account for the effects mentioned in item 3 is provided in Section S.5 of the online appendix.

The introductory paragraphs above motivate the purpose of the present work, which is to

i. Investigate the theoretically predicted effect of pathogen clustering on single-hit dose-response models in QMRA; i.e., what is fundamentally built into the single-hit risk framework with respect to the effects of pathogen clustering (or more generally, overdispersion in the dose distribution)?
ii. Simulate the effects of moderate clustering on single-hit risk estimates, a situation that may be particularly relevant for background risk levels in drinking water. Are single-hit models robust with respect to unaccounted for clustering?
iii. Introduce a risk bound (the Jensen bound) that emerged during the investigation of bullet point i, which could be useful for many situations where one has an overdispersed dose distribution.

For some of the technical derivations, we will draw upon the dose-response model formulation in terms of probability generating functions (pgfs) presented in the companion paper.[9]

In discussing dose-response modeling, we should distinguish between *dose-response assessment* and dose-response models as employed in practical *risk characterization* studies. The purpose of dose-response assessment is to estimate dose-response parameters for a particular pathogen, which can subsequently be used in a dose-response model to estimate infection risk in a risk characterization study, possibly undertaken as a simulation study using Monte Carlo methods. The dose distributions need not be the same in the two cases, and if it is non-Poisson due to clustering, it will not be known in any detail. For dose-response assessment, it is very convenient[1] if the dose-response model can be expressed in closed form, which limits the choice of dose distribution to simple ones. For risk characterization, this is less important since a complicated dose distribution may easily be specified in a Monte Carlo study, in conjunction with a conditional[7] dose-response model (Section 2.1). The material presented in this article should be useful for both purposes.

2. MODEL DEVELOPMENT

The semi-mechanistic single-hit dose-response framework has been described by many authors.[4–9] We first recapitulate the essentials of this framework (Section 2.1) using the formulations of our companion paper,[9] before introducing some basic concepts of clustering in Section 2.2. Section 2.3 introduces the *stuttering Poisson* distribution, which forms the basis for the analysis presented in Section 3.1.

[1] Although not necessary if the required quantities can be computed numerically with sufficient precision.

2.1. Single-Hit Dose-Response Framework

In a single-hit model, it is assumed that a *single* pathogen may be capable of causing an infection, and that individual pathogens act *independently* of each other. Under more precise assumptions stated at the end of this section, a randomly selected host that ingests a random[2] number of pathogens X has a probability P_{I} of becoming infected, which equals the probability that at least one pathogen establishes infection:

$$P_{\mathrm{I}} = 1 - (1 - R)^X. \tag{1}$$

Here, R is a random variable that equals the probability that a single pathogen establishes infection (the *single-hit probability*). We allow for the possibility that X (as in so-called conditional dose-response models) and/or R (as, e.g., in the exponential model) may degenerate to constants. It has been shown[(7–9)] that within single-hit theory, R derives its randomness from the variation in host susceptibility and that the variation in pathogen infectivity enters only indirectly through its modulating effect on the distribution of R.

The actual dose-response model is given by the marginal probability $\mathrm{E}(P_{\mathrm{I}})$ (the expected value of P_{I}) as a function of the dose distribution parameters. $\mathrm{E}(P_{\mathrm{I}})$ serves as a dose-dependent success probability in a binomial model for the number of infected hosts when one or more hosts are exposed. It can be written as:

$$\begin{aligned}\mathrm{E}(P_{\mathrm{I}}) &= 1 - \int_0^1 \sum_{x=0}^{\infty} (1-r)^x p_X(x) f_R(r)\,\mathrm{d}r \\ &= 1 - \int_0^1 G_X(1-r) f_R(r)\,\mathrm{d}r, \end{aligned} \tag{2}$$

where $p_X(x)$ is the probability mass function (pmf) of X, $f_R(r)$ is the probability density function[3] (pdf) of R, and $G_X(1-r)$ is the pgf of X, evaluated at $1-r$. The pgf is an alternative representation of the distribution of a count random variable, and the basics of pgfs are reviewed in Section S.1 of the online appendix since they play a central role in our dose-response models.

The baseline assumption in QMRA is that X is Poisson distributed with pmf:

$$\Pr(X = x) = p_X(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \tag{3}$$

which has a single parameter $\lambda > 0$ and $\mathrm{E}(X) = \mathrm{Var}(X) = \lambda$. The parameter λ can be interpreted as the product cv of the pathogen concentration c in the water source and the sample volume v. The pgf of a Poisson variable with mean λ is:

$$G_X(z) = e^{\lambda(z-1)}. \tag{4}$$

With this, Equation (2) reduces to:

$$\mathrm{E}(P_{\mathrm{I}}) = 1 - \int_0^1 e^{-\lambda r} f_R(r)\,\mathrm{d}r. \tag{5}$$

Various parameterized dose-response models will result for different choices of f_R.[(4,9)]

Inherent in the simple formulation in Equation (2) are several statistical independence assumptions on random variables representing host susceptibility, pathogen infectivity, and the dose. For their precise formulation, the companion paper[(9)] should be consulted. They can be summarized briefly as follows:

1. The probability that any single pathogen establishes infection is independent of the failure of one or more other pathogens within the same dose to do so.
2. The infectivities of the individual pathogens in the water sample are mutually independent.
3. The dose and the infectivity of each individual pathogen in the water sample are mutually independent.
4. The dose and the susceptibility of the host are mutually independent.

2.2. Nomenclature and Basic Concepts

One of the axioms used in a rigorous development of the Poisson distribution says that, roughly, for a very small sample volume, the probability of observing more than one pathogen is zero.[(20)] The presence of pathogen clustering will obviously invalidate this assumption and the dose distribution will not be Poisson anymore. In practice, deviations from the Poisson distribution may be identified by a statistically significant difference between the sample mean and variance. A useful statistic in this respect is the dispersion index:

$$\delta = \frac{\mathrm{Var}(X)}{\mathrm{E}(X)}. \tag{6}$$

[2] Throughout this article, strict adherence is made to the convention of denoting random variables with uppercase letters and particular instances of the same variables with the corresponding lowercase letters.

[3] R may also be represented as a mixed random variable[(9)] with both continuous and discrete parts (e.g., if some hosts are fully immune), in which case f_R is a mixed probability density/pmf.

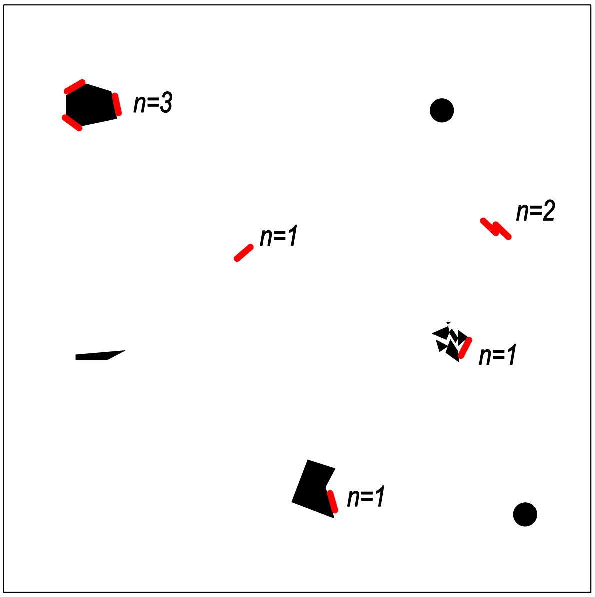

Fig. 1. Example of mild aggregation.

One may distinguish between the following situations:

- Underdispersion, i.e., $\delta < 1$. This can happen when there is a tendency toward special uniformity in the distribution of pathogens, and results in a pathogen count that is "less random" (its entropy is lower) than a Poisson variable.
- Poisson dispersion, i.e., $\delta = 1$.
- Overdispersion, i.e., $\delta > 1$. This is the type of deviation that is most commonly observed in practice,(10,11) and could be the effect of pathogen clustering.

Another useful measure of spread that will be employed below is the zero-inflation index, defined by:

$$\theta = 1 + \frac{\ln[p_X(0)]}{\mathrm{E}(X)}. \qquad (7)$$

In general, $\theta < 1$. For a Poisson variable, $\theta = 0$, while $\theta > 0$ for a situation with clustering, as discussed in Section 2.3.

Assume now that a sample is taken from a water source in which some of the pathogens may be clustered. Fig. 1 shows a conceptual example of how pathogens may be distributed in a sample from such a water source with (moderate) clustering. Some pathogens exist as single particles, some are clustered together, and some are attached to other types of particles of various sizes. We will use the term *n-cluster* for any collection of particles that contains n ($n \geq 1$) pathogens, in which the association between the pathogens is sufficiently strong that the cluster behaves as a single unit during sampling. With this terminology, the simplest cluster is the one consisting of a single pathogen (a 1-cluster). Furthermore, clusters are characterized only by the number of pathogens they contain, and *not* by the number and size of other types of particles included in the cluster.

The number of n-clusters contained in the water sample is a random variable and will be denoted as X_n. The total number of pathogens contained in the sample, X, and the total number of *clusters*, X_{cl}, are functions of the X_ns and given, respectively, by:

$$X = \sum_{n=1}^{\infty} nX_n = X_1 + 2X_2 + 3X_3 + \cdots \qquad (8)$$

$$X_{\mathrm{cl}} = \sum_{n=1}^{\infty} X_n = X_1 + X_2 + X_3 + \cdots \qquad (9)$$

The sums are over all cluster sizes with the assumption that $\mathrm{E}(X) < \infty$ (and hence $\mathrm{E}(X_{\mathrm{cl}}) < \infty$).

Since X_n represents the count of a specific type of cluster, clustering itself is no longer a source of overdispersion in the distribution of X_n (e.g., if two n-clusters form a new cluster, they are instead counted as a $2n$-cluster). Hence, if clusters can be considered to move about essentially randomly and independently, it is natural to assume that the distribution of each X_n is Poisson with corresponding parameter $\lambda_n = c_n v$, where c_n represents the concentration (number per unit volume) of n-clusters in the water source. The general distribution of X under this assumption is considered in Section 2.3.

It is worth emphasizing again the similarities and differences between clustering as defined above and other sources of spatiotemporal heterogeneity in the distribution of pathogens. If some of the pathogens tend to stay close in space or time, without actually being physically clustered, this will also contribute to overdispersion in the dose distribution and can be difficult, if not impossible, to distinguish from clustering only on the basis of observing pathogen counts. Temporal variation in the mean pathogen concentration on larger time scales will also induce overdispersion. While these sources of overdispersion may possibly also be representable by a distribution of the form of Equation (8), the interpretation of the parameters in terms of clusters is lost. The main focus of this article is on suspensions that have a given mean pathogen concentration $\lambda = \mathrm{E}(X)$, for which some of the pathogens are actually clustered, and the clusters themselves behave as Poisson particles.

2.3. A General Dose Distribution Accounting for Clustering

We are interested in the distribution of X as expressed in Equation (8), where the X_ns are assumed to be Poisson distributed. It is demonstrated in Section S.1 of the online appendix that the distribution of X is, in fact, a general *stuttering Poisson distribution*,[20–22] i.e., a Poisson-stopped sum of nonnegative discrete random variables. Special cases of this distribution have, for example, been used to model bulk arrivals in queuing theory[21] and the number of radiation-induced chromosome defects.[23] For the case where there is a fixed maximum cluster size $N > 1$, the distribution of X has been called the Nth-order (univariate) Hermite distribution.[24] For $N = 2$, it is known simply as the Hermite distribution.[25,26] This special case was used to model bacterial counts as early as 1926[27] (although the name "Hermite" distribution was coined later) and may be of particular importance for dilute suspensions, where larger clusters are unlikely to form. For the case where the only cluster sizes are 1 and N, it is known as the generalized Hermite distribution.[28]

The stuttering Poisson distribution may become very complicated (e.g., many modes), owing to the essentially combinatorial character of the problem of obtaining it (Section S.1 of the online appendix). Its pmf is generally not expressible in closed form, but can be obtained as a convenient recursive formula[22] that evaluates quite rapidly on an ordinary computer as long as the mean of the stuttering Poisson distribution is only moderately large. A proof of the following expression for the pmf is reproduced in Section S.1 of the online appendix (Lemma 1):

$$p_X(x) = \begin{cases} e^{-\sum_{n=1}^{\infty} \lambda_n} & \text{if } x = 0 \\ = \frac{1}{x} \sum_{n=1}^{x} n\lambda_n p_X(x-n) & \text{if } x \geq 1. \end{cases} \tag{10}$$

We may use Equation (10) to compute the dose distribution resulting from any given clustering state in the water source, which is specified by the set of parameters $\lambda_n = c_n v$, $n = 1, 2, \ldots$. Table S1 in Section S.1 of the online appendix gives expressions for δ and θ as a function of the parameters λ_n.

A special case of the stuttering Poisson is the two-parameter negative binomial distribution, which has been used to accommodate a larger than Poisson variance in QMRA studies.[10,16,18] For the negative binomial, the distribution of cluster sizes follows a logarithmic series distribution with parameter $0 < a < 1$. The details are given in Section S.1 of the online appendix. When parameterized in terms of the mean λ and a dispersion parameter $b = a/(1 - a)$, the pmf is given by:

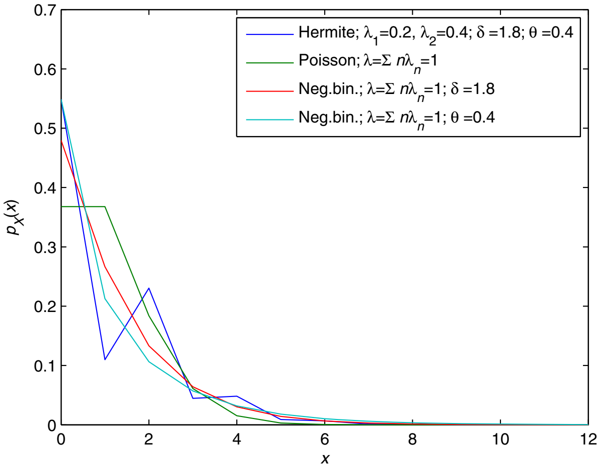

Fig. 2. Comparison of the Hermite distribution with the Poisson (equivalent λ) and the negative binomial (equivalent λ and δ, or λ and θ).

$$p_X(x) = \frac{\Gamma(x + \lambda/b)}{x!\Gamma(\lambda/b)} \left(\frac{b}{b+1}\right)^x \left(\frac{1}{b+1}\right)^{\lambda/b}. \tag{11}$$

The variance is $\mathrm{Var}(X) = \lambda(1 + b)$. The negative binomial reduces to the Poisson distribution with mean λ as $b \to 0$. Its pgf is:

$$G_X(z) = [1 + b(1 - z)]^{-\lambda/b}, \tag{12}$$

which we will use in Section 3.1.

In Fig. 2, an example of the Hermite distribution ($N = 2$) is compared with the Poisson distribution (identical means) and the negative binomial distribution (identical means/dispersion indexes or means/zero inflation indexes). The example represents a situation where as much as 80% of the pathogens are contained in 2-clusters, which accentuates the jagged nature of the Hermite distribution. It is seen that, compared to the Poisson, the three other distributions give a higher probability of obtaining zero pathogens and lower probability of obtaining exactly 1.

It is interesting to compare some key general properties of the Poisson distribution with mean λ and a stuttering Poisson with the same mean, i.e., $\lambda = \sum_{n=1}^{\infty} n\lambda_n$ (in terms of pathogen concentrations,

$c = \sum_{n=1}^{\infty} nc_n$). The detailed expressions for the moments have been left to Table S1 in Section S.1 of the online appendix. The important fact is that the variance in the dose distribution will always increase after clustering, and therefore the dispersion index δ also increases. If there is a maximum cluster size N, it can easily be shown that $1 \leq \delta \leq N$, where $\delta = 1$ if and only if X is simple Poisson and $\delta = N$ if and only if all the pathogens are contained exclusively in N-clusters. Therefore, if a reliable estimate of δ can be obtained experimentally, it gives an indication of cluster sizes: there are at least some clusters greater than or equal to δ. However, obtaining a reliable δ estimate may be difficult in practice, requiring that we sample from a stationary distribution for X and that analytical procedures have a constant 100% recovery efficiency.

Since $p_X(0)$ increases as a result of clustering (which means that the zero-inflation index θ also increases), the probability of getting at least one pathogen always decreases. Slightly counterintuitive, the probability of getting exactly one pathogen may increase or decrease, even though the concentration of 1-clusters always decreases. The direction of change depends on details of the clustering state. However, for low pathogen concentrations ($\lambda < 1$), we can show that $p_X(1)$ always decreases after clustering. Consider the fraction:

$$\frac{\Pr(X = 1)_{\mathrm{cl}}}{\Pr(X = 1)_{\mathrm{disp}}} = \frac{\lambda_1 e^{-\sum_{n=1}^{\infty} \lambda_n}}{\lambda e^{-\lambda}} = \frac{\lambda_1 e^{-\lambda_1}}{\lambda e^{-\lambda}} e^{-\sum_{n=2}^{\infty} \lambda_n}. \quad (13)$$

The last exponential is always less than 1. Inspection of the function $\lambda e^{-\lambda}$ will show that it is strictly increasing for $0 < \lambda < 1$. Thus, since $\lambda > \lambda_1$, the fraction $(\lambda_1 e^{-\lambda_1})/(\lambda e^{-\lambda})$ will always be less than 1 for $0 < \lambda < 1$. Typically, the expected pathogen dose in a glass of water will rarely exceed 1.

3. ANALYTICAL RESULTS AND EXAMPLES

3.1. Dose Response with Stuttering Poisson Doses

Fortunately, the dose-response expression in Equation (2) requires not the complicated pmf of X, but instead the pgf, which has a simple expression. In Section S.1 of the online appendix, it is shown that it is given by:

$$G_X(z) = \exp\left(\sum_{n=1}^{\infty} \lambda_n (z^n - 1)\right). \quad (14)$$

For any given λ, we may reparameterize the stuttering Poisson distribution by letting $q_n = \frac{n\lambda_n}{\lambda}$, i.e., q_n denotes the fraction of the total pathogen count that is contained in n-clusters. With this, Equation (14) becomes:

$$G_X(z) = \exp\left(\lambda \sum_{n=1}^{\infty} \frac{q_n}{n} (z^n - 1)\right). \quad (15)$$

Using the pgf of the stuttering Poisson (Equation (14)) in the general single-hit expression in Equation (2) gives us the dose-response relation:

$$\mathrm{E}(P_{\mathrm{I,sPo}}) = 1 - \int_0^1 \exp\left\{\sum_{n=1}^{\infty} \lambda_n [(1 - r)^n - 1]\right\} f_R(r)\,\mathrm{d}r. \quad (16)$$

Thus, within the single-hit theoretical framework, we may specify the parameters of the stuttering Poisson distribution corresponding to any given clustering state, and use Equation (16) to compute the (expected) probability of infection. Given the generality of the above expression, it is conjectured that it may encompass most, if not all, plausible "single-hit" dose-response relationships unless the dose distribution is underdispersed ($\delta < 1$), but this seems to be rare for microbial counts. It reduces to the conventional dose-response relationships (exponential, beta-Poisson) for specific choices of the parameters λ_n and the distribution f_R.[4,9]

The dose-response formulation in Equation (16) enables us to show quite generally that clustering, as represented by a stuttering Poisson distribution, always decreases the (expected) probability of infection in a single-hit model. We formulate this main result as a proposition, with the proof left to the Appendix.

Proposition 1 (Risk with stuttering Poisson doses). *Let the dose X be stuttering Poisson distributed with $\lambda_N > 0$ for some $N > 1$ (i.e., there exists some clusters) and fix the mean $\mathrm{E}(X) = \lambda = \sum_{n=1}^{\infty} n\lambda_n$. Then, the corresponding single-hit risk $\mathrm{E}(P_{\mathrm{I,sPo}})$ is bounded from above by $\mathrm{E}(P_{\mathrm{I,Po}})$, the single-hit risk computed using a Poisson distribution with the same mean λ.*

Proposition 1 is illustrated in Fig. 3, which shows a contour plot of the following ratio:

$$\frac{\mathrm{E}(P_{\mathrm{I}})}{\mathrm{E}(P_{\mathrm{I,He}})} = \frac{1 - e^{-\lambda r}}{1 - e^{-\lambda r(1 - \frac{1}{2} q_2 r)}}. \quad (17)$$

This is the ratio of the risk computed with a Poisson distribution (i.e., the exponential model)

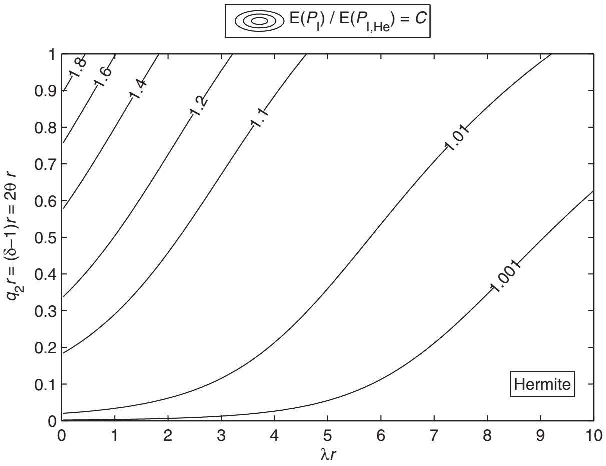

Fig. 3. Contour plot of the ratio in Equation (17), comparing the risk computed with the Poisson distribution (exponential model) to that computed with the Hermite distribution. q_2 is the proportion of pathogens in 2-clusters. A corresponding plot assuming beta-distributed R is given in Fig. S1 in the online appendix.

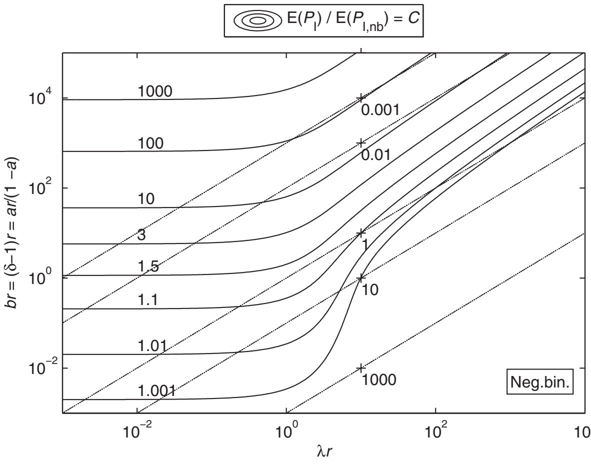

Fig. 4. Contour plot (solid lines) of the ratio in Equation (18), comparing the risk computed with the Poisson distribution (exponential model) to that computed with the negative binomial distribution. Dotted lines indicate a constant value of λ/b. A corresponding plot assuming beta-distributed R is given in Fig. S2 in the online appendix.

to the risk computed with the Hermite distribution, assuming a constant single-hit probability r. The denominator is obtained from Equation (16) with only λ_1 and λ_2 nonzero, r constant, and using $q_2 = 2\lambda_2/\lambda$. It is the simplest possible comparison between a clustered/non-clustered situation, but it may potentially be of practical relevance in dilute suspensions for pathogens that fit the exponential model. Furthermore, it uncovers some general tendencies of interest. First, for any r (single-hit probability) and q_2 (proportion of pathogens in 2-clusters), the effect of clustering becomes less important as the mean λ of the distributions increases. Second, for any λ and q_2, the effect of clustering becomes negligible when r becomes small since we are then approaching the lower-left corner of the plot. Third, even for small λ and large r, the effect of clustering is negligible unless q_2 is quite large. In summary, the effect of clustering only becomes important for jointly small λ, large r, and large q_2 ($rq_2 \gtrsim 0.2$ is required for a ratio of 1.1 or larger). The ratio in Equation (17) is bounded from above by 2. Fig. S1 in the online appendix generalizes Fig. 3 to the case of beta-distributed R, parameterized in terms of $\mathrm{E}(R) = \alpha/(\alpha+\beta)$ and α. The effect of clustering generally increases with $\mathrm{E}(R)$ and q_2 and decreases with α. For any given $\mathrm{E}(R)$ and q_2, the effect of clustering is relatively small unless both α *and* λ are small.

Fig. 4 is similar to Fig. 3 and shows a contour plot of:

$$\frac{\mathrm{E}(P_{\mathrm{I}})}{\mathrm{E}(P_{\mathrm{I,nb}})} = \frac{1 - e^{-\lambda r}}{1 - (1 + br)^{-\lambda/b}}, \tag{18}$$

where the denominator is the risk computed using the negative binomial distribution (no host heterogeneity), obtained by using Equation (2) with Equation (12). Here, the extent of clustering increases with the dispersion parameter $b = \delta - 1$. The situation is a bit more complicated than in Fig. 3. It is still correct that clustering becomes negligible as r or b becomes very small. When holding r and b constant while decreasing λ, the ratio reaches a near steady state for $\lambda r < 1$. The ratio is above 1.1 if $\lambda r < 1$ and $br > 0.25$. The effect of increasing r while holding b and λ constant (moving along dotted lines) depends on whether λr is below (ratio increases) or above (ratio decreases) 1. For drinking water applications, it will usually be below 1. Fig. S2 in the online appendix generalizes Fig. 4 to the case of beta-distributed R. The effect of clustering generally increases with $\mathrm{E}(R)$ and b and decreases with α.

3.2. Dose Response with Mixed Poisson Doses

For completeness, we now consider an alternative generalization of the dose distribution known as mixed Poisson distributions. Here, the Poisson parameter λ is considered to be randomly drawn from a so-called mixing distribution that represents the variation in λ. Such distributions have, e.g., been used to model seasonal variations in pathogen count in raw water (e.g., the Poisson log-normal distribution). In the case of mixed Poisson doses, the pmf $p_X(x)$ of the dose distribution is obtained by marginalizing the Poisson distribution over λ:

$$p_X(x) = \int_0^\infty p_{X_{\mathrm{Po}}}(x) f_\Lambda(\lambda)\,\mathrm{d}\lambda = \int_0^\infty \frac{\lambda^x e^{-\lambda}}{x!} f_\Lambda(\lambda)\,\mathrm{d}\lambda, \tag{19}$$

where $p_{X_{\mathrm{Po}}}$ is the pmf of a Poisson distribution with parameter λ and $f_\Lambda(\lambda)$ is the pdf of the mixing distribution. It can be demonstrated that the distribution in Equation (19) is indeed overdispersed relative to the Poisson distribution. By the law of total variance, we have $\mathrm{Var}(X) = \mathrm{Var}(\Lambda) + \mathrm{E}(\Lambda)$, which has a minimum when Λ is point mass distributed (i.e., X is Poisson and $\mathrm{Var}(\Lambda){=}0$). When f_Λ is specified, p_X may be used in the general single-hit expression (Equation (2)) to obtain a (possibly closed-form) marginal dose-response model. However, variation in λ is often more relevant for risk characterization (as opposed to dose-response assessment), for which it may be easier to sample sequentially from f_Λ and p_X during Monte Carlo simulations than it is to use a marginal dose-response model.

By an advanced theorem of probability,[20,29] a mixed Poisson distribution that is constructed from a so-called *infinitely divisible* mixing distribution will also be a stuttering Poisson distribution, so in many cases, the two families of distributions overlap (e.g., the negative binomial affords both interpretations). For completeness, though, we include the following proposition, which is the equivalent to Proposition 1, but for Poisson mixtures (the proof is left to the Appendix).

Proposition 2 (Risk with mixed Poisson doses). *Let the dose X be mixed Poisson distributed with mixing distribution $f_\Lambda(\lambda)$ and pmf given by Equation* (19). *Then, the corresponding single-hit risk* $\mathrm{E}(P_{\mathrm{I,mPo}})$ *is bounded from above by* $\mathrm{E}(P_{\mathrm{I,Po}})$*, the single-hit risk computed using a Poisson distribution with mean equal to the mean of the mixing distribution,* $\mathrm{E}(\Lambda)$.

In order to build some intuition for why Propositions 1 and 2 hold, note that the single-hit model in Equation (2) may be written:

$$\begin{aligned}\mathrm{E}(P_\mathrm{I}) &= \mathrm{E}_X[\mathrm{E}_R(P_\mathrm{I})] \\ &= \sum_{x=0}^{\infty} p_X(x) \int_0^1 [1-(1-r)^x] f_R(r)\,\mathrm{d}r,\end{aligned} \tag{20}$$

where the subscripts denote expectation with respect to the indicated random variables. The integral expression $\mathrm{E}_R(P_\mathrm{I})$ has been called a *conditional* dose-response model[7] since it gives the (expected) risk if exactly x pathogens are ingested. The essential property of $\mathrm{E}_R(P_\mathrm{I})$, which may be verified by twice differentiation under the integral sign, is that it is always *concave*[4] in x for $x \geq 0$. Furthermore, the variance of X increases when X is stuttering Poisson or mixed Poisson, as compared to a Poisson-distributed X with the same mean. Thus, in the weighted sum $\mathrm{E}_X[\mathrm{E}_R(P_\mathrm{I})]$ of conditional dose-response models, more weight is put on x-values far from the mean of X (on both sides of it). Since $\mathrm{E}_R(P_\mathrm{I})$ is concave in x (i.e., it becomes progressively flatter), the dispersion of weights may intuitively be expected to reduce the risk estimate. This property may be expected to not hold for a model that incorporates between-pathogen cooperation, which tends to introduce a convex region in the low-dose range of the conditional dose-response model (see Section S.4 of the online appendix).

While Propositions 1 and 2 agree with intuition, their strength is their generality: there exists no stuttering or mixed Poisson distribution, no matter how obscure, that increases the risk estimate compared to a Poisson distribution with the same mean. One may still ask how general these families of distributions are, and whether overdispersed count (dose) distributions that are not representable as stuttering or mixed Poisson lead to similar results as Propositions 1 and 2.[5] While we have not succeeded in finding a definitive answer, it has been shown[30] that any count random variable for which $\Pr(X = 0) > 0.5$ follows a *generalized*[6] stuttering Poisson distribution, but it is not clear to us whether the proof of Proposition 1 can be modified to cover this case.

Finally, we want to briefly mention the concept of *stochastic dominance*,[31,32] widely used in expected utility theory in economics, and a potentially useful tool also for microbial risk analysis. In particular, for any concave conditional dose-response model, second-order stochastic dominance dictates that the risk from dose distribution X_A is higher than the risk from dose distribution X_B if X_B is a so-called

[4] Often, dose-response models are plotted on log-log or semi-log plots, which gives the appearance of a convexity in the low-dose region.

[5] Section S.3 in the online appendix shows that it also holds when the X_is in Equation (8) are binomial random variables with identical success probabilities.

[6] The generalized version allows for negative λ_ns.

mean-preserving spread[33] of X_A, i.e., if $X_B = X_A + Z$ for some random variable Z and $\mathrm{E}(Z|x_A) = 0$ for all x_A.

4. AN APPROXIMATE DOSE-RESPONSE MODEL FROM JENSEN'S INEQUALITY

From Propositions 1 and 2, it is clear that the single-hit risk obtained with an overdispersed dose distribution in the form of stuttering or mixed Poisson-distributed doses is bounded from above by the risk obtained with Poisson-distributed doses. As shown in Figs. 3 and 4 and Figs. S1 and S2 in the online appendix, the difference in risk between the Poisson case and the overdispersed case may become substantial for extreme overdispersion. The following proposition gives another bound on risk that appears to be significantly closer (shown below) to the exact single-hit risk for highly overdispersed dose distributions, and could be useful for practical purposes. It is valid for *any* dose distribution (not necessarily stuttering or mixed Poisson) and *any* conditional dose-response model that is concave in the dose variable (not necessarily single-hit), and it requires only one additional parameter (the zero-inflation index) of the dose distribution compared to the Poisson distribution. The proof is again left to the Appendix.

Proposition 3. *Introduce the notation* $P_\mathrm{I}^0(x) \equiv \mathrm{E}_R(P_\mathrm{I})|_{X=x}$ *for a general concave (in x) conditional dose-response model. Then, the risk* $\mathrm{E}(P_\mathrm{I}) = \sum_{x=0}^{\infty} p_X(x) P_\mathrm{I}^0(x)$ *is bounded from above by:*

$$\begin{aligned}\mathrm{E}(P_{\mathrm{I,J}}) &= [1 - p_X(0)] \cdot P_\mathrm{I}^0\left(\frac{\lambda}{1 - p_X(0)}\right) \\ &= \left(1 - e^{\lambda(\theta-1)}\right) \cdot P_\mathrm{I}^0\left(\frac{\lambda}{1 - e^{\lambda(\theta-1)}}\right), \quad (21)\end{aligned}$$

where θ is the zero-inflation index of the distribution of X.

It is readily verified that Equation (21) satisfies some fundamental requirements of a dose-response model:

$$\begin{aligned} &0 \le \mathrm{E}(P_{\mathrm{I,J}}) \le 1, \\ &\lim_{\lambda\to 0} \mathrm{E}(P_{\mathrm{I,J}}) = 0, \quad (22) \\ &\lim_{\lambda\to \infty} \mathrm{E}(P_{\mathrm{I,J}}) = 1. \end{aligned}$$

The latter property holds only if there are no completely immune hosts.[9] The Jensen bound takes particular forms depending on which conditional dose-response model P_I^0 we choose. If R has a single point mass, the Jensen bound becomes:

$$\mathrm{E}(P_{\mathrm{I,J}}) = \left(1 - e^{\lambda(\theta-1)}\right) \cdot \left(1 - (1-r)^{\lambda/\left(1-e^{\lambda(\theta-1)}\right)}\right). \quad (23)$$

If R is beta distributed, the Jensen bound is:

$$\mathrm{E}(P_{\mathrm{I,J}}) = \left(1 - e^{\lambda(\theta-1)}\right) \cdot \left(1 - \frac{B\left[\alpha, \beta + \lambda/\left(1 - e^{\lambda(\theta-1)}\right)\right]}{B(\alpha,\beta)}\right), \quad (24)$$

where B denotes the beta function.

As mentioned, Equation (21) seems to be a very good risk bound in the single-hit case, i.e., it is quite close to the exact single-hit risk. Figs. 5 and 6 illustrate this. Fig. 5 shows a contour plot for the following ratio of the risk from the Jensen bound to the risk computed with a Hermite distribution (no host heterogeneity, i.e., a constant $R = r$):

$$\frac{\mathrm{E}(P_{\mathrm{I,J}})}{\mathrm{E}(P_{\mathrm{I,He}})} = \frac{\left(1 - e^{\lambda(\theta-1)}\right) \cdot \left(1 - (1-r)^{\lambda/\left(1-e^{\lambda(\theta-1)}\right)}\right)}{1 - e^{-\lambda r(1 - \frac{1}{2}q_2 r)}}. \quad (25)$$

Also shown (red curves (color visible in on-line version)) is the ratio in Equation (17) for comparison. For all parameter values, the Jensen bound stays within about 10% of the exact risk. In those cases where the exponential model (red curves) severely overestimates risk, the Jensen bound is markedly closer to the exact risk from the Hermite model. In other cases, where clustering is less pronounced, the exponential model tends to give a slightly more precise estimate of the exact risk than the Jensen bound. Fig. S1 in the online appendix generalizes Fig. 5 to the case of beta-distributed R, and the trends are similar; the Jensen bound performs very well overall and particularly in those cases where the exact beta-Poisson model overestimates risk.

Fig. 6 shows a contour plot for the following ratio of the risk from the Jensen bound to the risk computed with a negative binomial distribution (again, no host heterogeneity):

$$\frac{\mathrm{E}(P_{\mathrm{I,J}})}{\mathrm{E}(P_{\mathrm{I,nb}})} = \frac{\left(1 - e^{\lambda(\theta-1)}\right) \cdot \left(1 - (1-r)^{\lambda/\left(1-e^{\lambda(\theta-1)}\right)}\right)}{1 - (1+br)^{-\lambda/b}}. \quad (26)$$

Also shown (red curves) is the ratio in Equation (18) for comparison. Again, the bound seems to be very good in those cases where the exponential model (red curves) severely overestimates risk; for some parameter values close to two orders of magnitude better. Fig. S2 in the online appendix generalizes Fig. 6 to the case of beta-distributed R, and the trends are

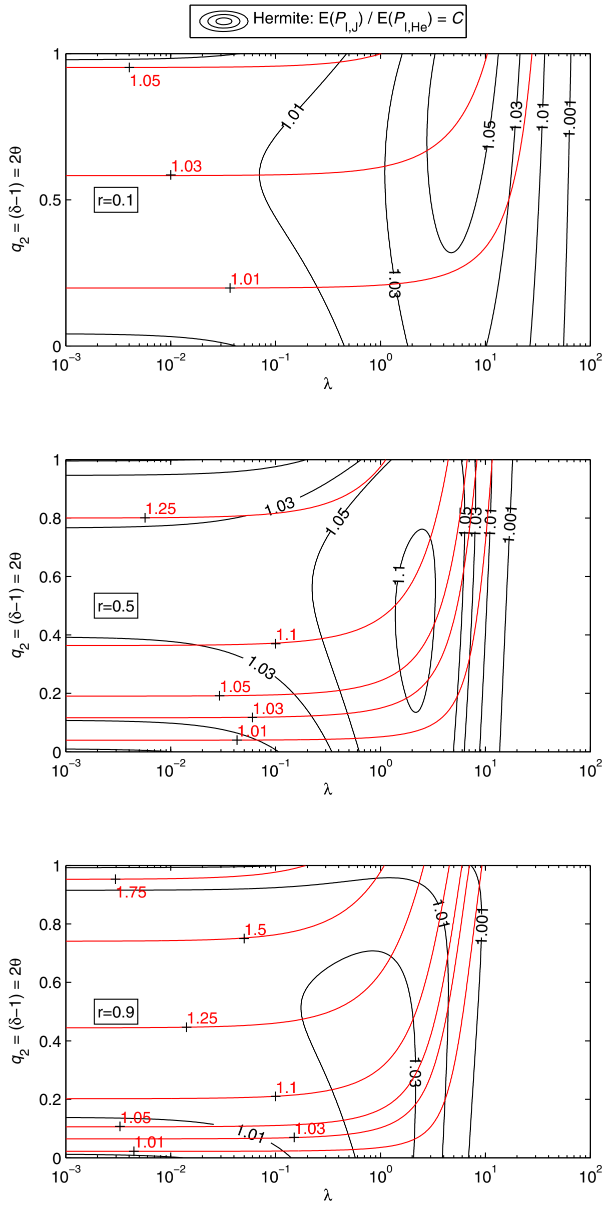

Fig. 5. Black curves are contours for the ratio in Equation (25), comparing the risk computed with the Jensen bound to that computed with the Hermite distribution. Red curves are contours for the ratio of risk computed with the Poisson distribution to that computed with the Hermite distribution. This figure is based on a constant $R = r$; Fig. S1 in the online appendix shows the case of beta-distributed R.

similar; the Jensen bound performs very well overall and particularly in those cases where the exact beta-Poisson model overestimates risk.

Finally, we compare the Jensen bound risk to the risk computed with the discrete Weibull distribution.[34] This distribution has been suggested[12,13,35] as a natural model for long-term pathogen counts in drinking water with the ability to account for rare, high-consequence events such as treatment plant failures. Hence, it can potentially model pathogen counts that are subject to *temporal* clustering. Its pmf, mean, and zero-inflation index are given by, respectively:

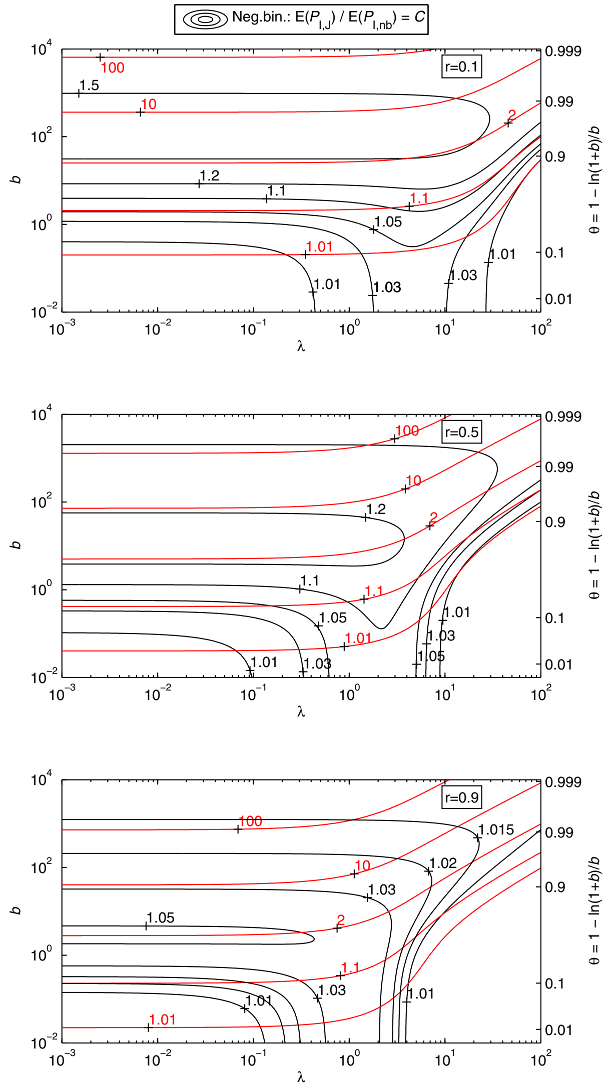

Fig. 6. Black curves are contours for the ratio in Equation (26), comparing the risk computed with the Jensen bound to that computed with the negative binomial distribution. Red curves are contours for the ratio of risk computed with the Poisson distribution to that computed with the neg.bin. distribution. This figure is based on a constant $R = r$; Fig. S2 in the online appendix shows the case of beta-distributed R.

$$p_X(x) = q^{x^\eta} - q^{(x+1)^\eta}, \tag{27}$$

$$\lambda = \sum_{x=1}^{\infty} q^{x^{\eta}}, \tag{28}$$

$$\theta = 1 + \frac{\ln(1-q)}{\lambda}, \tag{29}$$

with shape parameters $\eta > 0$ and $0 < q < 1$. The infinite sum for the mean was computed in this work by means of an approximation given by Englehardt and Li.[12] It can be shown that Equations (28) and (29) uniquely determine q and η for any given pair $\lambda > 0$ and $\theta < 1$; hence, we may reparameterize the distribution in terms of λ and θ. This was used in Figs. 7 (low θ values; θ on vertical axis) and 8 (high θ values; $1-\theta$ on vertical axis), which show contour plots of the following ratio:

$$\frac{\mathrm{E}(P_{\mathrm{I,J}})}{\mathrm{E}(P_{\mathrm{I,dW}})} = \frac{\left(1-e^{\lambda(\theta-1)}\right)\cdot\left(1-(1-r)^{\lambda/\left(1-e^{\lambda(\theta-1)}\right)}\right)}{1-\sum_{x=0}^{\infty} p_X(x)(1-r)^x}, \tag{30}$$

where X is discrete Weibull distributed. The denominator was computed to full numerical precision, i.e., until the term $p_X(x)(1-r)^x$ evaluated to 0. The trends in these figures are similar to those for the Hermite and negative binomial distribution; the Jensen bound performs very well in those cases where overdispersion causes a marked reduction in the exact risk, while it is also reasonably close to the exact risk when there is little overdispersion. Fig. S3 in the online appendix generalizes Figs. 7 and 8 to the case of beta-distributed R, and the trends are similar; the Jensen bound performs very well overall and particularly in those cases where the exact beta-Poisson model overestimates risk. In summary, the Jensen bound examples in this section appear to indicate that the single-hit risk is only moderately sensitive to the details of an overdispersed dose distribution, but quite sensitive to the overall *degree* of overdispersion, as expressed by the zero-inflation index.

5. APPLICATION OF THE APPROXIMATE MODEL: DOSE-RESPONSE FOR NOROVIRUS

Dose-response assessment for Norovirus[16–18] has been complicated by aggregation of viruses in the inoculum used for human feeding trials. Here, we fit the beta-Jensen bound (Equation (24)) to the available Norovirus dose-response data for the purposes of demonstrating its application, and for simple comparison with previous studies.

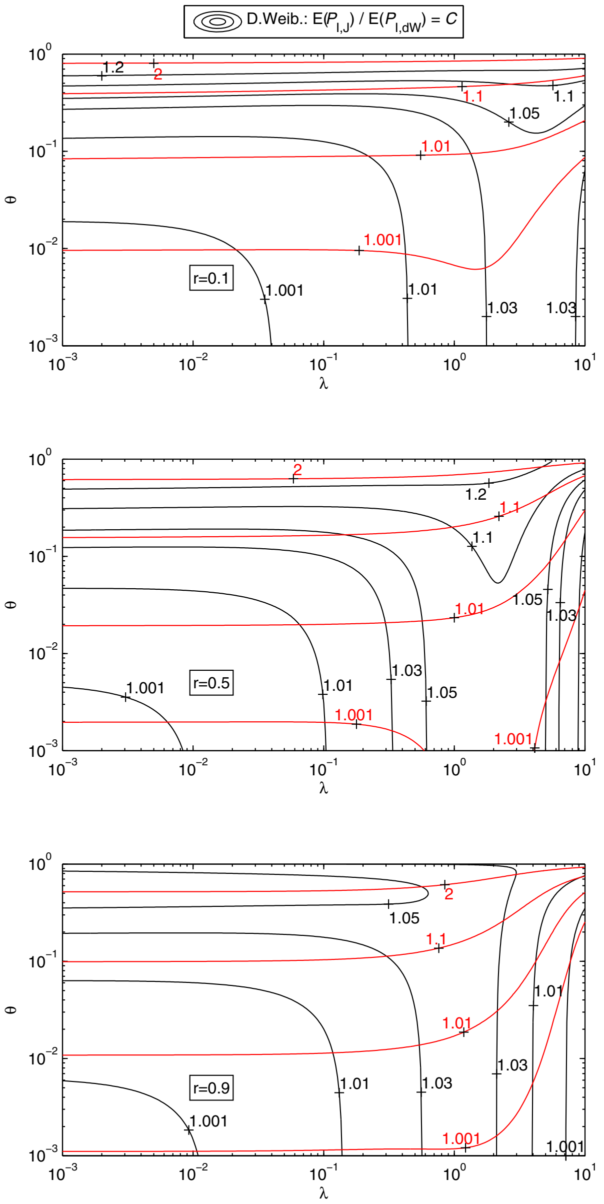

Fig. 7. Black curves are contours for the ratio in Equation (30), comparing the risk computed with the Jensen bound to that computed with the discrete Weibull distribution for low θ values. Red curves are contours for the ratio of risk computed with the Poisson distribution to that computed with the d.Wei. distribution. This figure is based on a constant $R = r$; Fig. S3 in the online appendix shows the case of beta-distributed R.

Several studies have reported Norovirus dose-response data from human feeding trials.[16,36–38] The essential data from those studies are given in Table I. In the study by Teunis *et al.*,[16] the suspension used as inoculum had been stored for a long time and, using electron microscopy, the viruses were observed to

Table I. Norovirus Dose-Response Data from Human Feeding Trials [16,36–38]

Designation	Aggregated	Source	Mean Dose (PCR Units)	Total Subjects	Infected Subjects
8fIIa GI.1	Y	Teunis *et al.*[16]			
			3.24×10^0	8	0
			3.24×10^1	9	0
			3.24×10^2	9	3
			3.24×10^3	3	2
			3.24×10^5	8	7
			3.24×10^6	7	3
			3.24×10^7	3	2
			3.24×10^8	6	5
8fIIb GI.1	N	Teunis *et al.*[16]			
			6.92×10^5	8	3
			6.92×10^6	18	14
			2.08×10^7	1	1
8fIIb GI.1	N (presumed)	Seitz *et al.*[36]			
			6.50×10^7	13	10
8fIIa GI.1	Y (presumed)	Atmar *et al.*[37]			
			1.92×10^2	13	1
			1.92×10^3	13	7
			1.92×10^4	8	7
			1.92×10^6	7	6
GII.4	Y (presumed)	Frenck *et al.*[38]			
			2.00×10^7	23	16

Note: 8fIIa: "Primary" inoculum from the original Norwalk isolate. 8fIIb: "Secondary" inoculum from stool samples of an infected individual. GI.1: Genogroup I/genotype 1. GII.4: Genogroup II/genotype 4.

be significantly clustered and could not be dispersed by sonication. The assumptions on aggregation in the other studies have been adopted here from Messner *et al.*[17] The dose levels in all these studies were determined by quantitative PCR. Recently, Norovirus was cultivated *in vitro* for the first time,[39] which may pave the way for quantification by culturing methods that will arguably be more relevant for dose-response assessment.

Table II gives an overview of the models that were fitted to the data in this work. The exact beta-Poisson model assumes completely dispersed pathogens and is included as a reference. The beta-negative binomial model was suggested and fitted by Teunis *et al.*[16] and refitted[7] to an extended data set by Messner *et al.*[17] Messner *et al.*[17] also suggested a model, termed *fractional Poisson*, in which R is Bernoulli distributed, i.e., hosts are either fully immune or fully susceptible. In that case, the model does not require the full dose distribution; only $p_X(0)$ is needed. The model contains two fitting parameters: the fraction of (fully) immune hosts, ϕ, and the mean aggregate size μ. Schmidt[18] investigated a range of models, including the previously mentioned ones, but extended all models to include a host immunity parameter and showed that the omission/inclusion of an immunity parameter may have a large effect on the results.

When fitting the Jensen bound, we have to assume that θ is constant across all dose levels, which is an assumption that warrants some attention. If we can assume that the only effect of dilution is to scale the concentration of each cluster size, it can be seen from the expression for θ in Table S1 (online appendix, Section S.1) that θ can be expected to be conserved across dilutions of a suspension, since every λ_n is scaled by the same (expected) factor. Equivalent assumptions have been made in the previously published models on Norovirus, either by stating the assumption explicitly[18] or implicitly by treating the aggregation parameter as a constant across all dose levels in a feeding trial.[16,17] In practice, however, the diluent may affect the colloidal stability and hence clustering state of the pathogens, and mechanical mixing procedures may also have an effect. Thus, there is some uncertainty

[7] Note that we arrive at parameter estimates for the beta-negative binomial model in this work that are different from those of Messner *et al.*,[17] using the same data set. We believe that the estimates reported here are correct, as our computed likelihood values agree to full reported precision with those of Schmidt.[18]

Table II. Dose-Response Models Fitted to Data in Table I

	Distr. of R	Dose distr. – agg.	$\mathrm{E}(P_{\mathrm{I}})$ – agg.	$\mathrm{E}(P_{\mathrm{I}})$ – disp.	θ
Exact beta-Poisson	Beta	Poisson	$1 - {}_1F_1(\alpha, \alpha+\beta, -\lambda)$	As for agg.	$\equiv 0$
Beta-neg.bin.	Beta	Neg.bin.	$1 - {}_2F_1(\lambda/b, \alpha; \alpha+\beta; -b)$	Ex. b.-Po.	$1 - \ln(b+1)/b$ $= 1 - 1/\mu$
Fractional Poisson	Bernoulli	$p_X(0) = e^{-\lambda/\mu}$	$(1-\phi)(1-e^{-\lambda/\mu})$	As for agg. with $\mu \equiv 1$	$1 - 1/\mu$
Beta-Jensen	Beta.	Not fully specified	Equation (24)	Ex. b.-Po.	$1 - 1/\mu$
Beta-Jensen with imm.	Beta.	Not fully specified	$(1-\phi)$ times Eq. (24)	Ex. b.-Po.	$1 - 1/\mu$

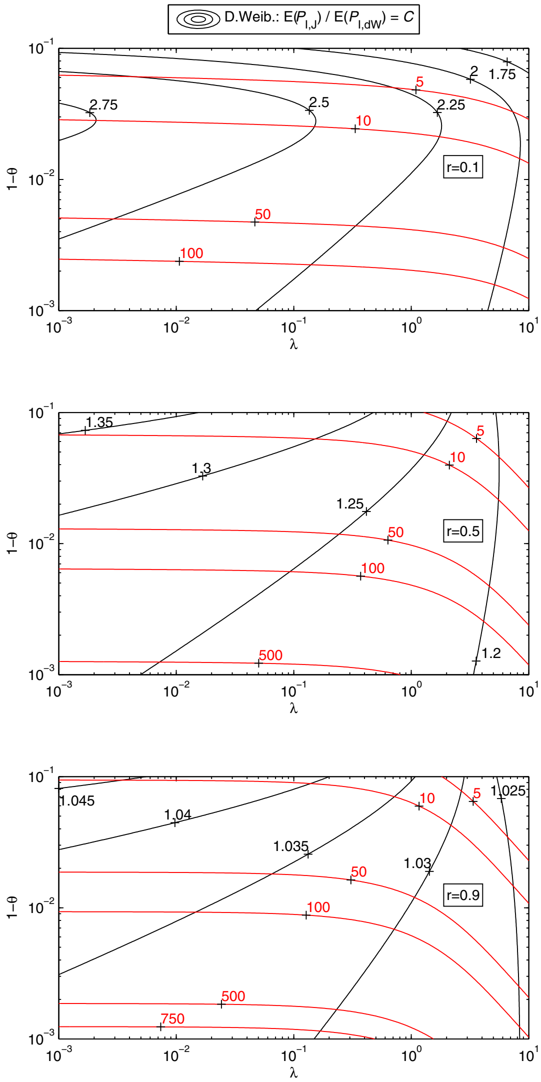

Fig. 8. Black curves are contours for the ratio in Equation (30), comparing the risk computed with the Jensen bound to that computed with the discrete Weibull distribution for high θ values. Red curves are contours for the ratio of risk computed with the Poisson distribution to that computed with the d.Wei. distribution. This figure is based on a constant $R = r$; Fig. S3 in the online appendix shows the case of beta-distributed R.

associated with treating the aggregation parameter as a constant.

For parameter fitting, maximum likelihood estimation was used. The likelihood function for this experimental setup is given by the product of binomial likelihood functions, where each factor corresponds to a certain dose level:

$$L(\omega) = \prod_{i=1}^{I} \binom{n_i}{w_i} \{\mathrm{E}(P_{\mathrm{I}})_i[\lambda_i, \omega]\}^{w_i} \{1 - \mathrm{E}(P_{\mathrm{I}})_i[\lambda_i, \omega]\}^{n_i - w_i}. \tag{31}$$

Here, ω is a parameter vector, I is the number of dose levels, λ_i, w_i, and n_i are the dose, number of positive (infected) subjects, and total number of subjects, respectively (at dose level indexed by i). $\mathrm{E}(P_{\mathrm{I}})_i[\lambda_i, \omega]$ is the dose-response model as a function of the mean dose and parameters to be fitted. Note that when constructing the likelihood function, we use different model formulations for data stemming from the use of aggregated and dispersed viruses, respectively (except in the case of the exact beta-Poisson model). Table II specifies which model formulation was used in each case. Thus, both the aggregated and dispersed data can be used simultaneously to estimate the parameters of the distribution of R, as well as the aggregation parameter. The maximum likelihood estimate of the unknown parameter vector ω is given by numerical optimization of Equation (31), which was performed in MATLAB.[40] The deviance, Y, also stated in Table III, is given by:

$$Y = -2 \ln\left(\frac{L(\omega)}{L^S}\right), \tag{32}$$

where L^S is the likelihood of the so-called saturated model:

$$L^S = \prod_{i=1}^{I} \binom{n_i}{w_i} \left(\frac{w_i}{n_i}\right)^{w_i} \left(1 - \frac{w_i}{n_i}\right)^{n_i - w_i}. \tag{33}$$

The p-value stated in Table III is for a chi-square goodness-of-fit test with the null hypothesis being

Table III. Parameter Estimates for the Dose-Response Models Fitted to Data in Table I

	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\theta}$	$\hat{\phi}$	$\mathrm{E}(\hat{R})$	Deviance	p-Value
Exact beta-Poisson	0.1103	29.55	$\equiv 0$	$\equiv 0$	3.719×10^{-3}	21.030	0.136
Beta-neg.bin.	8.128×10^{-3}	3.756×10^{-3}	0.999024	$\equiv 0$	0.6840	13.270	0.505
Fractional Poisson	–	–	0.999096	0.2775	0.7225	13.288	0.580
Jensen-beta	7.663×10^{-3}	3.504×10^{-3}	0.999045	$\equiv 0$	0.6862	13.273	0.505
Jensen-beta with imm.	2.478	2,186	0.993140	0.2756	8.200×10^{-4}	13.080	0.442

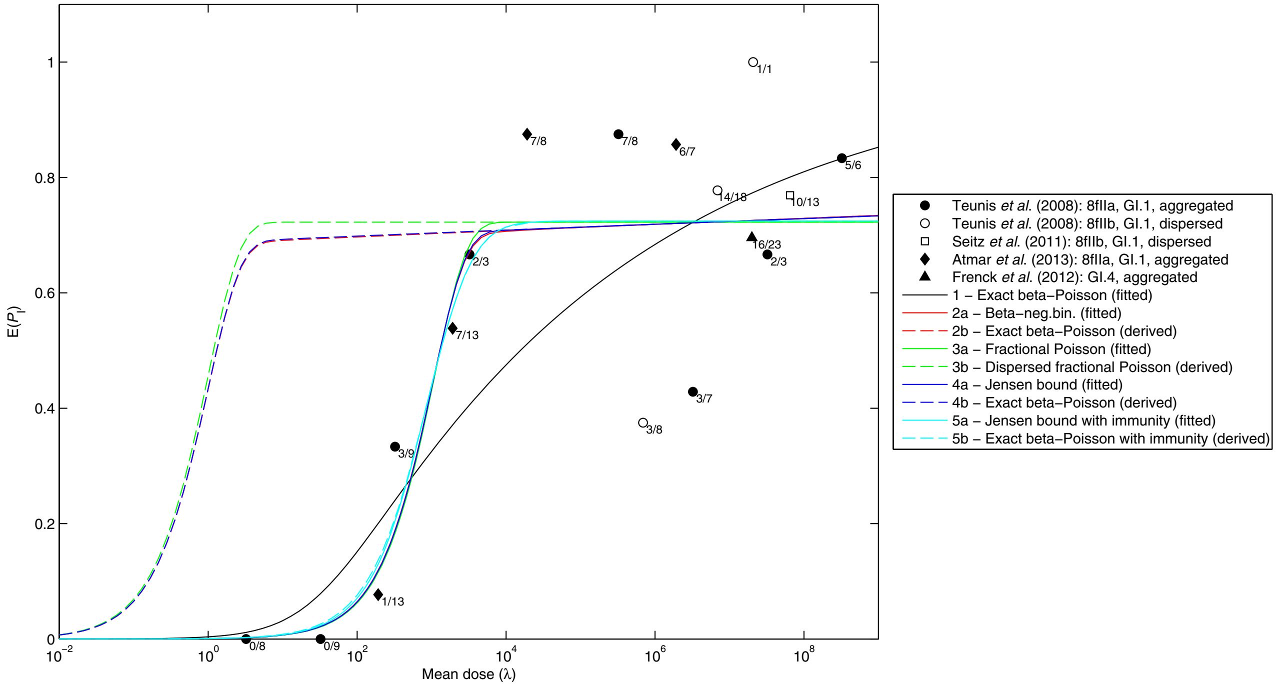

Fig. 9. Dose-response models (solid curves) in Table II fitted to the data in Table I. Dashed curves are derived from the respective fitted models by setting the value of the aggregation parameter to that corresponding to fully dispersed pathogens.

"acceptable fit" and the alternative hypothesis "lack of fit."

Table III and Fig. 9 give the results from parameter estimation. There are several points to note. First, the fitted models and their associated deviances are similar, except for the exact beta-Poisson model, which shows a somewhat poorer fit. Second, the fitted beta parameters of the beta-Jensen model (without immunity) are remarkably similar to those of the beta-negative binomial model. This result may not carry over to other cases, though, as the fitted beta distribution is quite extreme with almost all probability mass concentrated at 0 or 1.[17] Third, the mean single-hit probabilities of the exact beta-Poisson model and the beta-Jensen model with immunity deviate sharply from those in the three other models. Fourth, when eliminating the aggregation parameter from the fitted models (dashed curves in Fig. 9), the resulting dose-response curves are very different from their counterparts *with* the aggregation parameter, except for the beta-Jensen model with immunity, which almost does not change when compared with the exact beta-Poisson model (the corresponding dispersed model). This sensitivity to the inclusion/omission of an immunity parameter is consistent with what was reported by Schmidt.[18]

It should be noted, though, that the parameter estimates returned by the optimization routine for the beta-negative binomial model and the Jensen models seem quite sensitive to the initial guess that is supplied to the routine. The estimates reported here were obtained by maximizing the likelihood over a range of initial values until the routine delivered consistent results. Worryingly, there seems to exist a wide range of parameter sets, corresponding to a "ridge" or "plateau" in the likelihood surface, that gives approximately the same likelihood (i.e., changes in one parameter may be compensated by corresponding changes in (an)other parameter(s) without affecting the likelihood significantly). Similar challenges with nearly nonunique maximum likelihood estimates were reported by Messner *et al.*(17) when refitting the model used by Teunis *et al.*,(16) and by Schmidt(18) for several models incorporating aggregation. There is significant nonmonotonicity in the data, and there may not be enough information to fit three parameters (or even four, as for the beta-Jensen model with immunity) reliably.

6. DISCUSSION AND CONCLUDING REMARKS

In this work, we have argued that the stuttering Poisson distribution (Equation (10)) is a general and natural model for the dose distribution in the presence of pathogen clustering. By formulating the single-hit dose-response model in terms of a pgf, the stuttering Poisson leads to a simple expression for the dose-response model (Equation (16)). It was shown formally that the single-hit risk computed with a stuttering Poisson distribution is bounded from above by the risk computed with a Poisson distribution (Proposition 1) with the same mean. An equivalent result was obtained for mixed Poisson distributions (Proposition 2). We derived another risk bound (the Jensen bound; Proposition 3), valid for *any* dose distribution and *any* concave conditional dose-response model, which appears to approximate the single-hit risk quite closely for highly overdispersed dose distributions. This bound may also serve as an approximate dose-response model and its application to a real data set was demonstrated in Section 5.

Throughout this article, we have maintained the single-hit assumption of independently acting pathogens, even in the presence of pathogen clustering, as has been assumed in the published work on Norovirus clustered dose response.(16–18) This is a potentially unrealistic assumption that deserves some further attention in future work, although it may be challenging to test it experimentally with sufficient rigor. Propositions 1 and 2, as well as the Jensen bound, suggest that reduced risk from overdispersion (assuming equivalent mean doses) is a property that is fundamentally built into the single-hit framework, and as such is a theoretical prediction that can possibly be tested against data. In the remaining paragraphs, we make an attempt to discuss some potential practical implications of the theoretical results in the event that they actually *do* coincide with real-world effects. We should distinguish between *risk characterization* using an already calibrated dose-response model and *dose-response assessment* or parameter estimation in dose-response models.

The figures in Section 3.1 and in Section S.2 of the online appendix indicate that the effects of clustering in a single-hit model tend to be more pronounced at low doses, coinciding with typical background dose levels in most cases of drinking water risk characterization. However, the effects seem to become relevant only when there is pronounced clustering and when r is simultaneously large (or $\mathrm{E}(R)$ large and α small in the case of beta-distributed R). Therefore, it appears that moderate unaccounted for clustering in drinking water, as exemplified by the Hermite distribution, is unlikely to introduce much additional uncertainty or error into a single-hit risk characterization study, given that QMRA studies often have to quantify uncertainties by order-of-magnitude estimates. In the case of significant temporal variation in pathogen concentrations, periods/events of high doses may dominate the long-term mean risk. Since the theoretical effects of clustering become smaller at larger doses, it appears relatively unimportant to account for any physical clustering during these events.

Given the (likely) modest importance of accounting for clustering in single-hit drinking water studies, and the theoretical prediction that the risk computed using a Poisson distribution forms an upper bound for the risk computed using overdispersed distributions (stuttering or mixed Poisson), we have compelling arguments to direct our efforts at obtaining a correct mean dose rather than characterizing the dose distribution in greater detail. That is, *provided* that we can obtain a reliable estimate of the mean pathogen concentration (and, of course, the single-hit probability, or its distribution), using a Poisson distribution for the single-hit dose-response model during *risk characterization* will produce a higher (more conservative) mean risk estimate than using an overdispersed distribution with the same mean. Note, however, that it may be very difficult in practice to obtain a good estimate of the pathogen concentration, in particular for clustered suspensions or when temporal variation(12,35) is important, which may leave the risk estimate imprecise or even biased.

If one is interested in accounting for overdispersion, the Jensen bound in Equation (21) may prove useful. If a reliable estimate of both the mean concentration and the zero-inflation index (experimentally available from the proportion of zero counts) can be obtained, a significantly more precise single-hit risk estimate may be obtained for a situation with a highly overdispersed dose distribution without needing to consider further details of that dose distribution. For relatively dispersed suspensions, however, this bound may be more conservative than the risk obtained using a Poisson distribution, which means that the two should be compared before choosing which risk estimate to use.

When *fitting* a dose-response model to data, the implications of clustering are somewhat different. The risk computed with Poisson-distributed doses represents an upper bound on risk, so using the Poisson distribution when pathogens are, in fact, significantly clustered is likely to lead to an underestimation of the (mean) single-hit probability $\mathrm{E}(R)$, exemplified by the exact beta-Poisson parameters in Table III. This is because the parameters of the distribution for R will be chosen by the fitting procedure to compensate for the tendency toward increased risk enforced by the Poisson-distributed dose X. This problem may be somewhat counteracted by fitting the Jensen bound instead of a Poisson-based model, but only in those cases where the data allow reliable estimation of the additional parameter θ introduced in this model, which may represent a challenge.

The application of the Jensen bound in fitting dose-response data was illustrated in Section 5 with published data from human feeding trials on Norovirus. For this application, the Jensen bound model produced results that were very similar to the previously suggested beta-negative binomial model. However, like Schmidt,[18] we have some reservations regarding the possibility of reliably fitting three parameters to this data set. There appears to be a wide range of parameter values that gives roughly the same likelihood. Furthermore, Schmidt showed that the inclusion or omission of a host immunity parameter has a large effect on the results, which was also seen for the Jensen bound model in this work. Thus, there is still a need to obtain more dose-response data for Norovirus, preferably using nonaggregated viruses.

ACKNOWLEDGMENTS

The authors would like to thank four anonymous reviewers for their detailed and critical reviews.

APPENDIX

This appendix contains proofs of the three propositions that were presented in the main text. An overview of the contents of the online supplementary appendix can be found after the list of references.

A.1 Proofs of Propositions

Proposition 1 (Risk with stuttering Poisson doses). *Let the dose X be stuttering Poisson distributed with $\lambda_N > 0$ for some $N > 1$ (i.e., there exists some clusters) and fix the mean $\mathrm{E}(X) = \lambda = \sum_{n=1}^{\infty} n\lambda_n$. Then, the corresponding single-hit risk $\mathrm{E}(P_{\mathrm{I,sPo}})$ is bounded from above by $\mathrm{E}(P_{\mathrm{I,Po}})$, the single-hit risk computed using a Poisson distribution with the same mean λ.*

Proof. Consider the difference:

$$\mathrm{E}(P_{\mathrm{I,Po}}) - \mathrm{E}(P_{\mathrm{I,sPo}}) = \int_0^1 [G_X(1-r) - e^{-\lambda r}] f_R(r)\,\mathrm{d}r, \tag{A.1}$$

where we used the general expression of Equation (2) for $\mathrm{E}(P_{\mathrm{I,sPo}})$ and Equation (5) for $\mathrm{E}(P_{\mathrm{I,Po}})$. We need to show that Equation (A.1) is positive when X is stuttering Poisson. Since $f_R \geq 0$ for all $r \in [0, 1]$ and $f_R > 0$ on some subset of the unit interval, it will suffice to take $r > 0$ and show the positivity of the remaining factor in the integrand, ΔG:

$$\begin{aligned}\Delta G &= G_X(1-r) - e^{-\lambda r}\\ &= \left[G_X(1-r)e^{\lambda r} - 1\right] e^{-\lambda r}\\ &= \frac{\exp\left\{\sum_{n=1}^{\infty} \lambda_n \left[(1-r)^n - (1-nr)\right]\right\} - 1}{e^{\lambda r}}.\end{aligned} \tag{A.2}$$

Here, we used the pgf of a stuttering Poisson distribution (Equation (14)) and the identity (by assumption) $\lambda = \sum_{n=1}^{\infty} n\lambda_n$. We now show that the numerator in Equation (A.2) is positive. Let $h(n) = (1-r)^n - (1-nr)$. We have $h(1) = 0$ and for $n \geq 1$, we have the difference:

$$h(n+1) - h(n) = r[1-(1-r)^n] > 0, \tag{A.3}$$

since $(1-r)^n < 1$ for $n \geq 1$. By mathematical induction, $h(n) > 0$ for $n \geq 2$. Since there exists some $N > 1$ such that $\lambda_N > 0$, we have $\exp[\sum_{n=1}^{\infty} \lambda_n h(n)] > 1$, which shows that $\Delta G > 0$. This proves the proposition. □

Proposition 2 (Risk with mixed Poisson doses). *Let the dose X be mixed Poisson distributed with mixing distribution $f_\Lambda(\lambda)$ and pmf given by Equation* (19). *Then, the corresponding single-hit risk* $\mathrm{E}(P_{\mathrm{I,mPo}})$ *is bounded from above by* $\mathrm{E}(P_{\mathrm{I,Po}})$*, the single-hit risk computed using a Poisson distribution with mean equal to the mean of the mixing distribution,* $\mathrm{E}(\Lambda)$.

Proof. We need the pgf of X, which is given by:

$$\begin{aligned} G_X(z) &= \sum_{x=0}^{\infty} z^x \int_0^\infty p_{X_{\mathrm{Po}}}(x)\, f_\Lambda(\lambda)\,\mathrm{d}\lambda \\ &= \int_0^\infty \sum_{x=0}^{\infty} z^x p_{X_{\mathrm{Po}}}(x)\, f_\Lambda(\lambda)\,\mathrm{d}\lambda \qquad \text{(A.4)} \\ &= \int_0^\infty e^{\lambda(z-1)}\, f_\Lambda(\lambda)\,\mathrm{d}\lambda, \end{aligned}$$

where we assumed that we may interchange integration and summation. Inserting Equation (A.4) in the general single-hit expression (Equation (2)), we get:

$$\mathrm{E}(P_{\mathrm{I,mPo}}) = 1 - \int_0^1 \int_0^\infty e^{-\lambda r}\, f_\Lambda(\lambda)\,\mathrm{d}\lambda f_R(r)\,\mathrm{d}r. \tag*{(A.4)}$$

Since $e^{-\lambda r}$ is a strictly convex function of λ on $[0, \infty)$, we may use *Jensen's inequality* to conclude that:

$$\int_0^\infty e^{-\lambda r}\, f_\Lambda(\lambda)\,\mathrm{d}\lambda > e^{-\mathrm{E}(\Lambda) r}. \tag*{(A.6)}$$

This leads to:

$$\mathrm{E}(P_{\mathrm{I,mPo}}) < 1 - \int_0^1 e^{-\mathrm{E}(\Lambda) r}\, f_R(r)\,\mathrm{d}r, \tag*{(A.7)}$$

where the rhs. is recognized as $\mathrm{E}(P_{\mathrm{I,Po}})$, the single-hit risk computed with a Poisson distribution with mean $\mathrm{E}(\Lambda)$, which concludes the proof. □

Proposition 3 (The Jensen bound). *Introduce the notation* $P_{\mathrm{I}}^0(x) \equiv \mathrm{E}_R(P_{\mathrm{I}})\big|_{X=x}$ *for a general concave (in x) conditional dose-response model. Then, the risk* $\mathrm{E}(P_{\mathrm{I}}) = \sum_{x=0}^{\infty} p_X(x) P_{\mathrm{I}}^0(x)$ *is bounded from above by:*

$$\begin{aligned} \mathrm{E}(P_{\mathrm{I,J}}) &= [1 - p_X(0)] \cdot P_{\mathrm{I}}^0\left(\frac{\lambda}{1 - p_X(0)}\right) \\ &= \left(1 - e^{\lambda(\theta-1)}\right) \cdot P_{\mathrm{I}}^0\left(\frac{\lambda}{1 - e^{\lambda(\theta-1)}}\right), \end{aligned} \tag*{21}$$

where θ is the zero-inflation index of the distribution of X.

Proof. We need Jensen's inequality in the following form. Let ϕ be a concave function on $[0, \infty)$, u_x points in the domain of ϕ and $w_x \geq 0$ be weights such that $\sum w_x u_x < \infty$. Then, Jensen's inequality states (possibly involving infinite sums):

$$\frac{\sum w_x \phi(u_x)}{\sum w_x} \leq \phi\left(\frac{\sum w_x u_x}{\sum w_x}\right). \tag*{(A.8)}$$

Make the identifications $u_x = x$, $w_x = p_X(x)$, and $\phi(u_x) = \phi(x) = P_{\mathrm{I}}^0(x)$. Summing from $x = 1$ to infinity, inequality (A.8) becomes:

$$\frac{\sum_{x=1}^{\infty} p_X(x) P_{\mathrm{I}}^0(x)}{\sum_{x=1}^{\infty} p_X(x)} \leq P_{\mathrm{I}}^0\left(\frac{\sum_{x=1}^{\infty} x p_X(x)}{\sum_{x=1}^{\infty} p_X(x)}\right). \tag*{(A.9)}$$

Using $P_{\mathrm{I}}^0(0) = 0$, we thus have:

$$\begin{aligned} \mathrm{E}(P_{\mathrm{I}}) &= \sum_{x=0}^{\infty} p_X(x) P_{\mathrm{I}}^0(x) = \sum_{x=1}^{\infty} p_X(x) P_{\mathrm{I}}^0(x) \\ &\leq \left(\sum_{x=1}^{\infty} p_X(x)\right) \cdot P_{\mathrm{I}}^0\left(\frac{\sum_{x=1}^{\infty} x p_X(x)}{\sum_{x=1}^{\infty} p_X(x)}\right) \\ &= [1 - p_X(0)] \cdot P_{\mathrm{I}}^0\left(\frac{\lambda}{1 - p_X(0)}\right) \\ &= \left(1 - e^{\lambda(\theta-1)}\right) \cdot P_{\mathrm{I}}^0\left(\frac{\lambda}{1 - e^{\lambda(\theta-1)}}\right) = \mathrm{E}(P_{\mathrm{I,J}}), \end{aligned} \tag*{(A.10)}$$

where we used Equation (7) to introduce the zero-inflation index. □

REFERENCES

1. Elimelech M, Gregory J, Jia X, Williams R. Particle Deposition and Aggregation: Measurement, Modelling and Simulation. Oxford: Butterworth-Heinemann, 1995.
2. Gale P. Developments in microbiological risk assessment models for drinking water—A short review. Journal of Applied Microbiology, 1996; 81(4):403–410.
3. Gale P. Developments in microbiological risk assessment for drinking water. Journal of Applied Microbiology, 2001; 91(2):191–205.
4. Haas CN, Rose JB, Gerba CP. Quantitative Microbial Risk Assessment. New York: John Wiley & Sons, 1999.
5. Haas CN. Estimation of risk due to low doses of microorganisms: A comparison of alternative methodologies. American Journal of Epidemiology, 1983; 118(4):573–582.
6. Teunis P, Havelaar A. The beta Poisson dose-response model is not a single-hit model. Risk Analysis, 2000; 20(4):513–520.
7. Haas CN. Conditional dose-response relationships for microorganisms: Development and application. Risk Analysis, 2002; 22(3):455–463.
8. Schmidt PJ, Pintar KDM, Fazil AM, Topp E. Harnessing the theoretical foundations of the exponential and beta-Poisson dose-response models to quantify parameter uncertainty using Markov chain Monte Carlo. Risk Analysis, 2013; 33(9):1677–1693,.

9. Nilsen V, Wyller J. QMRA for drinking water: 1. Revisiting the mathematical structure of single-hit dose-response models. Risk Analysis, 2016; 36(1):145–162.
10. Gale P, Dijk Pv, Stanfield G. Drinking water treatment increases micro-organism clustering; the implications for microbiological risk assessment. Journal of Water Supply: Research and Technology-Aqua, 1997; 46(3):117–126.
11. Gale P, Pitchers R, Gray P. The effect of drinking water treatment on the spatial heterogeneity of micro-organisms: Implications for assessment of treatment efficiency and health risk. Water Research, 2002; 36(6):1640–1648.
12. Englehardt JD, Li R. The discrete Weibull distribution: An alternative for correlated counts with confirmation for microbial counts in water. Risk Analysis, 2011; 31(3):370–381.
13. Englehardt JD, Ashbolt NJ, Loewenstine C, Gadzinski ER, Ayenu-Prah AY. Methods for assessing long-term mean pathogen count in drinking water and risk management implications. Journal of Water and Health, 2012; 10(2):197–208.
14. Grant SB. Virus coagulation in aqueous environments. Environmental Science & Technology, 1994; 28(5):928–933.
15. Teunis P, Lodder W, Heisterkamp S, de Roda Husman A. Mixed plaques: Statistical evidence how plaque assays may underestimate virus concentrations. Water Research, 2005; 39(17):4240–4250.
16. Teunis PF, Moe CL, Liu P, Miller SE, Lindesmith L, Baric RS, Le Pendu J, Calderon RL. Norwalk virus: How infectious is it? Journal of Medical Virology, 2008; 80(8):1468–1476.
17. Messner MJ, Berger P, Nappier SP. Fractional Poisson—A simple dose-response model for human norovirus. Risk Analysis, 2014; 34(10):1820–1829.
18. Schmidt PJ. Norovirus dose–response: Are currently available data informative enough to determine how susceptible humans are to infection from a single virus? Risk Analysis, 2014; 35(7):1364–1383.
19. Thurston-Enriquez JA, Haas CN, Jacangelo J, Gerba CP. Chlorine inactivation of adenovirus type 40 and feline calicivirus. Applied and Environmental Microbiology, 2003; 69(7):3979–3985.
20. Johnson NL, Kemp AW, Kotz S. Univariate Discrete Distributions, 3rd ed. Hoboken, NJ: John Wiley & Sons, 2005.
21. Adelson R. Compound poisson distributions. OR, 1966; 17(1):73–75.
22. Kemp C. "Stuttering-Poisson" distributions. Journal of the Statistical and Social Inquiry Society of Ireland, 1967; 21(5):151–157.
23. Puig P, Barquinero JF. An application of compound Poisson modelling to biological dosimetry. Proceedings of the Royal Society A, 2011; 467(2127):897–910.
24. Milne RK, Westcott M. Generalized multivariate Hermite distributions and related point processes. Annals of the Institute of Statistical Mathematics, 1993; 45(2):367–381.
25. Kemp C, Kemp AW. Some properties of the "Hermite" distribution. Biometrika, 1965; 52(3–4):381–394.
26. Kemp AW, Kemp C. An alternative derivation of the Hermite distribution. Biometrika, 1966; 53(3–4):627–628.
27. McKendrick A. Applications of mathematics to medical problems. Proceedings of the Edinburgh Mathematical Society, 1926; 44:98–130.
28. Gupta R, Jain G. A generalized Hermite distribution and its properties. SIAM Journal on Applied Mathematics, 1974; 27(2):359–363.
29. Feller W. An Introduction to Probability Theory and its Applications, vol. 2, 2nd ed. New York: John Wiley & Sons, 1971.
30. Zhang H, Liu Y, Li B. Notes on discrete compound poisson model with applications to risk theory. Insurance: Mathematics and Economics, 2014; 59(1):325–336.
31. Hadar J, Russell WR. Rules for ordering uncertain prospects. American Economic Review, 1969; 25–34.
32. Bawa VS. Optimal rules for ordering uncertain prospects. Journal of Financial Economics, 1975; 2(1):95–121.
33. Rothschild M, Stiglitz JE. Increasing risk: I. A definition. Journal of Economic Theory, 1970; 2(3):225–243.
34. Nakagawa T, Osaki S. The discrete Weibull distribution. IEEE Transactions on Reliability, 1975; 5:300–301.
35. Englehardt J, Swartout J, Loewenstine C. A new theoretical discrete growth distribution with verification for microbial counts in water. Risk Analysis, 2009; 29(6):841–856.
36. Seitz SR, Leon JS, Schwab KJ, Lyon GM, Dowd M, McDaniels M, Abdulhafid G, Fernandez ML, Lindesmith LC, Baric RS, Moe CL. Norovirus infectivity in humans and persistence in water. Applied and Environmental Microbiology, 2011; 77(19):6884–6888.
37. Atmar RL, Opekun AR, Gilger MA, Estes MK, Crawford SE, Neill FH, Ramani S, Hill H, Ferreira J, Graham DY. Determination of the 50 norwalk virus. Journal of Infectious Diseases, 2014; 209(7):1016–1022. Available at: http://jid.oxfordjournals.org/content/209/7/1016.abstract.
38. Frenck R, Bernstein DI, Xia M, Huang P, Zhong W, Parker S, Dickey M, McNeal M, Jiang X. Predicting susceptibility to norovirus gii. 4 by use of a challenge model involving humans. Journal of Infectious Diseases, 2012; 206(9):1386–1393.
39. Jones MK, Watanabe M, Zhu S, Graves CL, Keyes LR, Grau KR, Gonzalez-Hernandez MB, Iovine NM, Wobus CE, VinjÃ© J, Tibbetts SA, Wallet SM, Karst SM. Enteric bacteria promote human and mouse norovirus infection of b cells. Science, 2014; 346(6210):755–759. Available at: http://www.sciencemag.org/content/346/6210/755.abstract.
40. MATLAB. Version 8.1.0.604 (R2013a). Natick, MA: MathWorks Inc., 2013.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's website:

S.1. A compact review of the mathematical concepts, in particular probability generating functions, needed to fully understand the main article.

S.2. A collection of numerical examples to show how clustering may affect single-hit models when R is beta distributed, and how the Jensen bound performs in these examples.

S.3. A parallel to Propositions 1 and 2 for binomially distributed clusters.

S.4. A simple example to show that the conclusion from Proposition 1 (reduced risk from clustering) fails if the conditional dose-response model has a convex portion in the low-dose range, as in the 2-hit model.

S.5. A primitive generalization of the single-hit concept to account for the effects discussed in bullet point 3 in the introduction of the main article, i.e., if the host-pathogen interaction for each pathogen depends on that pathogen being part of a cluster or not.

S.6. A section to show how a single-hit risk estimate may be affected by misinterpreting clusters as single pathogen during enumeration, as discussed in bullet point 1 in the introduction to the main article.