EPA/600/R-08/033
March 2008

# Development and Evaluation of Novel Dose-Response Models for Use in Microbial Risk Assessment

National Center for Environmental Assessment
Office of Research and Development
U.S. Environmental Protection Agency
Cincinnati, OH 45268

## NOTICE

The U.S. Environmental Protection Agency through its Office of Research and Development partially funded and collaborated on the research described here under contract no. CI0790101N with the Intergovernmental Personnel Act Agreement between U.S. EPA and Miami University and contract no. DW89921719-01 with the Interagency Agreement between the U.S. EPA and the Department of Energy (ORISE). It has been subjected to the Agency's peer and administrative review and has been approved for publication as an EPA document. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

## ABSTRACT

Dose-response relationships relating population infection and illness responses to drinking water dose are important to support development of drinking water and wastewater reuse policy. Illness endpoints, in particular, are of primary importance in economic benefits assessment and management of risk. Dose-response relationships for both endpoints, therefore, are needed for assessing the full extent of disease burden attributable to pathogens in drinking water and to evaluate the need for new regulation. The purpose of this document is to present the predictive Bayesian framework as an alternative to the current methods for expressing the risk of infection and illness resulting from exposure to pathogens in drinking water. Secondarily, an alternative non-Poisson approach for characterizing the exposure distribution at the tap is offered in the context of the dose-response function. Together, these new methods may provide a more realistic and rigorous depiction of the impact of water-borne pathogens on drinking water consumers.

**Preferred citation:**
U.S. EPA. 2008. Development and Evaluation of Novel Dose-Response Models for Use in Microbial Risk Assessment. U.S. Environmental Protection Agency, National Center for Environmental Assessment, Cincinnati, OH. EPA/600/R-08/033.

## TABLE OF CONTENTS

Page

LIST OF ABBREVIATIONS ..... iv
PREFACE ..... v
AUTHORS, CONTRIBUTORS AND REVIEWERS ..... vi

1. INTRODUCTION: SCOPE, PURPOSE AND OBJECTIVES ..... 1

   1.1. THE PREDICTIVE BAYESIAN FRAMEWORK FOR MICROBIAL RISK ASSESSMENT ..... 1
   1.2. A NON-POISSON DISTRIBUTION AND RELATED DOSE-RESPONSE FUNCTION FOR PATHOGENS IN TREATED DRINKING WATER ..... 4
   1.3. THE DISCRETE-SCALING DOSE-RESPONSE FUNCTION ..... 4

2. METHODS ..... 6

   2.1. PREDICTIVE BAYESIAN METHODS ..... 6
   2.2. DISCRETE-SCALING DISTRIBUTION ..... 7
   2.3. BETA DISCRETE-SCALING DISTRIBUTION DOSE-RESPONSE FUNCTION (beta-DSD) ..... 7

3. SUMMARY OF RESULTS ..... 8

   3.1. PREDICTIVE BAYESIAN DOSE-RESPONSE ASSESSMENT FOR APPRAISING ABSOLUTE HEALTH RISK FROM AVAILABLE INFORMATION ..... 10
   3.2. PREDICTIVE POPULATION DOSE-RESPONSE ASSESSMENT FOR CRYPTOSPORIDIUM PARVUM: INFECTION ENDPOINT ..... 11
   3.3. PREDICTIVE BAYESIAN MICROBIAL DOSE-RESPONSE ASSESSMENT BASED ON SUGGESTED SELF-ORGANIZATION IN PRIMARY ILLNESS RESPONSE: *CRYPTOSPORIDIUM PARVUM* ..... 11
   3.4. A NEW THEORETICAL DISCRETE “SCALING” PROBABILITY DISTRIBUTION WITH VERIFICATION FOR MICROBIAL COUNTS IN WATER ..... 12
   3.5. A DOSE-RESPONSE FUNCTION FOR DISCRETE-SCALING DISTRIBUTED PATHOGENS SUCH AS FOUND IN DRINKING WATER ..... 12

4. DISCUSSION AND CONCLUSIONS ..... 14
5. RESEARCH NEEDS ..... 19
6. REFERENCES ..... 21

## LIST OF ABBREVIATIONS

| | |
|---|---|
| CCL | Contaminant Candidate List |
| DSD | discrete-scaling distribution |
| PLN | Poisson lognormal |
| SOC | self-organized criticality |
| U.S. EPA | U.S. Environmental Protection Agency |

## PREFACE

This document was prepared by the National Center for Environmental Assessment for the Office of Ground Water and Drinking Water. The document is a summary and synthesis of five publication manuscripts authored by Jeff Swartout in collaboration with Jim Englehardt of the University of Miami. The document contains a description of dose-response modeling methods designed to provide a robust approach under uncertainty for predicting human population risk from exposure to pathogens in drinking water. There was no literature search specific to this document, only in conjunction with the publication manuscripts, three of which have been published in the peer-reviewed literature and two that have been submitted for publication as of May 7, 2007.

## AUTHORS, CONTRIBUTORS AND REVIEWERS

### AUTHORS

James D. Englehardt, Ph.D., P.E.
ORISE Fellow
University of Miami

Jeff Swartout
U.S. Environmental Protection Agency
Office of Research and Development
National Center for Environment Assessment
Cincinnati, OH 45268

### CONTRIBUTORS

Chad Loewenstine
U.S. Environmental Protection Agency
Office of Research and Development
National Center for Environment Assessment
Cincinnati, OH 45268

### REVIEWERS

Glenn Suter II, Ph.D.
U.S. Environmental Protection Agency
Office of Research and Development
National Center for Environment Assessment
Cincinnati, OH 45268

Nicholas J. Ashbolt, Ph.D.
U.S. Environmental Protection Agency
Office of Research and Development
National Exposure Research Laboratory
Cincinnati, OH 45268

# 1. INTRODUCTION: SCOPE, PURPOSE AND OBJECTIVES

Dose-response relationships relating population infection and illness responses to exposure to pathogens in drinking water are needed for the quantitative assessment of pathogen risks used in developing drinking water and wastewater reuse policy. Infection response is of importance in assessing secondary disease transmission and population-based pathogen risks (so-called dynamic risk models; Eisenberg et al., 2005). Illness endpoints have typically been used to indicate the health burden, of primary importance in economic benefits assessment and management of risk. Dose-response relationships for both endpoints, therefore, are needed for assessing the full extent of disease burden attributable to pathogens in drinking water and to evaluate the need for new regulation.

The purpose of this document is to describe a body of literature on a predictive (unconditional) Bayesian framework as an alternative to the currently used approach (variations on ILSI, 2000) to express the risk of infection and illness resulting from exposure to pathogens in drinking water. Secondarily, an alternative to the Poisson distribution for characterizing pathogens in tap water is also described. Together, these new methods may provide a more realistic and rigorous depiction of the impact of water-borne pathogens on drinking water consumers.

The objectives of this work are to:

(1) Present the unconditional (predictive) Bayesian forms of infection and illness dose-response functions as an alternative to confidence bounds for the assessment of drinking water pathogen risks;

(2) Derive a (discrete-scaling) form for the distribution of microbe counts in drinking water at the point of entry to the distribution system, based on a theoretical treatment process failure structure; and

(3) Derive general infection and illness dose-response functions for pathogens in drinking water and other media based on the new discrete-scaling count distribution.

## 1.1. THE PREDICTIVE BAYESIAN FRAMEWORK FOR MICROBIAL RISK ASSESSMENT

Pathogen dose-response data are fundamentally sparse and noisy due to limitations in the numbers of human subjects that can be tested, and a range of uncertainties resulting from defining infection and illness and pathogen strain variations (e.g., Teunis et al., 2002). To address limitations in empirical verification, confidence-

based approaches such as likelihood profiles, parametric bootstrapping and the Benchmark Dose have been used. Traditional confidence-based approaches based on frequentist methods, however, are not ideal for predictive modeling, as is needed for microbial risk assessment. Confidence bounds are dependent—or conditional—on a specific confidence level (e.g., 5% or 10%), such that the relative ranking of two health stressors often reverses, given another level of confidence (Englehardt, 2004). In addition, at a 95% level, the assessed bound may be far from the assessed mean risk and potentially subject to large errors. Therefore, confidence bounds are not generally comparable among health stressors.

Bayesian approaches have been replacing frequentist methods within predictive models, partially to overcome such limitations and partially from recognition that probability judgements are fundamentally subjective. In essence, Bayesian methods utilize more information in a more transparent manner than do frequentist methods. Bayesian methods are often confidence-based, but generally characterized by wider confidence bounds than the corresponding frequentist approach. In contrast, the predictive Bayesian method yields an unconditional result—one that is not dependent on a specific choice of confidence bounds. In the predictive Bayesian approach, the uncertainty represented in the confidence limits is integrated into the output, resulting in a more robust prediction. Figure 1 illustrates the results of an unconditional predictive Bayesian dose-response function compared to one (beta-Poisson) based on frequentist confidence limits. As the figure shows, the unconditional beta-Poisson response is slightly higher than the frequentist response determined by the direct fit of the beta-Poisson dose-response function to the data. For perfect information, the two curves would be the same. Otherwise, the unconditional response is more health protective (conservative), but still represents the expected risk, with uncertainty incorporated directly. Risk management decisions are often “quantified” largely on the expected value and are only “qualified” by the confidence bounds. That is, the key quantitative risk estimates and corresponding benefits are based on the central tendency, not the extremes. The unconditional risk estimate provided by the predictive Bayesian approach allows for a more conservative estimate of the central tendency when information is not perfect.

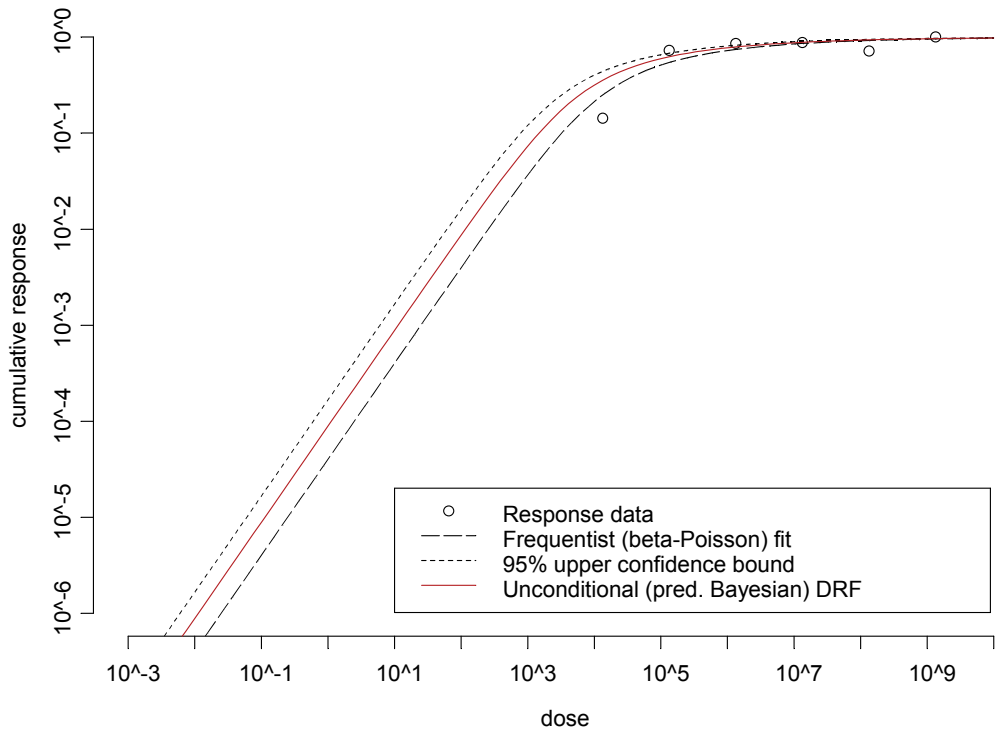



FIGURE 1

Comparison of an Unconditional Predictive-Bayesian Dose-Response Function (Englehardt, 2004) with Frequentist Confidence Bound Approach

An analogous unconditional risk can be assessed using frequentist methods, though efficient computational methods have been lacking and the approach has not found use. More important is that the frequentist approach precludes the use of any information other than numerical dose-response data, such as data on genetic prevalence that is rapidly finding application in dose-response assessment, and epidemiological information, which can substantially increase the knowledge base for a dose-response assessment. Equally important, the predictive Bayesian approach computed using Markov chain Monte Carlo methods developed over the last decade, gives the frequentist answer directly, if no non-traditional dose-response information is used in the analysis.

## 1.2. A NON-POISSON DISTRIBUTION AND RELATED DOSE-RESPONSE FUNCTION FOR PATHOGENS IN TREATED DRINKING WATER

The second objective of this work was to show how the distribution of pathogens in finished drinking water is likely to deviate from the Poisson distribution, and illustrate how that deviation will effect the dose-response function for consumers. The form of the function relating dose of pathogens in drinking water to illness response depends not only on the numerous characteristics of gastrointestinal pathogenesis, but also on the distribution of pathogens in the water. The pathogen distribution in finished drinking water over time is highly disperse, unlike the well-mixed uniformly-distributed (Poisson) laboratory sample used in controlled dose-response studies. For the low concentrations of pathogens typically found in drinking water, where the expected exposure at the tap is, at most, one organism at a time, a more disperse distribution (i.e., non-Poisson) will result in a reduction of the risk of infection in the population, compared to a Poisson distribution of pathogens, for any given long-term mean concentration. The discrete-scaling distribution (DSD; Englehardt et al., 2008b) was developed to account for this greater dispersion of microbes expected in finished drinking water. The DSD was based on a theoretical consideration of dependent multiplicative failure processes in drinking water treatment plants, as opposed to the conventional assumption of independent processes, resulting in lognormal means. The DSD provided the best fit to several finished and raw water data sets, the common characteristics being a large number of zeroes (non-detects) and low counts for positive samples. Conversely, data sets characterized by few or no zeroes and high positive sample counts was fitted better by the Poisson-lognormal distribution. Englehardt et al. (2008b) recommend the use of the DSD for any set of water samples where the number of zero-count samples is greater than the number of any other unique count samples. That is, as the DSD is a strictly decreasing probability mass function, the highest probability for any single count value resides at zero.

## 1.3. THE DISCRETE-SCALING DOSE-RESPONSE FUNCTION

In combining the results meeting the first two objectives, a new dose-response function was derived mathematically based on the new DSD count distribution. For the corresponding dose-response assessment of infection or illness risk, the DSD was integrated into a dose-response function by replacing the Poisson exposure component of the beta-Poisson function with the DSD. The beta-DSD dose-response function should be used whenever the DSD is the preferred distribution fit to water pathogen count data for a given site assessment. For over-dispersed (non-uniform) pathogen

distributions, compared to the DSD, the beta-Poisson dose-response function could overestimate the risk substantially, depending on the magnitude of the measured pathogen concentrations.

# 2. METHODS

The mathematical methods presented in the five associated papers (described in Section 3) are summarized here in two parts. Section 2.1 describes the predictive Bayesian approach, which is used in conjunction with both the beta-Poisson and beta-discrete dose-response functions. Section 2.2 describes the discrete-scaling distribution for modeling the occurrence of pathogens in finished drinking water. The DSD, in turn, is the basis for the exposure component of the beta-discrete dose-response function.

## 2.1. PREDICTIVE BAYESIAN METHODS

To address the nature of low-dose response assessment, an unconditional probability of illness, or other response, can be obtained as the conditional dose-response function (sampling distribution) multiplied by the distribution of uncertainty in model parameters, integrated over the full range of variability and uncertainty in the parameters (Englehardt, 2004). The result is a distribution for the quantity of interest that is quantitatively narrower as more information becomes available, consistent with principles of Shannon entropy (Shannon, 1948). Unconditional Bayesian distributions have been termed *predictive* distributions (Aitchison and Dunsmore, 1975). A predictive Bayesian probability distribution integrates all available information, as well as the full range of possible confidence levels. In addition, because the predictive distribution is sensitive to the assumed parametric sampling distribution, the method is powerful for the exploitation of information on the expected form of the distribution obtained by theoretical derivation. In the general case, the distribution of variability and uncertainty can be a Bayesian posterior, incorporating non-traditional information (e.g., genetic or epidemiological) as well as numerical dose-response data. The approach produces an unconditional probability distribution that is wider and more conservative when less information is available. The unconditional distribution is sensitive to, and exploitive of, knowledge of the form of the dose-response function.

For both the beta-Poisson (Englehardt and Swartout, 2006) and beta-discrete illness functions (Englehardt et al., 2008a), knowledge of the form of the function was obtained from simple models of the pathogenic process in the gut. The models were based on the principle of self-organized criticality (SOC) (Bak et al., 1988). The original model of species extinction proposed by Bak and coworker represents the process by which natural selection can drive evolutionary self-organization in stochastic systems, such that general patterns of outcomes emerge from specific, complex, small-scale

interactions. Although there is no direct practical utility of such models applied to the management of microbial risks, the approach was useful for investigation of integrative dose-response characteristics that may emerge from complex, adaptive host-pathogen interactions (Englehardt and Swartout, 2006; Englehardt et al., 2008a).

## 2.2. DISCRETE-SCALING DISTRIBUTION

The DSD was developed for assessing the occurrence-distribution of pathogens in finished drinking water, specifically for the case of dependent multiplicative-error processes in water treatment plants, where the failure of one process is, at least, partially dependent on the failure of the immediately preceding process (Englehardt, 1995; Englehardt et al., 2008b). The DSD can be generalized to cases of partially-dependent and independent causes. That is, the result of Frisch and Sornette (1997), showing the Weibull tail for incident sizes arising as the product of independent exponential cause sizes, was generalized by Englehardt et al. (2008b) to the case of discrete geometric cause sizes. The DSD is meant to be used for fitting to microbial count data from drinking water samples taken over time. Englehardt et al. (2008b) provided fits for the DSD to several example data sets.

## 2.3. BETA DISCRETE-SCALING DISTRIBUTION DOSE-RESPONSE FUNCTION (beta-DSD)

The beta-DSD integrates the beta distribution for inter-individual variability in response with the DSD as the exposure component. The beta-DSD is characterized by three parameters. Two of the parameters are identical to the beta parameters of the beta-Poisson, representing the distribution of response variability in the human population. The third parameter is the characteristic exponent, $\eta$, from the DSD function itself. The functional form is somewhat complex and is not reproduced here. The reader is referred to Englehardt et al. (2008a) for the technical details. The beta-DSD is linear at very low doses (dilutions of single organisms), but approaches linearity more slowly than the beta-Poisson because of the higher probability of exposure to more than one organism at a time. However, as explained previously, the beta-DSD predicts lower risk at low doses prior to becoming linear with respect to the beta-Poisson. Conversely, the beta-DSD predicts higher risk than the beta-Poisson at high doses.

## 3. SUMMARY OF RESULTS

The principal results of this work are as follows.

1. Englehardt (2004) and Englehardt and Swartout (2004, 2006) demonstrate that the relative importance of two health stressors may depend on the chosen level of confidence. In other words, because the confidence-bound approach is based on an isolated, somewhat arbitrary level of confidence, to the exclusion of the remaining distribution of parameter uncertainty, stressors may vary with respect to their relative importance.

2. A predictive Bayesian pathogen dose-response function that expresses the unconditional probability of infection was developed (Englehardt, 2004). The function is rigorously conservative and consistent among health stressors. The function is general for Poisson-distributed pathogens as found in laboratory samples. An unconditional assessment was presented based on available data for rotavirus.

3. An unconditional dose-response assessment, based on available infection data for *C. parvum*, was developed and published (Englehardt and Swartout, 2004).

4. A new general pathogen dose-response function for the illness endpoint was derived based on a model of self-organized pathogenesis in the human gastrointestinal tract and published (Englehardt and Swartout, 2006). The conditional function derived is a generalization of the beta-Poisson model, and was presented as a predictive Bayesian, unconditional model and demonstrated for *C. parvum*.

5. The assumption of Poisson-distributed pathogens underlying current dose-response assessments was shown to be inappropriate for environmental samples such as drinking water and source water (Englehardt et al., 2008a,b). A new *discrete scaling distribution* (DSD) for pathogen counts in drinking water was derived from theoretical principles and shown to describe well the distribution of pathogens in water samples characterized by a high number of zero-count observations (Englehardt et al., 2008b). The DSD is expected to accurately model distributions of pathogens in any media, including food and contaminated buildings, in which low pathogen counts are observed. The existing (Poisson lognormal) distribution was preferred for pathogen counts in environmental samples where higher counts (few zero counts) are observed.

6. A new dose-response function based on the new count distribution was derived for pathogens in environmental samples such as drinking water, for the infection and illness endpoints, not accounting for secondary transmission of infection, as shown in Figure 2 (Englehardt et al., 2008b). The predictive Bayesian version of the conditional function derived was presented and demonstrated for *C. parvum.*

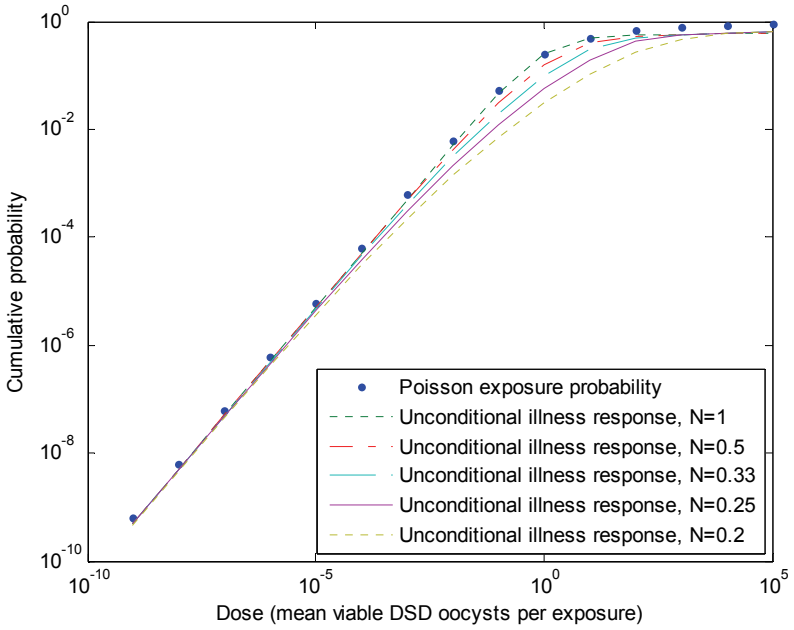



FIGURE 2

Predictive Bayesian Beta-Discrete Scaling Illness Dose-Response Function for *C. parvum* (all isolates) for the Inverse Number of Unit Processes $\eta$ = 1, 0.5, 0.33, 0.25, 0.2 Corresponding to Treatment Plant Sophistication Ranging from Basic ($\eta$ = 1) to Advanced ($\eta \leq 0.2$)

In the five papers summarized in the remainder of this chapter, unconditional infection/illness dose-response functions for both Poisson and discrete scaling-distributed pathogens are proposed for use in risk characterization of pathogen exposure in drinking water. In the first two papers (§3.1), the predictive Bayesian dose-response assessment approach is proposed for the infection endpoint and demonstrated using human dose-response data for rotavirus and *C. parvum*. In the third paper (§3.2), the generalized beta-Poisson is proposed for evaluating illness response in laboratory data, and demonstrated for *C. parvum*. In the fourth paper (§3.3), the discrete scaling distribution (DSD) is derived for assessing exposure to pathogens in drinking water, and verified versus simulated long-term drinking water data and validated versus field data. In the fifth paper (§3.4), the beta-discrete scaling dose-response function is derived for drinking water dose-response assessment and demonstrated for *Cryptosporidium*.

### 3.1. PREDICTIVE BAYESIAN DOSE-RESPONSE ASSESSMENT FOR APPRAISING ABSOLUTE HEALTH RISK FROM AVAILABLE INFORMATION (Englehardt, 2004)

Englehardt (2004) presents the predictive Bayesian approach for use in microbial dose-response assessment as an improvement over other methods. In contrast with other methods of handling limited information, Bayesian methods exploit available subjective and related information (unlike resampling plans such as the bootstrap, and likelihood methods such as the Benchmark Dose), as well as numeric data (unlike fuzzy-logic and unconditional frequentist methods). The approach allows quantitative assessment of probabilities (unlike interval bounding methods). Bayesian methods, in general, are becoming widely used to obtain distributions of confidence around the parameters of a function of interest.

In contrast to confidence bounds, a predictive Bayesian probability distribution is an unconditional one, integrated over the entire distribution of parameter uncertainty. That is, the answer takes into account all possible values of the parameters, effectively incorporating all confidence levels. The result can be thought of as the expected value under uncertainty. A primary property, and one of the principal advantages, of the predictive Bayesian distribution (dose-response function) is that it becomes quantitatively narrower as more information is obtained. This means that higher risks (relative to the true unknown risk) are predicted when information is limited. In addition, because the predictive distribution is sensitive to the assumed parametric sampling distribution (i.e., the dose-response model), the method is powerful for the exploitation of information on the expected form of the distribution obtained by theoretical derivation. As any number of unrelated empirical models will probably fit any given (high-dose) data set well, a strong theoretical basis for the conditional model is critical to any dose-response assessment, and the predictive Bayesian approach leverages this information.

Englehardt (2004) demonstrated the predictive Bayesian approach for human infection response data for orally-ingested rotavirus (Ward et al., 1986). The unconditional predictive Bayesian assessed risk was lower than the theoretical maximum risk (exposure risk), but higher than the risk based on fitting the conditional dose-response function to the data. Simulating additional data caused the unconditional risk to drop closer to the conditional result. That is, assessed risk was lower with increasing information availability, as expected, based on information theory. The exercise demonstrates the simplicity of evaluating the value of additional information. The change in risk and the corresponding monetary benefit can be estimated directly from the predictive Bayesian dose-response plot.

### 3.2. PREDICTIVE POPULATION DOSE-RESPONSE ASSESSMENT FOR CRYPTOSPORIDIUM PARVUM: INFECTION ENDPOINT (Englehardt and Swartout, 2004)

Englehardt and Swartout (2004) presented a predictive human population dose-response assessment for *C. parvum* for the infection endpoint, demonstrating a hierarchical predictive approach for microbial dose-response assessment. Available data on the infectivity of three isolates of *Cryptosporidium parvum* were adjusted for sensitive and resistant subpopulations not proportionately represented in the data by bootstrap analysis. The diverse mean infectivities of the isolates were used to obtain a predictive distribution for population infectivity, used in turn to obtain the predictive population dose-response function.

### 3.3. PREDICTIVE BAYESIAN MICROBIAL DOSE-RESPONSE ASSESSMENT BASED ON SUGGESTED SELF-ORGANIZATION IN PRIMARY ILLNESS RESPONSE: *CRYPTOSPORIDIUM PARVUM* (Englehardt and Swartout, 2006)

This paper describes the first probabilistic derivation of a general model of pathogen dose-response and is based on conceptual models of the process of gastroenteric pathogenesis. A self-organized, critical (SOC) model of host-pathogen interactions was postulated, such that once an infection is established in the GI tract, the wellness (i.e., fitness) of the least well GI tract segment will increase (host winning) or decrease (pathogen winning) randomly in a time step. The basic model was a linear 30-segment representation of the GI tract, the only parameters of which were those of the infection dose-response relationship (beta-Poisson) and a simulated clinical diagnosis of illness (e.g., diarrhea). Infection in the basic model was established according to the beta-Poisson relationship. Illness is determined to be over when the wellness value of all segments in the GI tract are above a self-organized critical value (0.678) that emerged from the simulation runs. Characteristics of the resulting probability of illness are determined entirely by self-organization—that is, by self-selection of the least well segment and its neighbors for random wellness revision in each time step. Simulations involved testing of one million “hosts” at multiple doses. The results of the simulation (probability of illness with respect to dose) was best described by a beta-Poisson distribution multiplied by a constant ($\leq$1). The constant fractional multiplier is mathematically equivalent to the susceptible proportion of the population, or one minus the immune fraction. Distributions of severity at a dose were characterized by power law distributions ranging over several orders of magnitude. That is, the severity of illness in this model behaved as an “incident,” the size of which is

proportional to the product of the preceding cause sizes according to early complex systems theory (Chow, 1954; Lomnitz, 1964; Benjamin and Cornell, 1970; Englehardt, 2002). This finding may have significance in the future development of more realistic morbidity models incorporating variable severities, such as attempted for by the disability adjusted life year metric for *Campylobacter jejuni* illnesses (Havelaar et al., 2000).

### 3.4. A NEW THEORETICAL DISCRETE "SCALING" PROBABILITY DISTRIBUTION WITH VERIFICATION FOR MICROBIAL COUNTS IN WATER (Englehardt et al., 2008b)

The form of the distribution of pathogen counts in drinking water and other low mean-count media is currently unknown, partly because most counts are zero. However, the mean of the count distribution is directly proportional to the dose of pathogens and, as the outcome of a complex system, the count distribution potentially "scales" over several orders of magnitude. Therefore, the long-term dose may be governed by rare, high-count events not generally represented in available short-term plant monitoring data, and the distributional form is needed to help assess this dose. The form of the count distribution is also needed in dose-response assessment, to extrapolate health response from high doses to low. In this paper a new discrete scaling distribution (DSD) was derived for assessing microbial counts in finished drinking water. The exponential parameter of the new distribution corresponds in concept to a real-valued inverse number of causes of pathogen counts. Scaling of the distribution is a consequence of interaction (dependence) among cause sizes. The DSD was shown to fit well to low mean-count data (containing many zeros) such as drinking water. Conversely, the Poisson-lognormal (PLN), a commonly used distribution (Haas et al., 1999; Masago et al., 2004) was more efficient at fitting high mean-count samples such as source water.

### 3.5. A DOSE-RESPONSE FUNCTION FOR DISCRETE-SCALING DISTRIBUTED PATHOGENS SUCH AS FOUND IN DRINKING WATER (Englehardt et al., 2008a)

This paper presents the derivation of a generalized dose-response function based on the new DSD count distribution, for the illness and infection endpoints, analogous to the generalized beta-Poisson function derived for Poisson-distributed organisms in Englehardt and Swartout (2006). The shape of the conditional function is established for the illness endpoint by means of simple, self-organized, critical models of gastro-enteric pathogenesis. The unconditional predictive Bayesian form of the

derived beta-discrete scaling dose-response function is then plotted using posterior distributions computed previously from available data for *Cryptosporidium*, for infection and illness assuming various numbers of conceptual unit processes. This range of values of the treatment parameter is typical of the number of unit treatment processes in drinking water treatment plants, and of microbe count distributions fitted to empirical data previously.

In contrast to the beta-Poisson, the beta-discrete scaling function is not linear at doses above $10^{-5}$. Therefore, linear low-dose extrapolation assuming a dose-response slope of unity, as assumed for Poisson-distributed organisms, may not be appropriate. In particular, long-term average drinking water concentrations accounting for plant upset events may be above the linear range (Englehardt et al., 2008b). In such cases, the implementation of additional treatment barriers at a plant would result in a lower-mean, higher-skew distribution of pathogen counts in the treated water, and a relaxation of the dose associated with an acceptable response level. On the other hand, at the level of $10^{-7}$ illnesses/exposure originally envisioned in the U.S. Environmental Protection Agency (U.S. EPA) Surface Water Treatment Rule, the allowable *C. parvum* concentration would be on the order of $10^{-6}$ oocysts/L for any treatment plant. These results do not consider secondary transmission of illness or exposure-driven immunity, both subjects of current research.

# 4. DISCUSSION AND CONCLUSIONS

The primary theme in this body of work is the utility of the predictive Bayesian framework for the evaluation of pathogen dose-response functions in a data-poor environment. Proponents of Bayesian methods, in general, argue that a Bayesian approach is a virtual requirement when trying to assess *probabilities* of future outcomes, rather than the *frequencies* of past events (Bernardo and Smith, 1994; Berry, 1996; Carlin and Louis, 2001). Bayesian methods provide a rigorous framework for common-sense interpretation of statistical conclusions. Frequentist approaches can only place confidence limits on a result that depends on specific conditions, leading to inferences that might be made in repeated practice. Bayesian proponents point out that most people interpret, erroneously, frequentist results in the Bayesian sense, anyway (Gelman et al., 2004). Bayesian methods generally provide much greater flexibility than do frequentist methods, allowing them to cope with very complex, data-limited problems.

The predictive Bayesian approach is different from "traditional" Bayesian methods in that the result is unconditional. In more standard Bayesian analyses, prior distributions are assigned to individual parameters of the model and the outputs are expressed similarly, as updated "posterior" distributions of the parameters based on a consideration of the empirical data. If no non-numerical dose-response information is available or desired for use, the prior can be made non-informative, resulting in the traditional frequentist assessment. The mean or mode of the posterior is then taken alone and used as a Bayesian estimate of the parameters of the conditional dose-response function. The predictive Bayesian approach, in contrast, involves integrating the dose-response function over the entire range of the posterior distribution, effectively averaging the dose-response function (not the posterior) over the range of uncertainty, to produce a dose-response function that is not conditional upon the uncertain parameters. Because the predictive Bayesian integrates across all values of the parameters, interpretation of the "answer" does not depend on consideration of some relatively arbitrary confidence limit. With respect to dose-response functions, the predictive Bayesian output is a single risk curve—devoid of confidence bounds—that can be described as the "believed" risk, but is more rigorously termed the unconditional (with respect to parameter uncertainty) probability of response, representing the expected value of the dose-response function under uncertainty. This unconditional probability of response is more conservative (higher) when information is scarce than when there is a surfeit of data. As shown in Section 1.1 (Figure 1), the predictive

Bayesian result is more conservative than the corresponding frequentist expected value. With "perfect" information, the predictive Bayesian result converges on the frequentist solution, which is why interpretation of the latter is more difficult. These characteristics make the predictive Bayesian approach an ideal choice for representing the expected outcomes of competing risk management options.

The predictive Bayesian approach has been demonstrated for both the infection and illness endpoints for rotavirus and *Cryptosporidium parvum*, under several different assumptions about the nature of the pathogen and the host and the exposure distribution. The rotavirus assessment (Englehardt, 2004) is the simpler case involving the infectivity of a single strain administered to a single population of susceptible adult hosts, primarily as an introduction to the use of the method for pathogen dose response. The assessment of *C. parvum* infection was the more complex application. Both pathogen and host variability were assessed in multiple dimensions. The assessment was unique in that host response was modeled for the entire human population, including sensitive and resistant sub-populations, as well as for susceptible adults. The human population response was evaluated separately across three separate isolates of *C. parvum* and integrated into a final unconditional population response curve by means of the predictive Bayesian method. The result was shown to be more health-protective than frequentist methods, but not extremely so (Englehardt and Swartout, 2004). A much more conservative result was obtained when all the *C. parvum* human response data are pooled, assuming no difference among the isolates (Englehardt and Swartout, 2006; Englehardt et al., 2008b). Under this assumption, the cross-isolate variability was subsumed within the inter-individual host-pathogen variability, resulting in a stretching out of the dose-response curve. The unconditional unit infectivity estimate increased by about an order of magnitude compared to the separate strain assumption for the individual isolates. In addition, as the morbidity ratio was very high in the study subjects, and given slightly greater noise in the illness data compared to infection, the unconditional probability of illness was virtually the same as for infection for the single strain scenario. Simply fitting the beta-Poisson dose-response function to the same data results in a slightly lower probability of infection at low dose (0.45 vs. 0.50), but a significantly lower probability of illness (0.38 vs. 0.50). The difference reflects the incorporation of uncertainty by the predictive Bayesian method into the unconditional result.

The self-organized-criticality (SOC) model used in both illness modeling papers (Englehardt and Swartout, 2006; Englehardt et al., 2008a) provides a unique and innovative approach by suggesting a general form of the illness dose-response

relationship. The behavior of the illness severity measure as an “incident,” similar to other phenomena in nature and engineered systems, may have significance in the future development of more realistic morbidity models incorporating variable severities. That is, insights into the distribution of various outcomes in a comprehensive model of continuously-variable severities, from infection to diarrhea to death, is of critical importance. The realization of such a model would greatly reduce the response classification error inherently present in the current categorical response approaches.

The discrete-scaling distribution developed in Englehardt et al. (2008b) is based on the consideration of water treatment processes having failure rates that are, at least, partially (positively) dependent. In this sense, failure rates have been likened to cause sizes in a complex system incident-size distribution analysis. Other published water treatment process simulations have considered the failure rates to be independent (Haas and Trussell, 1998; Masago et al., 2004; Signor and Ashbolt, 2005). The primary difference in the assumptions is that dependent processes will result in a more skewed distribution of pathogens in finished drinking water than for independent processes. The implications of this behavior are two-fold:

> First, adding more treatment processes may lower mean occurrence but will likely increase dispersion, resulting in a higher probability of larger events. The latter could mean a greater risk of outbreaks than would be predicted under the assumption of a less-skewed distribution. Longer monitoring records are needed to predict long-term exposure. Perhaps of more significance is the potential for significant underestimation of the long-term mean pathogen concentration when the temporal distribution is highly skewed, where high-count rare events greatly influence the mean. Short-term monitoring efforts, even with adequate sample volumes, may underestimate the longer-term population risk significantly.
>
> Second, a more disperse distribution will generally result in a lower aggregate risk of infection with a given mean pathogen concentration. The latter is the result of an increase in the probability of exposure per exposure event to more than one pathogen unit (organism) at low exposure concentrations (<<1 organism per liter), but a corresponding increase in the number of null exposures (failure to ingest any pathogens). Because the increase of unexposed individuals results in a proportional decrease in population risk and the increase in risk from exposure to more than one unit per event is less than proportional, the net result is a decrease in overall population risk for more disperse exposure distributions. The difference becomes more pronounced with increasing skewness and increasing dose.

Follow-up work will necessarily include long-term data simulation to investigate the impact of distribution parameter assumptions on population risk.

In a practical sense, the DSD appears to be favored over the PLN for modeling exposure to sparse, high-zero, low-count data (i.e., occurrence data characterized by mostly non-detects and positive counts covering a relatively narrow range). This conclusion is not based so much on the results presented in Englehardt et al. (2008b) where a slight preference for the DSD was found, but rather on preliminary results from a more detailed analysis of New York City reservoir data and simulations of observation records of various lengths (Englehardt et al., 2007). In the latter study, a strong preference for the DSD over the PLN is shown by the ability of the model to predict the longer-term distribution from short-term records. Additional work is being conducted to clarify these relationships.

The impact of discrete-scaling distributed pathogens in drinking water was investigated in Englehardt et al. (2008a), by assuming that exposure at the tap for drinking water consumers would be similar to the one found at the treatment plant. The effect of the use of the proposed function, as opposed to the beta-Poisson dose-response function, was shown to be a partial relaxation of the apparent allowable concentration of pathogens in drinking water for a fixed level of acceptable risk. This difference was more noticeable for microbial count distributions representing the case of treatment plants that present a more effective overall barrier to pathogen breakthrough (i.e., more treatment processes). However, the difference between treatment plants may or may not be appreciable at very low exposure levels depending on the maximum acceptable risk level assumed. For example, given the results of this analysis (Englehardt et al., 2008a), the concentration corresponding to an acceptable risk level of $10^{-4}$ illnesses/year would be on the order of $10^{-6}$ oocysts/L for any treatment plant, assuming a drinking water consumption of one liter per day. This exposure estimate does not consider secondary transmission of illness or acquired immunity, both subjects of current research. With “perfect” information, that is knowledge of the true population response, the dose corresponding to an unconditional risk of $10^{-4}$ illnesses per year might be an order of magnitude higher. The extent to which the predictive Bayesian underestimates that dose (i.e., overestimates risk) reflects the degree to which the data are imperfect. Frequentist methods typically reflect only particular types of experimental and physical variability, without addressing the often-dominant uncertainties rooted in information limitations.

With respect to *Cryptosporidium parvum*, used as an example in most of the papers cited above, the single most significant factor is pathogen variability. Treating the *C. parvum* isolates as distinct strains results in an unconditional unit risk of infection of about 0.05 (Englehardt and Swartout, 2004). If the isolates, instead, are assumed to

be random samples of the same strain, the unconditional unit risk of infection is an order of magnitude higher (Englehardt and Swartout, 2006; Englehardt et al., 2008b). Although part of the increased risk in the latter scenario was due to the addition of new data, most was a result of the single strain assumption. In contrast, human population variability does not appear to be as big a factor. Taken individually, most of the *C. parvum* isolate response data sets are fit adequately by the exponential model, implying a lack of inter-individual variability in the human hosts. The simulation of sensitive and resistant subpopulations had little impact on the overall population risk (Englehardt and Swartout, 2004). Changing the underlying pathogen distribution to a more disperse one resulted in significant impact in the extreme case at higher exposures but, in most cases, did not change the risk substantially at expected exposure levels in drinking water ($<10^{-4}$ oocysts per liter). It would be premature to generalize these conclusions, however, as *C. parvum* may be somewhat unique with respect to other microorganisms. Follow-up work in evaluating the nature of the response differences across the *C. parvum* isolates is indicated.

The methods developed in this body of work are designed to provide scientifically-defensible and rapid (essentially on-demand) dose-response assessments with the following caveats: (a) currently, only primary response to an ingested dose (e.g., in drinking water) can be assessed, (b) the temporal effects of acquired immunity on population response are not well-understood and (c) the applicability of laboratory human dose-response data to actual response to environmental strains of pathogens is not well-understood. First, to understand the total disease burden in the population, modeling of the secondary transmission of disease and dose-response for other sequellae (other disease endpoints from the one pathogen infection) in the population is necessary. Second, perhaps using the same model, the long-term effect of acquired immunity on the primary response needs to be evaluated. Third, questions remain as to whether strains of pathogens cultured in the lab over various time periods accurately represent the infectivity of the strains existing in the environment.

## 5. RESEARCH NEEDS

The following list of research needs is not necessarily comprehensive or in order of priority. The highest priority tasks will depend on the specific needs of the U.S. EPA program offices as overarching regulatory processes develop. The first two research needs, however, address specific regulatory programs within the Office of Water and may have higher priority.

1. Assess the relative virulence and the effect of laboratory culturing on strain virulence for all *C. parvum* isolates and determine their prevalence in the environment. Fully assess the relationship among *C. parvum* and C. *hominis* isolates (one strain or many?) by means of statistical modeling. Published assessments of the infectivity and morbidity of *C. parvum* vary by more than an order of magnitude. The choice of parameter values will have a significant impact on the 6-year review of the Long-Term 2 Surface Water Treatment Rule (due in 2012).

2. Develop unconditional dose-response functions for CCL organisms based on available information. There are a few organisms on the proposed CCL3 for which some data exist.

3. Model the effect of long-term source-water variability on finished-water pathogen distributions. The current version of the DSD does not take into account, explicitly, source-water variability. Many pathogens exhibit seasonal variability in occurrence. Preliminary analysis of the New York City source water supply suggests that other temporal cycles may exist. This research will have direct impact on human exposure estimation.

4. Evaluate the monitoring strategy necessary for predicting the long-term distribution of pathogens in drinking water (length of record, frequency of measurements, sample volume). The DSD fits to some data sets suggest highly disperse distributions, characterized mostly by zero or low-count samples and long-term mean values dependent on large rare events, suggesting longer monitoring periods, more frequent sampling and larger sample volumes. Alternatively, quantification of pathogens in raw waters and their removal (largely using surrogates) may provide more accurate assessment of their distributions in treated waters (Signor et al., 2007).

5. Evaluate the effect of secondary transmission on the shape of the dose-response function to estimate the total disease burden from primary drinking water exposure.

6. Evaluate the effect of exposure-induced immunity in the population to assess the impact of prior exposure on primary response to pathogens in drinking water.

7. Demonstrate the incorporation of information other than numerical dose-response data in the predictive Bayesian posterior, such as genetic prevalence data and epidemiological information. This work allows for a fuller exploitation of existing data in a rigorous framework, with an expectation of reduced uncertainty in the prediction.

8. Further evaluate the behavior of the predictive Bayesian with respect to change in shape relative to amount of data, using simulated data, to evaluate the value of information.

## 6. REFERENCES


Aitchison, J. and I. Dunsmore. 1975. Statistical Prediction Analysis. Cambridge University Press, New York, NY. 273 pp.

Bak, P., C. Tang and K. Wiesenfeld. 1988. Self-organized criticality - an explanation of 1/*f* noise. Phys. Rev. Lett. 59(4):381-384.

Benjamin, J. and C. Cornell. 1970. Probability, Statistics, and Decision for Civil Engineers. McGraw-Hill Publishing Co., New York, NY.

Bernardo, J.M. and J.F.M. Smith. 1994. Bayesian Theory. John Wiley & Sons Inc., New York, NY.

Berry, D.A. 1996. Statistics: A Bayesian Perspective. Wadsworth Publishing Co., Belmont, CA.

Carlin, B.P. and T.A. Louis. 2001. Bayes and Empirical Bayes Methods for Data Analysis, 2nd ed. Chapman and Hall, New York, NY.

Chow, V. 1954. The log-probability law and its engineering applications. Proceed. Am. Soc. Civil Engr. 80(536):1-25.

Eisenberg, J.N., X. Lei, A.H. Hubbard, M.A. Brookhart and J.M. Colford, Jr. 2005. The role of disease transmission and conferred immunity in outbreaks: analysis of the 1993 *Cryptosporidium* outbreak in Milwaukee, Wisconsin. Am. J. Epidemiol. 161:62-72.

Englehardt, J. 1995. Predicting incident size from limited information. J. Environ. Engrg. 121(6):455-464.

Englehardt, J. 2002. Scale invariance of incident size distributions in response to sizes of their causes. Risk Anal. 22:369-381.

Englehardt, J.D. 2004. Predictive Bayesian dose-response assessment for appraising absolute health risk from available information. Hum. Ecol. Risk Assess. 10:69-78.

Englehardt, J.D. and J. Swartout. 2004. Predictive population dose-response assessment for *Cryptosporidium parvum*: Infection endpoint. J. Toxicol. Environ. Health A. 67:651-666.

Englehardt, J.D. and J. Swartout. 2006. Predictive Bayesian microbial dose-response assessment based on suggested self-organization in primary illness response: *C. parvum.* Risk Anal. 26(2):543-554.

Englehardt, J.D., J. Swartout and C. Loewenstine. 2007. The Effect of Record Length on the Assessed Microbial Dose in Drinking Water. Presented at the Toxicology and Risk Assessment Conference, Cincinnati, OH. April 25.

Englehardt, J.D., C. Loewenstine and J. Swartout. 2008a. A dose-response function for discrete-scaling distributed pathogens such as found in drinking water. Environ. Sci. Tech. (submitted).

Englehardt, J.D., J. Swartout and C. Loewenstine. 2008b. A new theoretical discrete scaling probability distribution with verification for microbial counts in water. Environ. Sci. Tech. (submitted).

Frisch, U. and D. Sornette. 1997. Extreme deviations and applications. J. Phys. I. 7(9):1155-1171.

Gelman, A., J.B. Carlin, H.S. Stern and D.B. Rubin. 2004. Bayesian Data Analysis, 2nd ed. Chapman and Hall, New York, NY.

Haas, C. and R. Trussell. 1998. Frameworks for assessing reliability of multiple independent barriers in potable water reuse. Water Sci. Technol. 38(6):1-8.

Haas, C., J. Rose and C. Gerba. 1999. Quantitative Microbial Risk Assessment. John Wiley & Sons, Inc., New York, NY.

Havelaar, A.H., M.A. de Wit, R. van Koningsveld and E. van Kempen. 2000. Health burden in the Netherlands due to infection with thermophilic *Campylobacter* spp. Epidemiol. Infect. 125:505-522.

ILSI (International Life Sciences Institute). 2000. Revised framework for microbial risk assessment. An ILSI Risk Science Institute workshop report. International Life Sciences Institute. ILSI Press, Washington DC. Available at http://rsi.ilsi.org/file/mrabook.pdf.

Lomnitz, C. 1964. Estimation problems in earthquake series. Tectnophysics. 2:193-203.

Masago, Y., K. Oguma, H. Katayama, T. Hirata and S. Ohgaki. 2004. *Cryptosporidium* monitoring system at a water treatment plant, based on waterborne risk assessment. Water Sci. Technol. 50(1):293-299.

Shannon, C. 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27:379-423, 623-656.

Signor, R.S. and N.J. Ashbolt. 2005. Pathogen monitoring offers questionable protection against drinking-water risks: a QMRA (Quantitative Microbial Risk Analysis) approach to assess management strategies. Water Sci. Technol. 54(3):261-268.

Signor, R.S., N.J. Ashbolt and D.J. Roser. 2007. Microbial risk implications of rainfall-induced runoff events entering a reservoir used as a drinking-water source. J. Water Supply – AQUA. 56:515-531.

Teunis, P., C. Chappell and P. Okhuysen. 2002. *Cryptosporidium* dose response studies: Variation between hosts. Risk Anal. 22:475-485.

Ward, R.L., D.I. Bernstein, E.C. Young, J.R. Sherwood, D.R. Knowlton and G.M. Schiff. 1986. Human rotavirus studies in volunteers: Determination of infectious dose and serological response to infection. J. Infect. Dis. 154(5):871–880.