MICRORISK

Microbiological risk assessment: a scientific basis for managing drinking water safety from source to tap

QMRA methodology

April 2006

QMRA methodology

Susan Petterson, Ryan Signor, Nicholas Ashbolt and David Roser
University of New South Wales
Sydney, Australia

Contents

Contents **5**

Acknowledgements **7**

Introduction **9**
Step 1 – Context 10
Step 2 - Exposure 11
Source Water 11
Treatment Efficacy 11
Distribution 12
Consumption 13
Step 3 – Dose-Response 13
Dose –response models 13
Experimental Data 16
Step 4 – Risk Characterisation 17
Incorporating events into overall probability of infection 17
Predicting the number of infections from multiple exposures 17
Annual probability of one or more infections 18
Example: Implementation of DALY metric for interpreting probability of infection estimates for *Cryptosporidium*. 20

Tiered approach to qmra **21**
(Higher level) probabilistic analyses 22

Quantifying Variability from Microbiological Datasets **23**
Characteristics of microbial data 24
Detection and Quantification 24
Recovery and Imperfect Detection 26
Sampling Effects 27
Model Fitting and Parameter Uncertainty 28
Parametric distributions 29
Triangular Distributions 34

Quantifying uncertainty **34**
Expert opinion and worst case sensitivity 34
Auditing Score 36

Implementation for CTS's **37**
Source water pathogen densities 38
Microbial Counts 38
Incorporation of Events 39
Presence/Absence results 41
Index Organisms 43
Quantifying Method Recovery 45
Microbial counts 46
Presence/Absence results 50
Quantifying Treatment Performance 51

Transformation model assumptions 52
Ratio of outflow concentration to inflow concentration 54
Binomial models 55
Surrogates 57
Dose-Response 57
Campylobacter 57
E. coli 0157 59
Viruses 60
Giardia 60
Cryptosporidium 61

References 62

Acknowledgements

This study has been performed as part of the MicroRisk project that is co-funded by the European Commission under the Fifth Framework Programme, Theme 4: "Energy, environment and sustainable development" (contract EVK1-CT-2002-00123).

Introduction

Quantitative microbial risk assessment (QMRA) of drinking water systems requires the quantification of pathogen occurrence in source water and their removal through various treatment barriers to the consumer's tap. When pathogen occurrence at the tap is combined with consumption patterns and pathogen dose-response relationships, the risk of infection (or other end-points) can be estimated.

In this report, a framework for calculating and characterising the microbial risk from drinking water is presented (Figure 1). The process of quantifying model inputs and choosing numerical values for each variable is critical to the QMRA process, and yet potentially daunting for the risk analyst. Pathogen sources, transport and removal are complex processes dependent on many factors including hydrology, climate, land use, hydrodynamics, disease incidence, process design and performance, and unpredictable human behaviour.

Experimental data is key, however datasets that relate directly to the variables of interest are limited. Analyses are costly, and pathogens generally occur at low densities in environmental waters making detection difficult. Datasets are characteristically small in size and often contain low numbers and many non-detects. Statistically, analysis of microbial datasets for characterising inputs to QMRA is therefore a less usual case. Whereas most traditional statistical methods are designed to analyse large datasets of relatively low variance, microbial datasets are generally small datasets of high variance with large uncertainties.

The role of the risk analyst is to determine what the limited monitoring or experimental results reveal regarding the magnitude and variability of each quantitative input to the risk model. Statistical approaches appropriate for describing model inputs for QMRA from microbial data are therefore presented. Examples are given that require a specific and sometimes rigorous consideration of the relevant experimental data collected during the MicroRisk project. The authors argue that the greatest value will be obtained from the experimental data when the statistical analysis approach is tailored to the individual dataset. The aim of these detailed analyses is to learn as much as possible from the available information regarding appropriate quantitative estimates of model inputs, and the uncertainties associated with these estimates.

Uncertainties[1] in microbial risk modelling are important to untangle from variability[2]. The risk management implications of an isolated probability of infection estimate can be difficult to interpret without further understanding of how and why model inputs may vary and where the most important sources of uncertainty lie. The aim of this report is not only to demonstrate how risk calculations were undertaken, but also to give some guidance regarding the interpretation of uncertainty associated with modelling results.

In the MicroRisk project, a pragmatic approach to characterising uncertainty was applied by combining traditional quantitative methods and a more semi-quantitative

[1] Uncertainties arise due to lack of precise knowledge of the input values or to lack of knowledge of the system being modelled, more data typically reduces these uncertainties.

[2] Variability refers to observed differences attributable to true heterogeneity or natural diversity in a parameter that cannot be reduced by additional data collection (but can be better characterised).

approach drawing on expert opinion. An expert may have prior knowledge that the estimated value for a model variable (quantified based on a small or surrogate dataset) is unlikely to be representative. In fact, discussion of this prior knowledge, aimed at ensuring the representativeness of data and assumptions to the real systems and processes being studied, was identified as a critical component of the quantitative risk assessment process. The importance of these kinds of uncertainties (or scepticisms) on risk calculations was explored using sensitivity analysis.
A general framework for estimating pathogen risks from drinking water is illustrated in Figure 1.

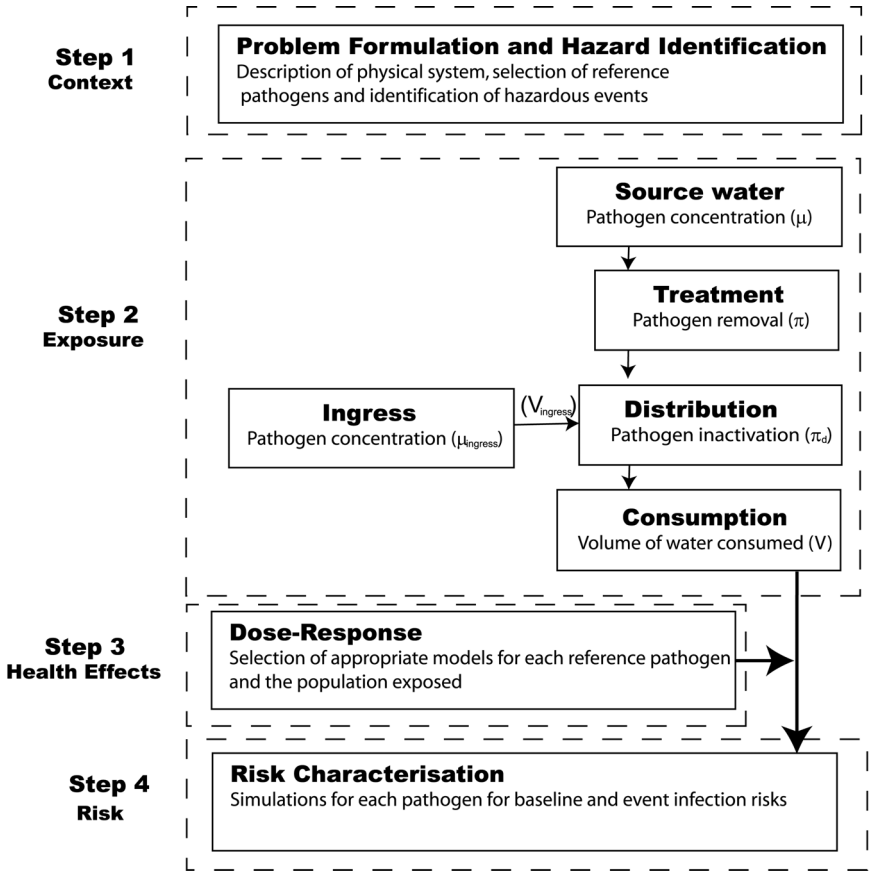


Figure 1 General framework for calculating microbial risk from drinking water

Step 1 – Context

The first step in the QMRA process is to define which pathogens will be modelled and what conditions will be investigated. These choices should be made so that the range of pathogen types are modelled under both baseline and hazardous event conditions, providing the context for the QMRA.

Step 2 - Exposure

Source Water

Pathogens may be present in the water column at the treatment plant off-take due to human and/or animal inputs (waterborne enteric viruses being assumed to only come from human excreta) within the catchment. The density of pathogens at the treatment plant off-take is dependent upon the magnitude of pathogen inputs and the environmental processes affecting their transport and inactivation; and is expected to vary both over time and between pathogen groups as described in Chapters 3-4[3].

To calculate microbial risk the density of pathogens (number of micro-organisms per litre) in the source water must be quantified and entered into the risk model. The main sources of information for quantifying pathogen density in source waters are:

- Water samples collected from the site and analysed directly for the presence of pathogens;
- Water samples collected from the site and analysed for (pathogen) index organisms combined with some assumptions regarding the ratio of index organisms to pathogens (direct analysis of pathogens is often not undertaken as analytical methods for detection are costly, and pathogens are often present in low densities requiring the collection of large volume samples); and
- Literature data from a catchment of similar pathogen sources and physical characteristics.

For some systems, significant changes in pathogen density can be linked to specific events affecting the mobilisation of micro-organisms from their source to surface waters such as rainfall induced runoff, and discharge of sewage overflows. Identifying the occurrence and impact of such hazardous events can facilitate understanding and management of the microbial risks for a given system. Describing the source water concentration for such a system by a single estimate may underestimate the peak risks, as high pathogen densities can be dampened by nominal low densities. A simple approach to address this is to describe separately pathogen densities under "event" and "nominal" conditions, leading to a bimodal description of pathogen density.

Treatment Efficacy

A wide range of treatment processes exist for the physical, chemical and microbiological purification of drinking water as indicated in Chapter 4. Each of these processes contributes to the removal or inactivation of pathogens from the water column. The effectiveness of each process in removing pathogens is variable: between different types of processes; between the same processes operated at different treatment facilities; and even variable over time for an individual process at a specified treatment plant.

Quantifying treatment removal performance for a drinking water CTS, accounting for the individual characteristics of the system being studied, and the expected temporal variability in performance for each process unit is a great challenge. Careful consideration of the available data is essential. In the QMRA framework, removal performance is represented by π (Figure 1) which is the fraction of organisms passing any treatment barrier (or barriers). When multiplied by the source water concentration ($\mu\square\pi$), the pathogen density in finished (treated) water may be estimated.

The primary sources of data for quantifying treatment performance (π) include:

[3] "Chapter" in this report refers to the corresponding chapter in the final report of MicroRisk.

- Pathogen densities at the inlet and outlet for a process or treatment plant;
- Surrogate densities at the inlet and outlet for a process or treatment plant; and
- Online performance data including turbidity, and chlorine residual.

Treatment efficacy is often reported as decimal elimination or Log_{10} reduction in micro-organism density. Log_{10} reduction is simply the Log_{10} of π, and therefore can be directly transformed to an estimate of π for input to the risk model ($\pi = 10^{Log10 reduction}$).

Distribution

An ideal distribution system protects water quality as it transports treated water from the plant to the consumer's tap. The only effect on pathogen density should be a reduction due to inactivation with travel time (increased in the presence of a disinfectant residual), and incorporation into pipe biofilms.

In reality however two types of events in the distribution system may lead to an increase in the pathogen concentration between the treatment plant and the consumer:

1. **Deficiencies in the distribution system** may lead to the ingress of pathogen contaminated material including cross-connections, contamination while in storage, contamination during construction or repair, and broken or leaking mains. In addition, common hydraulic transients may lead to contamination through negative pressure and subsequent intrusion of soil water.
2. **Biofilm sloughing events** (caused by shear force from changes in water flow or change in disinfectant concentration) may lead to incorporation of pathogen rich material from the internal pipe surface into the water column [Storey and Ashbolt, 2003].

Calculation of the impact of events within the distribution system on the microbial risk to the consumer requires the quantification of the frequency and duration of each type of event, along with the numbers of pathogens incorporated into the drinking water. Techniques for identifying the occurrence and impact of these events are still in their infancy and there is a great need for research in this area, as described in Chapter 5.

The current risk model oversimplifies the problem by looking at the impact of an ingress event on the pathogen concentration at the consumer's tap by considering the volume of contaminated material entering the water ($V_{ingress}$) and the pathogen concentration in the contaminated material ($\mu_{ingress}$). Within this framework, the relative importance of ingress events on consumer risk can be explored. The concentration of pathogens at the tap may be calculated using Equation 1.

$$\mu_{tap} = X\mu_{ingress} + (1 - X)(\mu.\pi)$$

Equation 1

Where: $X = \frac{V_{ingress}}{Q \times t}$ if no ingress event occurred, then $X = 0$

(X) represents the proportion of external (ingressed) material present in the water column at the tap; $V_{ingress}$ is the volume of contaminated material entering the water column over time t for an ingress event; and Q is the flow rate in the pipe at the time of the ingress event.

No attempt has been made to quantify the impact of pathogen incorporation into and subsequent sloughing of biofilms. Given the oversimplified framework, however, the sloughing of biofilm could be tested as a special case of an ingress event where the estimated pathogen density in the biofilm is combined with the volume of material sloughed into the water with the subsequent concentration at the tap calculated using Equation 1. The likelihood and overall impact of these ingress events on the probability of infection to the consumer may then also be characterised (see section 0).

Consumption

The volume of water consumed must be quantified in order to estimate the dose of pathogens. Results from analysis of unboiled tap water consumption patterns (Chapter 6) indicate that the amount of water consumed is influenced by many factors including age, culture (or nationality) and level of physical activity. The volume of water consumed (litres per day) is multiplied by the pathogen concentration at the tap to calculate the total exposure or dose ($Dose = \mu_{tap} \times V_{consumed}$) per day.

Step 3 – Dose-Response

Dose-response modelling is the key to microbial risk assessment as it provides a link between exposure dose and the probability of infection. Prior to dose-response relationships, human feeding experiments were only used to estimate infectious doses such as ID_{50} or minimum infective dose (MID). However, in more recent years it has become clear that infection is theoretically possible from exposure to a single organism, and the use of models based on the 'single-hit' theory of dose-response have increased [Regli *et al.*, 1991; Haas *et al.*, 1993; Gerba *et al.*, 1996b].

Dose –response models

Quantitative dose-response models have been developed to estimate the probability of infection based on the average pathogen dose [Haas *et al.*, 1983]. While the average dose of pathogens is continuous and can potentially take any value, the actual number of organism that an individual may consume is a discrete quantity (i.e. it is not possible to consume 2.67 *Cryptosporidium* oocysts, but rather given an average dose of 2.67 most individuals would consume 2 or 3 oocysts with a fewer number consuming lower [0, 1] or higher numbers [6, 7]). Beginning with the average dose, the calculation of probability of infection is a two step process, being the combined probability of exposure and infection shown in Equation 2.

$$P(\text{inf} \mid \mu) = \sum_{n=0}^{\infty} P(n \mid \mu) \times P(\text{inf} \mid n)$$

Equation 2

Where: $P(\text{inf} \mid \mu)$ is the probability of infection given the mean pathogen density.
$P(n \mid \mu)$ is the probability of exposure to *n* organisms given the mean pathogen density μ.
$P(\text{inf} \mid n)$ is the probability of infection given exposure to *n* organisms

The distribution of pathogens in the exposure media is assumed to be random, and therefore the probability of exposure to *n* organisms when the mean concentration is equal to μ (P(n| μ)) is given by the Poisson distribution.

When an individual organism is ingested, the probability of that organism successfully overcoming host barriers and reaching a site for infection may be represented by *r*. If every organism is assumed to behave independently from other organisms within the host, then the overall probability of infection may be described as a binomial process. That is, each ingested organism may result in one of two outcomes; infection or not infection. If the probability that an individual organism may cause infection is denoted by *r*, then the probability of not being infection is equal to (1-*r*). Over a series of *n* independent trials (in this case, number of organisms consumed), the probability of not being infection is equal to $(1-r)^n$, and hence the probability of at least one organism being successful in causing infection is the complement:

$$P(\text{inf} \mid n) = 1-(1-r)^n$$

Equation 3

The implementation of the Poisson pathogen distribution and binomial probability of infection (Equation 3) leads to a family of models referred to as single-hit models, where the name relates to the concept that only a single organism is necessary to cause infection. The simplest form of the single-hit model assumes that for a given pathogen, every pathogenic particle within every host has the same constant probability of survival, given by *r*. When combined with the P(n| μ), the dose-response relationship is the exponential model.

Exponential model: When organisms are distributed randomly (Poisson) and the probability of infection for any organism equals r then:

$$P_{\text{inf}} = 1-e^{-r\mu}$$

Equation 4

While the exponential model is simple, the practical implications are unsatisfying since the between pathogen variation in infectivity, and between host variation in susceptibility is ignored. This limitation is partially overcome by Beta Poisson model.

Beta Poisson model: When *r* is assumed vary according to a beta distribution, a complicated dose-response relationship emerges containing a confluent hypergeometric function [Haas *et al.*, 1999]. Furumoto and Mickey [1967] made some simplifying assumptions to this relationship, and derived a simple dose-response relationship referred to as the Beta Poisson:

$$P_{\text{inf}} \approx 1-\left(1+\frac{\mu}{\beta}\right)^{-\alpha} \quad \text{which holds when } \beta \geq 1 \text{ and } \alpha \leq \beta$$

Equation 5

The Beta Poisson approximation has been widely applied for describing dose-response relationships for QMRA. In some studies, the Beta-Poisson approximation has been applied even when the criteria for the parameter values (Equation 5) are not satisfied. A notable example is the dose-response relationship for Rotavirus infection fitted to data from Ward *et al.* [1986] with maximum likelihood parameters of ($\hat{\alpha}$ = 0.253, $\hat{\beta}$ = 0.422). The implications of this inappropriate application, particularly as it relates to the maximum risk curve are discussed below.

Maximum Risk Model: An important property of the single-hit relationship is that a maximum risk curve exists. The maximum risk curve is calculated when the probability that an ingested organism will pass the host's defense mechanisms and find a site suitable for colonisation is maximised and assumed equal to 1. The resulting equation is therefore the exponential dose-response function with $r = 1$.
This property is not retained by the Beta-Poisson approximation. In a study aimed at investigating the Rotavirus Beta-Poisson model fitted to data from Ward *et al.* [1986] with maximum likelihood parameters of ($\hat{\alpha}$ = 0.253, $\hat{\beta}$ = 0.422), the upper confidence level of the dose-response relation was shown to exceed the maximum risk curve [Teunis and Havelaar, 2000].
In addition for some models used in the MicroRisk calculations, the Beta-Poisson approximation was shown to exceed the maximum risk curve at low doses. This exceedance is illustrated for the *Campylobacter* model (again $\hat{\alpha}$ and $\hat{\beta}$ do not satisfy the criteria of approximation), in Figure 2[4]. The implication was that at low doses, the dose-response model was predicting theoretically impossible probability of infection estimates. As an alternative, for low doses (< 0.1 org.L^{-1}) the exact Beta Poisson model can be approximated by setting r (Equation 4) equal to the expected value of the Beta distribution ($\alpha/\alpha+\beta$)), thus avoiding this complication.

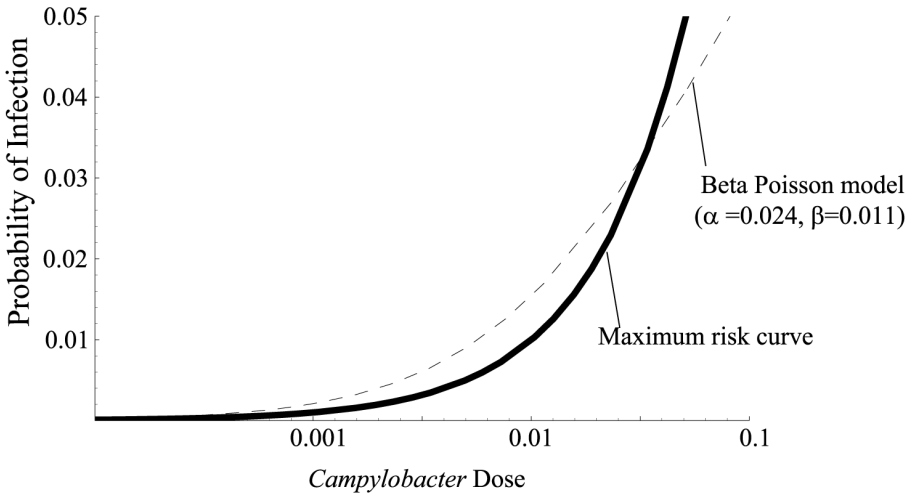


Figure 2. *Campylobacter* and maximum risk dose response curves at low doses

[4] Teunis *et al.* (2005) fitted the exact Beta Poisson model to the dose-response data to find estimates of the parameter values α and β. Here, the parameter values have been used in the Beta Poisson approximation (Equation 5).

The maximum risk curve is also an important tool for uncertainty analysis, providing the upper bound of possible infection response. The importance of uncertainties in the calculation of the dose-response relationship can be screened using the maximum risk curve as a worst case sensitivity input.
The maximum risk curve could also be applied for risk assessment of pathogens with unknown properties. While for highly infectious pathogens the maximum risk curve appears to be a reasonable conservative assumption, it is however important to be aware that for less infectious pathogens, the maximum risk curve may significantly overestimate infection risk.

Experimental Data

In order to estimate the parameters of a dose-response model it is necessary to fit the dose-response relationship to some data. There are two primary sources of relevant data; they are from human feeding trials and unsolicited outbreaks.

Human Feeding Trials: Human feeding trials are controlled experiments where "volunteers" are administered doses of different pathogen concentrations. The number of volunteers who then exhibit an infective response are recorded. Important uncertainties associated with these studies include:

- The number of *viable* particles in the dose is unknown. Depending on the source of the inoculum and the individual pathogen, there is uncertainty as to how many of the administered particles were actually infectious at the time of consumption.
- *Strain* of the micro-organisms contained in the inoculum. Practicalities drive the sourcing of pathogens for the feeding trials. In some circumstances the strain of the administered organisms varied from the strain most likely to cause infection in humans, for example, most *Cryptosporidium* feeding trials have been undertaken using *Cryptosporidium parvum,* whereas most human infections are thought to be caused by various strains of *Cryptosporidium hominis*.
- Representativeness of volunteers. For ethical reasons human feeding trials are conducted on healthy adults who's immune response may not be representative of the entire population.

Outbreak data: In more recent years, information from outbreaks of enteric illness has been used to estimate dose-response parameters [Teunis *et al.*, 2005; Teunis *et al.*, 2004]. The great advantage of data from a real outbreak is that it demonstrates an actual response to exposure to human pathogens, without the constraints and simplifications necessary for a controlled study; pathogens are native to the system, and those exposed are a true sample from the susceptible population. Conversely however, additional uncertainties are introduced including:

- Estimating the dose. There is an incubation period between the time a pathogen is ingested and when a response (illness) is identified. Due to this incubation period, by the time an outbreak is identified, the source material is unlikely to be available for direct analysis. If it is available, the pathogen density may no longer be representative of the density at the time of exposure (due to inactivation or growth).
- Illness rather than infection is the endpoint. In a controlled feeding trial, blood serum can be analysed on a daily interval following exposure to identify

whether or not an individual has been infected. For a real outbreak, identification of a response is limited to those who report *symptoms* of infection (illness) which is only a portion of the total infected population.

Step 4 – Risk Characterisation

The aim of risk characterisation is to integrate information from exposure and dose-response assessment to express public health outcomes. Dose-response models are concerned with estimating probability of infection. Infection has been defined as a situation in which the pathogen, after ingestion and surviving all host barriers, actively grows at its target site [Last, 1995]. Infection may or may not result in illness, as asymptomatic infection can be common for some pathogens.

Incorporating events into overall probability of infection

Probability of infection estimates are based on the calculated exposure to pathogens. While the baseline (nominal) exposure can be calculated based on the expected variability in model inputs, it is often desirable to incorporate the likelihood and magnitude of certain events into the overall probability of infection estimate. One approach for undertaking this analysis is to calculate the probability of infection for each event condition that is to be investigated, and then to combine all events and nominal conditions based on their probability of occurrence (Equation 6).

$$P\inf = \sum_{i=1}^{n} Pevent_i \times P\inf i + (1 - \sum_{i=1}^{n} Pevent_i) \times P\inf_{no\min al}$$

Equation 6

Where: Pinf is the overall probability of infection
n is the total number of event conditions to be included
$Pevent_i$ is the probability of event *i* occurring
$Pinf_i$ is the probability of infection given that event *i* has occurred
$Pinf_{nominal}$ is the probability of infection under baseline or nominal conditions.

Predicting the number of infections from multiple exposures

When multiple exposures (either due to many individuals being exposed at the one time, one individual being exposed on multiple occasions, or a combination) are assumed to be independent events, then the number of infections (successes) may be described as a binomial random variable (X). The probability that the number of infections will equal a given number (k) is:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Equation 7

Where: *k* is the number of infections
n is the number of trials (i.e. for the number of infections per year for an individual, n = 365; for the number of infections per year for a population of 10 000, n = 3 650 000)
p is the probability of infection

This distribution can be maximized to find the most likely number of infections based on the calculated P_{inf}.

Annual probability of one or more infections

If consecutive exposures are assumed to be independent, the annual probability of 1 or more infections may be calculated under the assumptions of a binomial process (a series of trials with one of two possible outcomes – infection or not infection). If the probability of infection for an individual exposure is given by P_{inf}, then the probability of not being infected is $(1-P_{inf})$. For *n* exposures, the probability of not being infected is given by $(1-P_{inf})^n$. The annual probability of one or more infections is the corollary of this for $n = 365$, and is given by Equation 8 :

$$P_{ann} = 1 - (1 - P_{\inf})^{365}$$

Equation 8

When $P_{inf} << 1$, this may be approximated as $P_{ann} = 365 \times P_{inf}$

Incorporating the impact of events into the annual probability of one or more infections

Equation 6 calculated the probability of infection given the likelihood of a range of possible event scenarios. It is also possible to consider the impact of events on the yearly probability of one or more infections when it is assumed that one (or more) events occurred during the year for a known duration (days). In this situation, the binomial assumption can be expanded:

$$P_{ann} = 1 - (1 - P_{\inf(no\,\min al)})^{t(no\,\min al)} \prod_{n=1}^{i} (1 - P_{\inf(n)})^{t(n)}$$

Equation 9

For example, consider a scenario when an event was known to occur in a given treatment plant for 2 days during the year. The probability of infection during that event was calculated to be 0.01. For the remainder of the year (363 days) the probability of infection was calculated to be 0.00001 (1×10^{-5}). The overall probability of one or more infections during that year was = $1- (1-0.00001)^{363} \times (1-0.01)^2 = 0.023$, if the event had not occurred the probability of one or more infections would have been = $1 - (1-0.00001)^{365} = 0.0036$.

Disease Outcomes

Infection is necessary to cause disease, however not all infections will result in symptoms of illness. While asymptomatic infections may be important for disease transmission, they do not in themselves contribute to the disease burden on a community. Evaluating the disease burden requires consideration of illness outcomes including the likelihood, severity and duration.

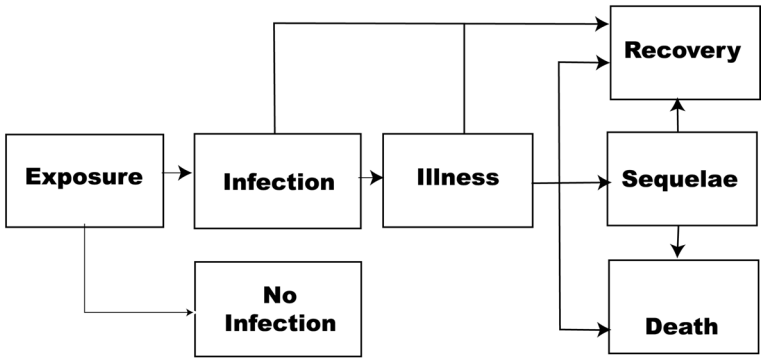


Figure 3 Outcomes of exposure to pathogens

Disability Adjusted Life Years (DALYs) is as a metric for translating the risk of disease burden a general health burden per case of illness, as discussed in Chapter 2. The DALY accounts for the years lived with a disability (YLD) plus the years of life lost (YLL) due to the hazard (compared to the average expected age of death in a community). One DALY per million people a year roughly equates to one cancer death per 100 000 in a 70 year lifetime (a benchmark often used in chemical risk assessments) [WHO, 2004]. The DALY is calculated as the product of the probability of each illness outcome with a severity factor and the duration (years). Calculation of the DALY contribution per infection is undertaken using Equation 10.

$$DALY = \sum_{i=1}^{n} P(ill \mid \text{inf}) \times P(outcome_i \mid ill) \times Duration_i \times Severity_i$$

Equation 10

Where n is the total number of outcomes considered

$P\ (ill|inf)$ is the probability of illness given infection
$P\ (outcome|ill)$ is the probability of outcome i given illness
$Duration_i$ is the duration (years) of outcome i
$Severity_i$ is the severity weighting for outcome i

The advantage of using DALYs over an infection risk end point is that it not only reflects the effects of acute end-points (e.g. diarrhoeal illness) but also the likelihood and severity of more serious disease outcomes (e.g. Guillain-Barré syndrome associated with *Campylobacter*). Disease burden per case varies widely, but can be focused on a locality. For example, the disease burden per 1000 cases of rotavirus diarrhoea is 480 DALYs in low-income regions, where child mortality frequently occurs. However, it is only 14 DALYs per 1000 cases in high-income regions, where hospital facilities are accessible to the great majority of the population. Disease burden estimates for different drinking water contaminants is summarised in Table 1.

Table 1 Summary of disease burden estimates for different drinking-water contaminants*

	Disease burden per 1000 cases		
	YLD	**YLL**	**DALY**
Cryptosporidium parvum	1.34	0.13	1.47
Campylobacter spp	3.2	1.4	4.6
STEC O157	13.8	40.9	54.7
Rotavirus			
High income countries	2.0	12	14
Low income countries	2.2	480	482
Hepatitis-A virus			
High income countries, 15-49yr	5	250	255
Low income countries	3	74	77

* Reproduced from Havelaar and Melse [2003]

While the use of DALYs has many conceptual advantages, research is necessary to facilitate its implementation. Estimates of incidence, severity and duration of disease outcomes based on epidemiologic data have only been presented in the literature for Rotavirus [Havelaar and Melse, 2003], *Campylobacter* [Havelaar *et al.*, 2000b]*, E. coli* O157 [Havelaar *et al.*, 2003] and *Cryptosporidium* [Havelaar *et al*., 2000]. These inputs for DALY calculations are extremely uncertain and the variability in severity and duration between cases is still poorly understood. Havelaar *et al.* [2000b] however, argue that for *Campylobacter* spp that the uncertainty is relatively small and that the DALY remains a robust measure even when input parameters are varied. Nevertheless, risk calculations undertaken as part of the MicroRisk project maintained probability of infection as the endpoint. The implementation of the DALY metric is demonstrated in an illustrative example only.

Example: Implementation of DALY metric for interpreting probability of infection estimates for *Cryptosporidium*.

Estimates of severity and duration of health outcomes following infection with *Cryptosporidium*, based primarily on the Global Burden of Disease (GBD) project, have been presented by Havelaar *et al.* [2000a] and reviewed by Havelaar and Melse [2003]. In developed countries 71% of immunocompetent individuals infected with *Cryptosporidium* develop gastroenteritis. The mean duration and severity weightings are summarised in Table 2.

Table 2 Summary of input assumptions for Cryptosporidium DALY calculations

Outcome	**Probability of outcome given illness**	**Duration**	**Severity**
Diarrhoea (mild)	1	7.2 (days)	0.067
Death	0.00001	13.2 (years)	1

Implementing
Equation 10 with data for *Cryptosporidium* from Table 2:

$$DALY = \underbrace{0.71 \times 1 \times 7.2/365 \times 0.067}_{Diarrhoea} + \underbrace{0.71 \times 10^{-5} \times 13.2 \times 1}_{Death} = 0.00103$$

Note: the probability of diarrhoea given illness is 1 since in the case of *Cryptosporidium* infection, all ill individuals are assumed to have diarrhoea.

The disease burden based on DALYs would therefore be calculated using the expected number of infections per year (maximising Equation 7) for the population multiplied by the DALY contribution per infection (0.00103).

Tiered approach to qmra

QMRA can be undertaken at various levels of detail, from a deterministic analysis aiming to characterise, say, worst or best case risk scenarios, to a full scale stochastic analysis. More detail is not always advantageous, but rather the QMRA scope and the perceived risk level of the system should govern what an assessor considers an appropriate level of detail. Figure 4 illustrates an iterative approach for conducting QMRA that aims to:

i) Assess the health risks associated with a water supply system;
ii) compare the estimated risks to the health targets; and
iii) if necessary, identify points in the system whereby either more data is required to better characterise the risks, or, where management strategies could best be deployed to improve the overall system performance.

The level of detail required for each iteration cannot be prescriptive; instead it will depend on the exercise scope and the available data and resources. While any level of detail or method sophistication can be employed for the first or subsequent iterations, it is advisable to begin with a simple approach, *i.e.* to conduct a screening-level assessment, the simplest approach for QMRA is to describe each model input as a point estimate. As the aim of each iteration is to identify if further consideration of the microbial risk from the system of interest is necessary, any deterministic parameter estimates should be based on the best information available at the time, and be conservative. Should the infection risk estimate be well below some health target level, the management outcome may be interpreted as that the system is performing adequately, and that current practices are adequately safe. Alternatively, it may be that even following simple analyses while the infection risk is not below the target, that management options can be identified to reduce the risk to an acceptable level. In this circumstance, further data collection or analysis may also be considered unnecessary.

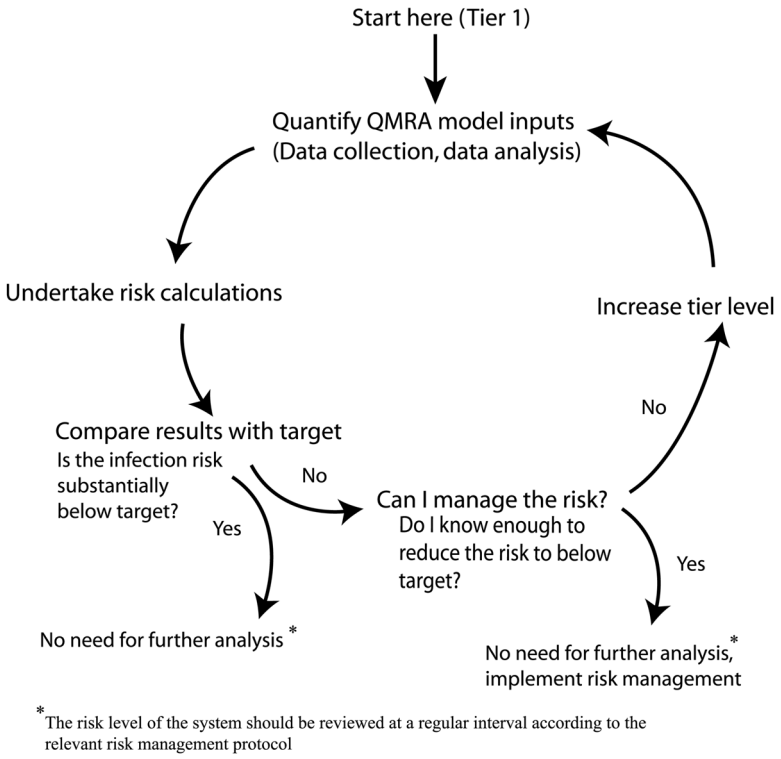


Figure 4 Iterative tiered approach for undertaking QMRA

In many situations however, an effective risk management approach is not clear unless further analysis is undertaken to characterise the variability in risk and the important determinants that drive such variability for the specific system. In this case, higher level analyses, whereby parameters are described not as point estimates, but as variable quantities may be useful.

(Higher level) probabilistic analyses

All inputs in a QMRA model are likely to vary. Source water quality varies with time dependent on catchment activity, seasonal climate changes, or specific point-source contamination such as a sewer overflow entering the waterway. Treatment efficacy varies depending on a host of factors (Chapter 4) such as plant design, treatment methods, and plant breakdown. Water consumption and susceptibility to pathogens varies between consumers. Understanding the impact of this variability on consumer risk is important, especially in management terms, as such understanding will aid answering why and how higher risk periods may occur, and provide insight into controlling those effects.

The central tool for describing variability is the Probability Density Function (PDF). When a model input is considered to be a variable rather than a constant, the input may be quantified using a PDF. When described by a PDF, the variable may take one of a range of values, each with a known probability of occurrence. The variable risk estimates may then be made using Monte Carlo simulation.

It is necessary to distinguish between the true PDF and an estimated PDF of an input variable. Each variable could be considered to have a true PDF; that is the actual frequency/duration of the range of values that the variable may take. This true PDF however, is unknown. For example, consider the *Campylobacter* density in a particular source water, which is known to be constantly changing. We may also know that it is higher at some times than others, perhaps due to season or hydrology. What we do not know is how high it can become, or for exactly how long it may be elevated.

At best, we can use available data along with some understanding of the system and formulate a PDF that is an *estimate* of the true PDF. Hopefully the estimate will encapsulate the key features of the true distribution and provide a realistic representation of the variable of interest. In order to construct an estimate of the true PDF we rely on experimental data and process or statistical models.

The manner in which a parameter is estimated and described depends on various factors, not least of all an understanding of the processes and mechanisms that may dictate the 'true' value of a parameter or the nature of its variability. The following sections (1.3 & 1.4) detail manners in which variable parameters (and the related concept of uncertainty) may be estimated using different types of relevant datasets.

QUANTIFYING VARIABILITY FROM MICROBIOLOGICAL DATASETS

Experimental data provides the most important insight into the quantitative value of each model variable. Numbers can be comforting, and tend to create an aura of certainty and accuracy; however numbers can be easily misinterpreted and inappropriately applied. Some elementary considerations when approaching a dataset for analysis include:

1. Is the dataset a random representative sample? It is important to consider if the available dataset is representative of the variable being quantified. For example: The aim may be to estimate the PDF for source water density of a particular pathogen. Were the samples taken randomly? If the dataset was collected as part of a short term study and all results were collected say during winter, or if the data was collected to investigate the impact of rainfall events and every sample was collected following rain, then the results would not represent a random sample of the source water microbial density. Conversely, if samples were collected according to an arbitrary (or randomised) time factor, unrelated to water quality processes (e.g. the first Tuesday of every month) then the sample would be assumed to be random.

2. Is the dataset representative of my system? There are countless factors that vary between water supply systems that affect microbial risk, including: catchment land use, climate, hydrology, water chemistry and treatment process performance. Direct application of published literature data, or data provided by a colleague from an apparently similar system is appealing, however processes that affect pathogen risk are complex and one system cannot necessarily be directly applied elsewhere.

3. What was the method of detection and how accurate may I expect it to be? Microbiological detection methods are constantly improving, but may differ due to changing water (matrix) effects and technician performance.

4. What is the source of the numbers being used? Reported laboratory results may have already undergone transformation from their raw state due to averaging of replicate samples or translation into concentration estimates. Translation of raw laboratory results into reported densities can ignore underlying uncertainties arising from the detection and quantification process. It is extremely relevant to understand exactly what reported numbers represent.

Characteristics of microbial data

Microbiological datasets have many unique characteristics and represent a less usual case for statistical analysis. Understanding the source of microbial data, and where uncertainties may lie in their generation and analysis is important for QMRA.

Detection and Quantification

Microbiological species are small, and present in highly variable densities in environmental samples (ranging from <1 micro-organism per L to potentially $>10^8$ micro-organism per L (e.g. for sewage)). Technical assay procedures rely on a range of approaches for developing a quantitative estimate of the micro-organism density in a water sample. Understanding the basis of quantitative density estimates is important for interpreting the inherent uncertainties associated with reported results. Hence, it is important to provide a brief description of microbiological methods employed so as to give the necessary background to understand the statistical difference between different types of microbiological data.

There are three approaches for identifying the presence of an organism in the analysed sample including:

- **Visual identification:** The presence of the organism is identified using a microscope. For example, analysis of *Cryptosporidium* and *Giardia* typically involves filtering a water sample concentrate through a membrane, staining oo/cysts then counting them with the aid of a microscope. The particular strain and infectivity status are not resolved.
- **Culture:** The most common methods of pathogen detection and enumeration rely on culturing organisms (allowing organism to multiply under favourable conditions) in the laboratory. Viable organisms are quantified by the growth of colonies or diagnostic changes in liquid media (bacteria), or by measuring their effect on established host cell lines (viruses).
- **Molecular methods:** Molecular methods are designed to detect and analyse specific genetic material unique to the group being enumerated. The genetic material is present in the sample whether or not the pathogen is infectious, and therefore routine molecular methods cannot distinguish between infectious and non-infectious organisms. Polymerase Chain Reaction (PCR) involves the specific amplification of DNA from the genome of the organism with the aid of primers. PCR can be undertaken as a non-quantitative presence/absence test, or as a semi-quantitative analysis (real-time PCR).

A combination of methods may be implemented such as the culture/ enrichment of organisms prior to PCR identification. In this case, the organisms are cultured to increase their concentration prior to PCR, therefore improving the sensitivity of the PCR identification and largely detecting viable micro-organisms.

The results of these analytical processes are translated into a quantitative estimate of micro-organism densities using:

- **Count**: A directly quantitative approach where the number of micro-organisms (*Giardia* (cysts), *Cryptosporidium* (oocysts)), plaques (viruses), or colonies are counted (bacteria). The concentration is then estimated based on the volume of original sample analysed.
- **Most Probable Number**: Results from a series of presence/absence analysis are used to predict the most probable number of organisms in the original sample based on the assumption of a Poisson distribution.

Table 3. Summary of analytical approaches and datatypes obtained from MicroRisk CTSs

	Detection method:	**Quantitative: P/A or Count**	**Reported value:**
Campylobacter spp.	Culture or enrichment + PCR	P/A	$MPN.vol^{-1}$
E. coli O157	Culture or enrichment + PCR	P/A	Estimated concentration
Norovirus	PCR	P/A	Estimated concentration
Enteroviruses	Cell Culture	P/A or count	PFU or $TCID_{50}$. L^{-1}
Cryptosporidium spp.	Visual identification	count	$Oocysts.10L^{-1}$
Giardia spp.	Visual identification	count	$Cysts.10L^{-1}$
Indicators and Surrogate Organisms			
E. coli	Plate Culture	count	$CFU.100mL^{-1}$
	Culture	P/A	$MPN.100mL^{-1}$
Clostridium perfringens	Plate Culture	count	$CFU.100mL^{-1}$
	Culture	P/A	$MPN.100mL^{-1}$

Key sources of uncertainty associated with these methods include:

- **One micro-organism or a cluster?** In many of these methods, one organism or a cluster can initiate a positive reaction. For example, when counting colonies (bacteria) or plaques (viruses) either one or a cluster of cells/virions may have contributed to each colony or plaque. When a result is reported as PFU (plaque forming units) or CFU (colony forming units) in a sample volume, this is interpreted directly as the concentration estimate. In reality, the PFU or CFU may in fact underestimate the number of cells originally present in the sample. Similarly, for a Presence/Absence MPN a positive result may have been caused by either one or a cluster of organism. The estimated MPN is based on the assumption of a Poisson distribution which only accounts for random distribution of cells in the sample rather than clustering (see Section 0). If micro-organisms are clustered, the MPN may also underestimate the original concentration.
- **Non-culturable but still infectious?** Under environmental conditions, micro-organisms may become stressed and as a result may be non-culturable in the laboratory. There is evidence however to suggest that such non-culturable organisms may still be infectious [McFeters, 1990; Barer and Harwood, 1999]. Though still controversial, the concept of cells being infectious but not

culturable has been raised for a number of the bacterial pathogens [Federighi *et al.*, 1998].

- Lastly, visual identification of pathogens, such as oo/cysts of *Cryptosporidium* or *Giardia* is fraught with additional problems to those outline above. Primarily, standard methods estimate total oo/cysts, or at best the presence of potentially infectious structures within these oo/cysts [US-EPA, 1999; Smith *et al.*, 2004). Nonetheless, differences between strains, and indeed species can be missed, if not totally miss-identified by confounding microorganisms, such as algae [Rodgers *et al.*, 1995].

Recovery and Imperfect Detection

Analytical methods for identifying microbial species in water are imperfect. Imperfect methods are evidenced by the presence of a detection limit; a value below which organisms cannot be detected. For example, consider an assay known to have a detection limit of 5 pathogens.L^{-1}. If a sample containing 3 pathogens.L^{-1} is analysed, the result will be zero as the sample density is below the limit of detection. This detection limit may be caused by:

- **Method sensitivity -** The detection limit may exist because a critical mass is necessary to perform a successful analysis. Once the detection limit is exceeded the analytical result is a direct reflection of the original organism density. Consider a sample containing 6 pathogens.L^{-1} to be analysed by a method with a detection limit of 5 pathogens.L^{-1}. Under this explanation, the expected analytical result would be 6 pathogens.L^{-1}, as the density is greater than the detection limit. This explanation may be true of many chemical analysis methods, and may also be true of microbiological analysis that aims to identify the presence or absence of the target organism in a sample volume.
- **Inactivation or loss-** A portion of original microorganisms may be inactivated or lost during the assay process. Consider again the previous example of a sample containing 6 pathogens.L^{-1}, and a detection limit of 5 pathogens.L^{-1}. The result would be expected to be 2 pathogens.L^{-1} where four pathogens may be "lost" and two detected. The inactivation or loss of organisms throughout the analytical process potentially affects all methods.

These underlying mechanisms[5] are relevant for predicting original source water density from analytical results. Many results may be interpreted as though the cause of the detection limit was method sensitivity, i.e. any value above the detection limit is assumed representative. It may be more realistic to assume, particularly with microbiological species, that the detection limit exists due to inactivation or loss, and that those "lost" organisms should be accounted for over all reported values. This loss is described by the method recovery.

Techniques for assaying microbial constituents in water samples can involve many processes and steps each of which may lead to loss or inactivation of some micro-organisms. Recovery is the portion of micro-organisms "recovered" by a particular method. If the recovery was 100%, then there would be no loss, and the analytical result would be a direct refection of the original micro-organism density.

[5] A third notable interpretation of the detection limit relates to the sample volume. The lower limit of detectable concentration is also limited by the size of sample, for example, if the sample volume was 200mL, and one organism was found the estimated concentration would be reported as 5 org.L^{-1}. Consequently, if no organisms were found, the concentration would be reported as < 5 org.L^{-1}. This interpretation does not directly relate to the recovery, but represents negative results.

Alternatively, if the recovery was say 40%, then the original density would be estimated at 60% higher than the analytical result. For example, if the analytical result predicted a *Cryptosporidium* density of 10 oocysts.L^{-1}, and the recovery of the method was 40%, then the original sample density would be estimated to be $10 \times 1/0.4 = 25$ oocysts.L^{-1}

Little has been reported regarding the recovery of *Campylobacter* and *E. coli* 0157, however recoveries of *Cryptosporidium* and *Giardia* may vary from <10% to >80% (US-EPA Method 1623) and viruses enumerated by plaque assays from ~10% to 90%. *Cryptosporidium* and *Giardia* have received the most attention in the literature regarding recovery experimentation. US-EPA methods 1622 and 1623 for enumerating oocysts and oo/cysts respectively from environmental samples have been shown to yield highly variable recoveries [Kuhn and Oshima, 2002]. Many studies have sought to quantify the dependency between sample characteristics and recovery, however the results have been inconclusive. While some studies have identified a drop in recovery at high turbidities (e.g. 159 NTU) [Kuhn and Oshima, 2002; Digiorgio *et al*., 2002] a continuous relationship is not easily defined and may not exist. Digiorgio *et al.* [2002] noted that the nature of the turbidity and the background water matrix is likely to be just as important as the absolute NTU. Consequently, there is currently no easily measurable native surrogate for estimating recovery of *Cryptosporidium* and *Giardia* in water samples; hence the recommendation of an internal control with each sample assayed.

Sampling Effects

Microbiological species consist of discrete entities or particles that cannot be assumed to be uniformly distributed throughout the water body. Rather, due to the random variation in the location of microbial particles, microbiological counts enumerated from a single well-mixed sample will rarely yield a series of identical numbers [Tillet and Lightfoot, 1995]. At low densities, the impact of sampling variability may be large.

For example, consider a volume of water containing an unknown density of *Cryptosporidium*. Suppose that the density of organisms in the volume of water is to be estimated by taking several 1L samples at random. The first sample contains 5, second 3, third 2 and finally 1. Each of these counts is an estimate of the actual mean pathogen density, illustrated in Figure 5.

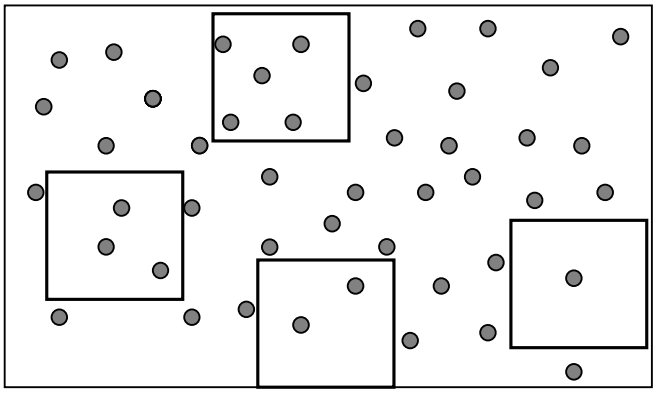


Figure 5. Random sampling of oocysts in a fixed sample volume

If the volume of water is considered well-mixed, the counts may be expected to follow a Poisson distribution [Haas *et al*., 1999]. An illustration of the Poisson count distribution with a mean (μ) of three is shown in Figure 6.

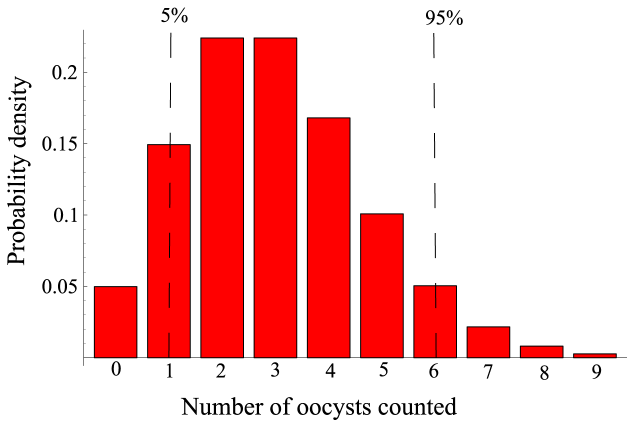


Figure 6 Poisson distribution (μ=3) highlighting 5th and 95th percentiles

In a well-mixed water body, with a mean *Cryptosporidium* density of three oocysts per litre, replicate counts would be expected to vary from 1 to 6, 90% of the time, with 10% of samples outside these limits. It is therefore possible that while the mean density is three, samples may be collected from which eight oocysts are enumerated. This expected variability has implications for interpreting a pathogen's density in source water from analytical results. Suppose the number of organisms enumerated from a 1 L sample was three, what was the actual organism density in the water body at that time? A common assumption would be to consider the result a direct measure of the mean organism density, at 3.oocysts L^{-1}. However it is evident that due to sampling variability this count could have been enumerated from waters with a much higher or lower mean organism density.
Similarly, for analysis techniques that rely on identifying the presence or absence of a target organism in the sample volume, sampling variability leads to uncertainty in interpreting analytical results. While the target organism may not have been identified in a particular sample volume, it is possible due to sampling variability that the mean density in the original sample was greater than zero.

Model Fitting and Parameter Uncertainty

Statistical models can be used to enable variability and uncertainty associated with model inputs to be quantified from microbiological data. Models are idealisations of reality that facilitate a description of the true situation. No model presents reality, however certain models are more useful descriptors than others. The aim is to choose a model that facilitates the description of the target variable for the purposes of the QMRA.
The type of statistical model selected and implemented to describe a model variable will depend on the experimental data including the type of data available (e.g. continuous or discrete; raw data or reported densities); the size of the dataset (number of data points); and perhaps the appearance (the data may appear to have come from a particular type of underlying distribution).

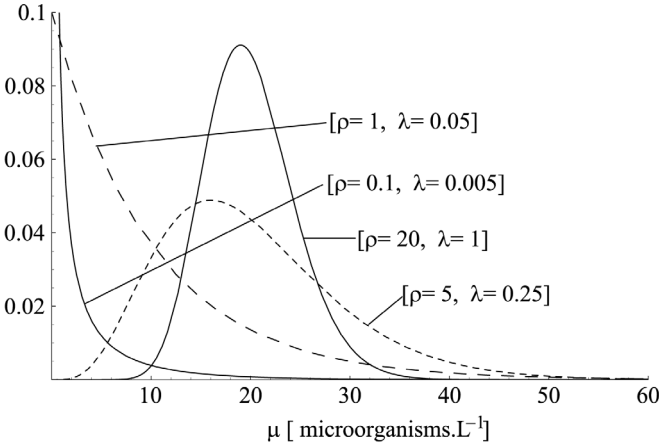
Model choice will also depend on how much is known about the process or system being studied. If the process is poorly understood, a simple empirical model that simply describes the dataset may be selected. Alternatively, if the underlying processes are well known, a model may account for the environmental, mechanistic or social processes that drive the value of the variable. Finally, the choice of model will depend on the aims of the individual risk investigation. A screening-level (tier 1) risk assessment, may intentionally select an overly simple approach.

Parametric distributions

Parametric distributions are important modelling tools for describing variability. A great number of distributions are available; however this section is limited to a description of the distributions applied as part of the MicroRisk project. In this context, the choice of distribution depended on the type of data (continuous or discrete), and the constraints (or domain) of the target variable.

Continuous data

A continuous variable can take on any value within a specified range and is not limited to discrete integer values. For continuous variables limited to positive values (such as pathogen density in source water, which cannot be negative), the Gamma distribution was applied for describing the probability density. The gamma distribution is a family of curves described by two parameters, shape (ρ) and scale (λ), of which the exponential and Chi-square distributions are special cases. The gamma distribution is particularly flexible for describing PDFs of different shapes (Figure 7). When ρ is large, the gamma distribution closely approximates the normal distribution however gamma only has density for positive numbers.



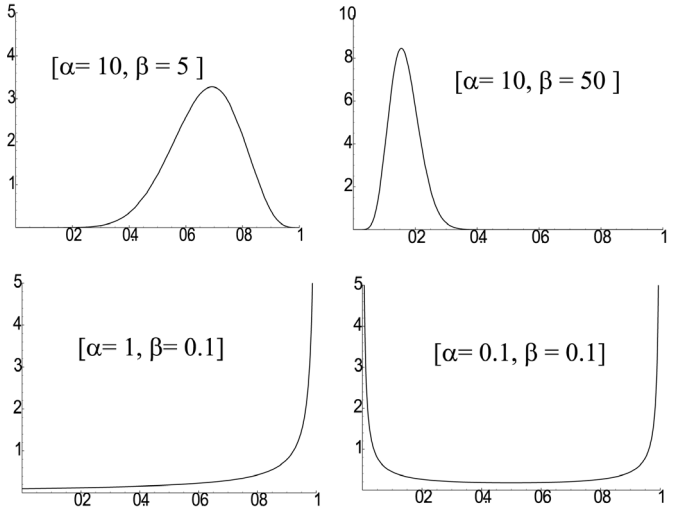
Gamma distribution:

$$g(\mu \mid \rho, \lambda) = \frac{\lambda^{\rho}}{\Gamma(\rho)} \mu^{\rho-1} e^{-\lambda\mu}$$

Where Γ represents the Euler gamma function

Figure 7. Shape of the gamma distribution for different combinations of shape and scale parameter values

For continuous variables limited to values between 0 and 1, the Beta distribution was applied for describing the probability density. The beta distribution is described by two parameters α and β and is extremely flexible for describing PDFs for binomial probabilities which will always lie between 0 and 1 (such as method recovery and probability of passage through a treatment barrier), some parameter combinations are illustrated in Figure 8.



Beta distribution:

$$g(\pi \mid \alpha, \beta) = \frac{(1-\pi)^{\beta-1}\pi^{\alpha-1}}{Beta(\alpha, \beta)}$$

Where $Beta(\alpha, \beta)$ represents the Euler beta function

Figure 8. Shape of the Beta distribution for combinations of parameter values

Discrete data

Microbial datasets frequently consist of discrete counts of micro-organisms, colonies or plaques in a certain sample volume. At low microbial densities, sampling variability should be incorporated into the statistical model structure using discrete distributions. Apart from satisfying statistical correctness, there are two practical advantages associated with accounting for sampling variability:

1. Provides greater flexibility in describing the target variable. If a particular analytical result could have eventuated from a range of source water densities, then to only consider the most likely density limits the flexibility of the statistical model to predict the most likely parameter values for the PDF; and
2. Handling zero counts. Under the assumption of a discrete counting distribution, zeros are a result with a known probability of occurrence and can therefore be directly included within the model. There is no need to substitute zero values with a less than, or a detection limit, this approach describes what actually happened.

Two types of discrete distributions are presented here for describing microbial counts: the Poisson and the negative binomial distributions.

When particles are assumed to be randomly distributed in the water body, then a series of counts enumerated from water samples may be described by a Poisson distribution. The Poisson distribution assumes that the mean density of particles is a constant value. In reality, the mean density of micro-organisms in a water body may be expected to vary both spatially and temporally. This variability in mean density implies that micro-organisms are overdispersed, rather than randomly dispersed, in the waterbody (Figure 9). When that variability is described by a gamma distribution, the result is a Poisson-gamma mixture model, which is a form of the negative binomial distribution (BOX 1). The negative binomial distribution has been widely used to describe microbial count data [Haas *et al.*, 1999; Teunis *et al*., 1999a; Teunis *et al.*, 1999b; DeVires and Hamilton, 1999; Pipes *et al.*, 1977].

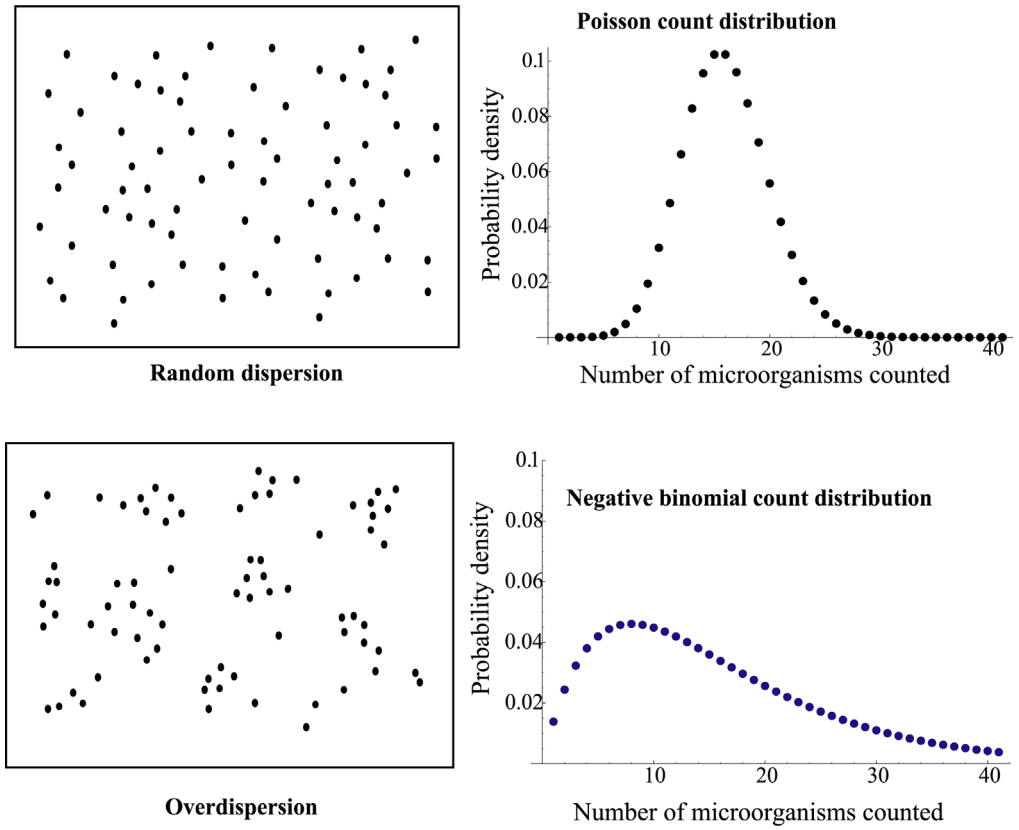


Figure 9. Illustration of the expected distribution of counts resulting from random and over-dispersion of micro-organisms.

BOX 1 - Parameterisation of the Poisson-gamma mixture model

While a range of equivalent parameterisations are available, the following description is reproduced from the work of Teunis *et al.* [1999a, b].

When counts are assumed to be generated from a Poisson (random) process, then the probability of counting *n* organisms given a mean concentration (μ) and sample volume (V) is given by:

$$P(n|V) = \frac{(\mu.V)^n e^{-(\mu.V)}}{n!}$$

Equation 11

If that mean concentration (μ) is assumed to follow a gamma distribution, then the distribution of counts (n) is given by:

$$g(n|\lambda,\rho,V) = \int \frac{(\mu.V)^n e^{-(\mu.V)}}{n!} \frac{\lambda^\rho}{\Gamma(\rho)} \mu^{\rho-1} e^{-\lambda\mu} d\mu$$

Equation 12

The solution to the integral can be rearranged into the form of the negative binomial count distribution described by gamma parameters ρ and λ:

$$g(n|\lambda,\rho,V)=\frac{\Gamma(\rho+n)}{n!\Gamma(\rho)}\frac{\lambda^{\rho}V^{n}}{(\lambda+V)^{\rho+n}}$$

Equation 13

This function can be used to construct a likelihood function based on measured counts. The maximum likelihood estimators for the gamma distribution parameters describe the variability in mean concentration μ.

Parameter estimation and uncertainty

Models used to predict and describe process variables are defined by *parameters*. Once a model has been selected, and it is hypothesised to be a useful representation of the underlying variable or system, appropriate values for the model parameters need to be estimated.

Given the experimental data (observations), the aim is to infer the parameter values of the selected distribution describing them. Several combinations of parameter values may be possible, and could have led to the observations, however the objective is to find the *most likely* parameter values, along with the probable region within which the parameter values may be expected to lie. The size of this region is reduced as the number of observations is increased. This uncertainty is referred to as *parameter uncertainty* and can be significant for small datasets.

Method of Maximum Likelihood

The concept of likelihood has been widely applied in the development of statistical models, and refers to the probability that the experimental data was generated from the assumed model [Edwards, 1992]. Construction of the likelihood function facilitates the inference of parameters values and evaluation of their uncertainty. Values of the model parameters that maximise the value of the likelihood function are termed the Maximum Likelihood Estimators (MLE), and are deemed the parameter values that are most consistent with the observations (data). It is also possible to construct a confidence region for a parameter vector based on the likelihood function[6]. For a full explanation of constructing likelihood functions and the method of maximum likelihood see a standard text, such as Montgomery and Runger [1999].

Bayesian Inference and MCMC

For complex models containing large numbers of parameters, numerical optimization of the likelihood function can be laborious. Simulation techniques using Markov

---

[6] Confidence region for parameter vector $\theta=(\theta_1,\theta_2,....\theta_k)$ consists of all parameter vector values that do not lead to rejection of the hypothesis $H_0: \theta = \theta_0$. Leading to a 100(1-α)% confidence region: $\left\{\theta,2Ln\left(\frac{L(\hat{\theta})}{L(\theta)}\right)<\chi^{2}_{[v;1-\alpha]}\right\}$

Chain Monte Carlo (MCMC) analysis, are available that allow the characteristics of the likelihood function to be explored within a Bayesian framework[7]
MCMC methods are well established and have been used for parameter estimation and uncertainty analysis in a range of modelling applications [Gilks *et al.*, 1996; Gelman *et al.*, 2004], particularly hydrology [Campbell *et al.*, 1999; Bates and Campbell, 2001]. The approach is well suited to risk assessment for evaluating uncertainty associated with models fitted to small datasets [Teunis *et al*., 1997; Teunis *et al*., 1999]. In the examples presented in this chapter, MCMC has been applied to quantify the uncertainty associated with parameter estimates. For a detailed explanation of the MCMC techniques and applications see Gilks *et al.* [1996] and Gelman *et al.* [2004].
For MicroRisk, models were constructed in Mathematica® software package (Wolfram Research, Inc.) and a Markov Chain Monte Carlo approach using the Metropolis-Hastings algorithm was used to obtain a sample of the posterior distribution for model parameters. These samples were used to construct credible intervals for the PDF (see Figure 10). The posterior sample of parameter values was used to construct a sample of PDFs, one PDF representing each sample of the parameter vector (gray lines, Figure 10). For each value of the given variable (x-axis), the lower 2.5% and upper 97.5% quantiles of the sample of PDFs were selected. These quantiles were joined, resulting in a 95% credible region for the PDF of interest. The credible region is a representation of the parameter uncertainty, and represents the region within which the PDF is expected to lie – with 95% confidence.

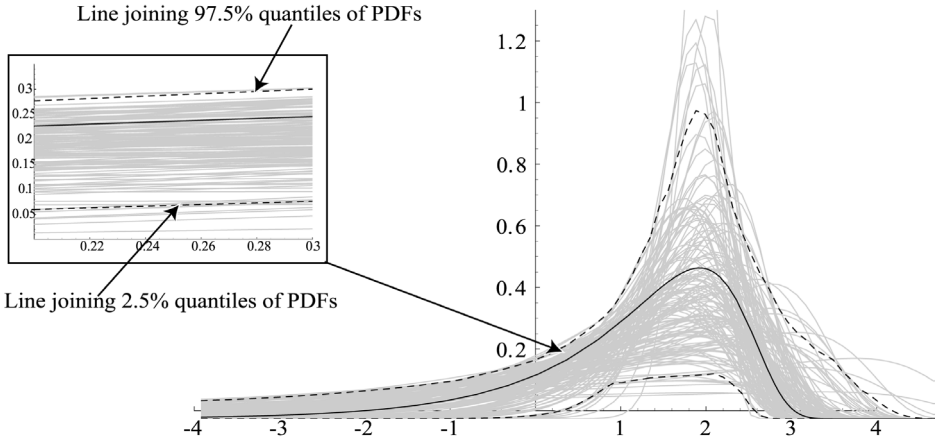


Figure 10 Illustration of method for constructing 95% credible interval (dashed line) from posterior sample of parameter pairs (PDFs constructed from posterior sample shown in grey)

[7] Statistical methods have great difficulty in determining uncertainty distributions for two or more parameters from the same data set when these distributions are correlated. Classical statistical methods either assume that the uncertainty distributions are Normally distributed, and then use a covariance matrix to create the correlation, or use resampling methods (bootstrapping). MCMC is a technique to obtain a required Bayesian posterior distribution and is particularly useful for multi-parameter models where it is difficult to algebraically define, normalise and draw from a posterior distribution. The method is based on Markov chain simulation: a technique that creates a Markov process (a type of random walk) whose stationary distribution (the distribution of the values it will take after a very large number of steps) is the required posterior distribution. The technique requires that one runs the Markov chain a sufficiently large number of steps to be close to the stationary distribution, and then record the generated values [Vose, 2004].

Triangular Distributions

For some of the MicroRisk systems examined representative pathogen data was completely lacking and attempting to predict the shape of the PDF for certain variables was considered to be inappropriate. Other forms of information were however often accessible including related literature data, MicroRisk data from other similar systems and expert opinion. It was desirable to be able to quantitatively describe this prior knowledge or expert opinion in a simple way for the purposes of a low tier assessment. While many formal approaches for incorporating prior knowledge into risk calculations are available in the statistical literature – the complexity of implementation did not fit well with the objective of undertaking a simple analysis. In these situations the triangular distribution was considered to be a useful representation of the region within which the variable may be expected to lie. The triangular distribution is defined by a minimum, most likely and maximum value, these limits of the distribution could be estimated based on general information. While the true distribution of environmental variables may never be expected to be triangularly distributed, it was considered to be useful representation of existing knowledge surrounding the value of variables at a low tier level.

Quantifying uncertainty

Once a PDF has been constructed to estimate the target variable – it is relevant to ask, how good is the estimate? How confident may I be that the estimate is a realistic representation of reality? There are many sources of uncertainty associated with predicting PDFs from experimental data. The key to accounting for uncertainty is to be precise about the uncertainty: that is, to be precise about the source of uncertainty and quantify it accordingly.
Two approaches are presented here for quantifying uncertainty. The first is to use statistical methods to account for and quantify uncertainty based on experimental data and model selection. These methods answer the question of: given my selected model, and experimental data, how much could the predictions deviate from the best estimate. These methods are presented in Section 0.
But what if the selected model is not only wrong (as all models are) but is a major misrepresentation of the system? or, What if the dataset is not-representative (perhaps all source water samples were collected under low flow conditions missing potential event driven spikes)? The second approach to quantifying uncertainty relies on expert opinion to explore the impact of these possible underlying errors or inadequacies on the overall risk estimates. Sensitivity analysis is used to investigate the sensitivity of the risk model to such underlying assumptions.

Expert opinion and worst case sensitivity

Limited datasets available for QMRA rarely tell the whole story. Expert opinion has an important role to play in interpreting environmental and risk implications of the

data available. In particular, incorporating known sources of uncertainty, even when they cannot be easily quantified is desirable. Two such sources include:

Uncertainty regarding the representativeness of experimental data: The risk analyst may consider a small dataset of pathogen counts from their water supply system. Due to some prior experience or knowledge (for example the range of pathogen densities expected given catchment sources), the analyst may question whether the dataset is in fact representative of the system. Incorporation of this uncertainty or scepticism is relevant for understanding the risk. Ignoring all prior knowledge from literature studies, other datasets or epidemiologic experience in favour of local data alone is irrational, particularly if the local dataset was extremely small, or from an unknown source. The data itself is subject to many uncertainties including influences from random sampling, and method recovery. It is therefore desirable to be able to use small local datasets for estimating the PDF, but then also to test the importance of any perceived inadequacies.

Uncertainty regarding model selection: The basis of some models may be relatively poorly understood, containing necessary but questionable simplifications. While the risk analyst may believe that the selected model is the most appropriate choice given the available data and understanding of the system, they may also be interested to test the importance of this model choice on the calculated probability of infection. The selection of a second possible model may lead to much higher probability of infection, highlighting the need to consider carefully which model is chosen, and perhaps the need for further data collection to understand which model is likely to be more representative.
In order to provide a quantitative framework for the consideration of these uncertainties, a pragmatic approach, using a sensitivity analysis calculation was proposed.
In a model that contains a series of steps, sensitivity analysis may be used to identify which components or variables within the model are most important to the outcome. Sensitivity analysis allows for the effect of changing assumptions to be assessed and is a valuable tool for determining the critical drivers of microbial risk within the system. Using sensitivity analysis tools, uncertainties can be evaluated for the purpose of prioritising data collection and research. Methods for undertaking sensitivity analysis have been reviewed by Frey and Patil [2002]. In that article, sensitivity analysis methods were categorized into three groups: mathematical, statistical or graphical. The method adopted to evaluate the sensitivity of the model to uncertainty in variable estimation was the worst case sensitivity. This method was presented by Zwietering and van Gerwen [2000] in the context of food safety and risk assessment.

Worst case sensitivity: The importance of uncertainty in each model component may be evaluated by calculating the factor sensitivity at each step. The factor sensitivity compares the impact of worst case, or extreme assumptions relative to the average.

$$FS_K = Log\left(\frac{N_k(extreme)}{N_k(average)}\right)$$

Equation 14

The risk model was initially constructed and simulated using best-estimates of all model variables. The results from this analysis were used to find the dose under "average" conditions (N_k (*average*)). A worst-case value was then selected for each of the model variables. Keeping all other variables at their average or "best" estimates, the model was simulated to find the dose under "extreme" conditions (N_k (*extreme*)), with reference to each individual variable. The resulting factor sensitivity for each step indicates the relative importance of uncertainties associated with each model variable. Given that the dose under average and extreme conditions is described by a PDF, the FS is also represented by a PDF, and was calculated for both the average and the 95th percentile of the PDF.

Selection of worst case value: For each model variable, a "worst case" value was selected, the basis of which depended on the particular variable in question and the perceived uncertainty associated with the estimation of that variable. Sources included:

1. **Parameter uncertainty:** There is uncertainty associated with the parameter values fitted to local datasets. For small datasets, this uncertainty can be significant. The worst case value was selected as the conservative (i.e. for source water concentration the upper, for treatment performance the lower) 95% credible limit of the estimated PDF.
2. **Data from another system or literature** – Perhaps the *Cryptosporidium* density in source waters was estimated from a small experimental dataset to be 0.001 oocysts. L^{-1}. However results from another similar catchment indicate densities closer to 2 oocysts.L^{-1}. This higher density of 2 could be adopted as a worst case value, to test the sensitivity of the risk model to the uncertainty in source water *Cryptosporidium* density. If found to be important, further investigation of the source waters may then be justified.
3. **Event impact** – Particularly for treatment performance, the impact of loss of a treatment barrier in the process was estimated by assuming that removal performance was zero as an extreme value.

The aim of the framework is to allow the sensitivity of any assumed variable value to be tested. Whether selected arbitrarily or as a set percentile from parameter uncertainty, the influence of any hypothesis or assumption can be tested in a structure way allowing for uncertainties that are important to the risk outcomes to be identified and prioritised.

Auditing Score

Each model input, for each system studied was quantified with a varying degree of precision and complexity. This variation depended on the quantity and quality of data available, the importance of the model input to the overall quantitative risk outcomes and the tier level of the investigation (Figure 4). The level of knowledge and uncertainty associated with quantifying each model input needs to be weighed in the light of the QMRA outcomes. For example, consider the case where a system's source water pathogen content was estimated from measured pathogens in the specific source water, while removal by sedimentation was based on observed removals in particle count data at the treatment plant, and removal by filtration was estimated based on removals of an indicator organism at a similar plant elsewhere. A

sensitivity analysis result might imply that the source water content of pathogens most heavily dictates the risks, and so should be the first point for a water manager to begin implementing management resources. However, there is greater unquantified uncertainty regarding how representative the estimates of removal by sedimentation, and removal by filtration were compared to the true values for the system of interest. A framework for testing the importance of these uncertainties on the risk outcomes has been presented (see section 0), however there is also a need to *evaluate* and *document* the level of detail and confidence associated with each model input, alongside the QMRA calculations. The need to consider and communicate such information suitability uncertainties for QMRAs has been raised before (*e.g.* Fewtrell, *et al*. [2000]).

To facilitate this auditing process in the MicroRisk project, each variable in the QMRA model was given an audit score as described in (Appendix 2, Chapter 8). When determining data needs for future iterations of the QMRA models, consideration needs to be given both to the quantified assessment results as well as the data quality audit scores.

Implementation for CTS's

Ideally, each variable should be described by a PDF, with the resulting variability in exposure characterized by Monte Carlo simulation. Variability and uncertainty are inevitably woven together, however at least conceptually, they were separated for this risk analysis. PDFs were limited to the description of variability. Second order analysis (where PDFs are also used to describe uncertainty) was not undertaken. This was a deliberate choice, given the magnitude of variability, and the limited datasets available (leading to high uncertainty). An exposure characterisation that incorporates all variability and uncertainty was considered to be so broad as to limit its practical application.

As an alternative, PDFs for model variables were estimated based on relevant data and parameter uncertainty associated with those PDFs was predicted. The importance of uncertainty due to model assumptions and adequacy of experimental data was evaluated within a more pragmatic framework. The sensitivity of the risk model to uncertainty associated with each variable was estimated by choosing a worst case value (i.e. how high/low could an expert realistically expect this variable to be) and calculating the Factor Sensitivity.

The resulting risk characterisation, therefore reflects variability in model inputs, described using the best available data. Factor sensitivity results identify the most important sources of uncertainty in terms of risk outcomes, therefore identifying where additional information is required to improve risk predictions.

The following sections outline specific techniques applied for estimating PDFs from microbiological data for source water concentration, recovery and treatment performance; distribution and consumption are not included in this section since results from Chapters 5 and 6 were directly applied for the risk calculations in Chapter 8. Published dose-response parameter values are included along with some discussion of the influence of model choice on the estimated probability of infection estimates.

Source water pathogen densities

Pathogens may be present in the source water due to human or animal inputs (waterborne enteric viruses being assumed to only come from human excreta). The density of pathogens at the treatment plant off-take is dependent upon the magnitude of pathogen inputs and the environmental processes affecting the transport and inactivation; and is expected to vary both over time and between pathogen groups. Literature and Chapter 3 data was used to describe pathogen densities at the off-take for each studied system. With few exceptions, the Gamma distribution was selected for describing variability in source water pathogen density due to its flexibility. The modelling approach adopted for fitting a PDF for source water density depended upon the type of experimental data provided.

Microbial Counts

Consider *Giardia* counts from the raw water source for CTS 7 (shown in Table 4). Direct conversion of these counts to concentrations (i.e. number cysts counted/Volume = cysts.L^{-1}) leads to a mean cyst concentration of 0.117 cysts. L^{-1} with a maximum concentration of 0.97 cysts L^{-1}. Describing these counts directly as densities ignores the influence of sampling variability (the mean bulk water density at the time the sample was taken is assumed to be exactly equal to No. cysts/Volume), and necessitates the substitution of zero counts with some positive value. To obtain a more realistic picture of the source water concentration, these data should be analysed as a discrete dataset using counting statistics (relying on each raw count in the measured sample volume) rather than a continuous distribution. Assuming that these discrete counts reflect random samples (Poisson process) from the source water with mean *Giardia* concentration (μ), and that the mean concentration varies according to a gamma distribution leads to a negative binomial count distribution (Equation 13). When the negative binomial distribution was fitted to the counts and volumes (Equation 3), the maximum likelihood gamma distribution describing *Giardia* cyst concentration in source water for CTS 7 was found and is illustrated in Figure 11.

Table 4. *Giardia* cysts counts from CTS7 source water

Count	Volume (L)
8	16.25
9	9.25
8	65
7	67.5
9	92.5
1	110
3	130.75
4	134
5	105
2	76.25
0	137.5
3	125
2	125
2	125
0	125
0	125
1	112.25

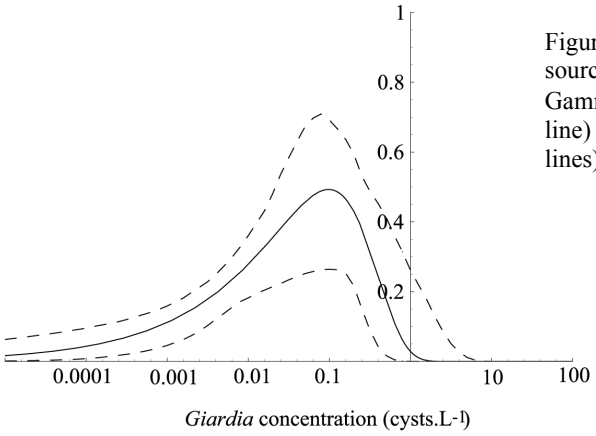


Figure 11. PDF for the *Giardia* cyst density in the source water for CTS 7 – Maximum likelihood Gamma distribution $\rho = 0.41$ and $\lambda = 0.24$ (solid line) and 95% Bayesian credible intervals (dashed lines) constructed from posterior MCMC samples.

Parameter uncertainty was explored by constructing a sample of the Bayesian posterior distribution of ρ and λ using MCMC simulation with uninformative priors. The variable *Giardia* cyst concentration in source water for CTS7 was included in the risk model as a gamma distribution defined by maximum likelihood values of $\rho = 0.41$ and $\lambda = 0.24$. While the upper 95% of the best fit PDF was 1.12 cysts.L^{-1}, the uncertainty analysis indicates that given the data, the mean *Giardia* cyst concentration could reach concentrations as high as 10 cysts.L^{-1} (upper credible interval).

Incorporation of Events

For some source waters, elevated concentrations of pathogens may be directly linked to events that mobilise pathogens in the catchment such as rainfall induced runoff, or sewage discharges. In these situations, rather than fitting one distribution to all data points, it may be more representative to describe the source water pathogen concentration separately for event and nominal (baseline) conditions.
For the *Giardia* dataset from CTS 7, elevated pathogen concentrations were hypothesised to be associated with periods following a rapid rise in water level in the source river. To investigate this hypothesis, when the samples were collected, the operator identified whether the conditions were classified as “event” or “nominal”. The same dataset from Table 4 is categorised as event and nominal in Table 5.

Table 5. *Giardia* cyst counts from CTS7 source water. Samples classified as "event affected" and "nominal"

Event affected	
Count	Analysed Volume (L)
8	16.25
9	9.25
8	65
7	67.5
9	92.5
1	110
3	130.75
4	134
5	105
2	76.25
Nominal	
Count	Analysed Volume (L)
0	137.5
3	125
2	125
2	125
0	125
0	125
1	122.25

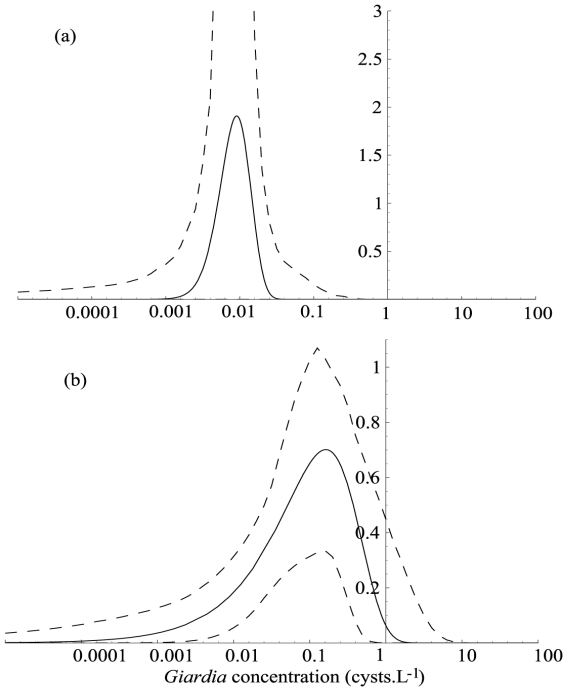


Figure 12 PDF for variability in the mean *Giardia* density (μ) under nominal (a) [$\hat{\lambda}$ =0.002 , $\hat{\rho}$ =4.48] and event (b) [$\hat{\lambda}$ = 0.22, $\hat{\rho}$ = 0.72] conditions. Maximum likelihood gamma distribution (solid line), with 95% credible intervals from MCMC analysis (dashed lines).

The aim of the analysis was firstly to determine if there was a significant difference between the *Giardia* density under event and nominal conditions; and secondly if appropriate, to estimate the PDF for each condition. The maximum likelihood estimates of the parameter values and deviance were calculated for each separate condition and the pooled (combined) dataset. A comparison of the deviance indicated that there was a significant improvement in fit achieved by separating the datasets (Combined (101.23) - (Nominal (20.33) + Event (68.34)) = 12.56 > 5.991, Chi squared distribution at 95% level with 2 degrees of freedom), and describing the PDF for nominal and event conditions separately. The PDFs for nominal and event conditions are illustrated in Figure 12 with their credible intervals.
The expected value of the *Giardia* density under event conditions was 0.16 cysts.L^{-1} in comparison to 0.009 cysts.L^{-1} under nominal conditions. *Giardia* density was clearly higher during events.

The separation of data for describing event and nominal conditions affected the spread of the distribution, and hence the assumed variability associated with the mean density. Firstly, when counts measured under event conditions were removed, the baseline or nominal dataset showed very little variability and was equally well fit by the Poisson distribution (Deviance = 0.11 < Chi-Square at 95% level with 1 degree of freedom = 3.841), indicating no evidence in the nominal data for variability in the mean density (μ). Under nominal conditions, the source water could be assumed constant at 0.009 cysts.L^{-1} (The upper 95% quantile of the MCMC posterior sample of the Poisson parameter μ was 0.016 cyst.L^{-1}).
Secondly, predictions of upper concentration values were reduced by considering event conditions separately. Comparison of the upper 95 quantile of the distribution for event conditions (0.69 cysts.L^{-1}) with the upper 95 quantile of the pooled dataset (Table 4 and Figure 11) (1.12 cysts.L^{-1}) demonstrates that the assumed peaks in concentration are in this case reduced by considering events separately. Similarly, the upper credible interval of parameter uncertainty was lower for the event only dataset (1.57 cysts.L^{-1}) in comparison to the pooled dataset, where the upper credible interval for the upper 95 quantile approached 10 cysts.L^{-1}.
Exposure to pathogens could be calculated in the risk model for event and nominal conditions separately, in which case the relative importance of each condition could be examined during risk characterisation. Alternatively, the two distributions could be combined to form one single PDF for source water *Giardia* concentration.
Combining the two PDFs for a single model input requires a representation of the proportion of time that the water quality is represented by each condition. If the dataset itself was representative, the river would be under event conditions 10/17= 58.8% of the time. More accurate data relating to the cause of events would be expected to exist, that would provide a better estimate of this parameter [Signor *et al.*, 2005 application to CTS 8]. The overall PDF for source water density would be given by Equation 15:

$$\mu_{Sourcewater} = A \times \mu_{event} + (1 - A) \times \mu_{no\,\min al}$$

Equation 15

Where:
A is the proportion of time that source water is under event conditions
μ_{event} is the *Giardia* concentration under event conditions
$\mu_{nominal}$ is the *Giardia* concentration under nominal conditions
Given the analysis of the *Giardia* count data, the best estimate for implementing Equation 15 and describing source water density of *Giardia* at CTS 7 is: $\mu_{Sourcewater} = 0.588 \times GammaDistribution[\rho = 0.72, \lambda = 0.22] + 0.422 \times 0.009$.

Presence/Absence results

For many pathogens, the number of pathogens present cannot be directly identified and analytical methods are limited to identifying the presence or absence of the target organism in a sample volume. A presence/absence approach can however be used quantitatively when several replicate samples at different dilutions are analysed in parallel.
One such organism that is analysed for presence or absence is *E. coli* O157. Sampling and analytical procedure for quantifying *E. coli* O157 concentration in source water

for CTS 10 included 15 samples, each of which was sub-sampled at three ten-fold dilutions. The presence or absence of *E. coli* O157 was identified in each sub-sample. The results, along with reported density estimates are included in Table 6.

Table 6. Presence/Absence results for *E. coli* O157 from CTS 10 source water samples

Volume (L)			*Estimate org.L^{-1}
0.01	0.1	1	
0	0	0	<1
0	0	0	<1
0	0	1	1-10
0	0	0	<1
0	0	0	<1
0	0	0	<1
0	0	0	<1
0	0	0	<1
0	0	1	1-10
1	1	1	>100
0	1	1	10-100
1	1	1	>100
1	1	1	>100
0	0	1	1-10
0	0	0	<1

*Without taking into account recovery of the method

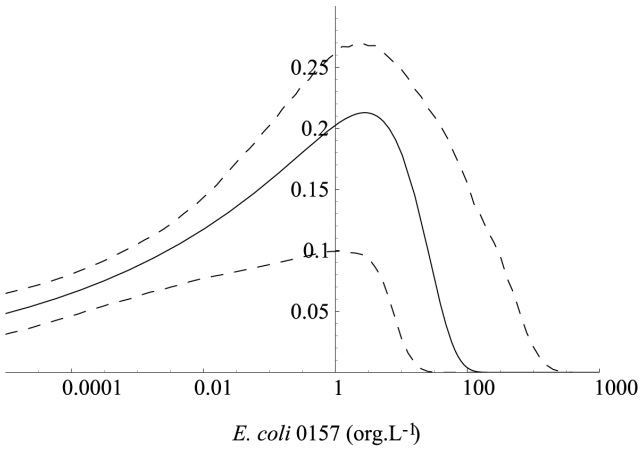


Figure 13 PDF for *E. coli* O157 density based on presence/absence data from CTS10. Maximum posterior gamma distribution (solid line) with 95% credible intervals (dashed line) from MCMC analysis.

The estimated concentrations give some idea of the expected range of how many *E. coli* O157 may have been present in the source water; however quantifying the PDF of *E. coli* O157 concentration for input into the risk model is more complicated. A statistical approach is required that allows the shape of the PDF to be estimated (including a realistic representation of the parameter uncertainty) based on the presence/absence results.

In order to undertake this analysis, some relatively simple assumptions regarding the underlying processes influencing the pathogen density were made:

1. The three sub-samples for each sampling day were assumed to be random samples (Poisson process,) from the source water with mean *E. coli* O157 concentration μ.
2. The mean density (μ) was assumed to vary between sampling occasions according to a gamma distribution.

The model is no-longer a straight forward Poisson-gamma mixed model but rather a special case where gamma dispersion is only assumed between sampling days; on any individual sampling day, the dispersion between sub-samples is assumed random (Poisson). Implementing these assumptions by constructing a likelihood function is

mathematically complex, however when the model is constructed within a Bayesian hierarchical framework, the calculations are simplified.

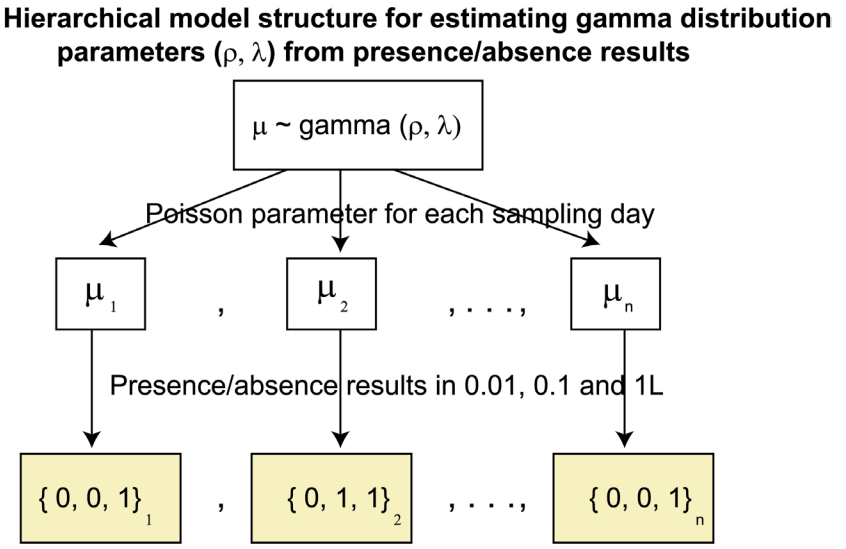


Figure 14 Structure of the hierarchical model for estimating gamma distribution parameters from Presence/Absence results

The posterior distribution of ρ and λ can be investigated by simulation using MCMC. The maximum posterior gamma distribution with 95% credible intervals for the PDF of *E. coli* O157 density illustrated in Figure 13. Sampling days that consisted of all negative results led to a lack of convergence in the posterior samples of ρ and λ. To avoid this problem, sampling days where all results were negative, were modelled as half the detection limit, rather than zero. This was achieved by substituting a positive result for the 1L sample volume, and estimating μ as half the predicted value for that day. The expected value of the *E. coli* O157 density was 2.78 org.L^{-1} with an upper 95% quantile of variability of 15.73 org.L^{-1}. The upper 95% quantile of parameter uncertainty was 134.8 org.L^{-1}. The clear benefit of this approach is that the shape of the PDF, along with associated uncertainty can be estimated directly from the presence/absence results.

Index Organisms

Index organisms are microbial species that are present in water samples at a known ratio to one or more human pathogens [Ashbolt *et al.*, 2001]. Not only do index organisms indicate the presence of human pathogens, but they can be used to quantitatively estimate the concentration of a particular pathogen using the ratio between their densities. In order to be useful as an index organism, a microbial species should be from the same source as the human pathogen, and respond similarly to environmental conditions.

Given the complexity and hence expense associated with analysing directly for pathogens in source water samples, the use of index organisms for quantifying source water pathogen concentration is desirable.

Within the MicroRisk project, *E. coli* and thermotolerant coliforms (Coli 44C) have been used as an index for quantifying pathogen densities in surface waters and sewage from reported concentrations (Chapter 5). These estimated pathogen to *E. coli* or thermotolerant coliform ratios were used to predict possible pathogen densities in

distribution systems based on *E. coli* measurements. The data used to estimate these ratios are summarised in Figures 15 & 16 for surface waters and sewage respectively. The illustrated PDFs were constructed by fitting a gamma distribution separately to the sample of pathogen and index densities, and then calculating the PDF for the ratio between the two gamma distributions. The estimated ratios based on paired index and pathogen reported densities are also shown for each figure.

The illustrations demonstrate the variability in the estimated ratios based on reported concentrations, spanning several orders of magnitude for all organisms. For some organisms such as *Campylobacter* in sewage, the ratio with *E. coli* varied by nearly 6 orders of magnitude. The uncertainty associated with the estimated ratios was not possible to capture since the underlying (raw) data was not available for analysis.

While the theory behind the use of index organisms is attractive, the practical application is subject to both the existence and quantitative description of the ratio between the particular index organism and pathogen under consideration. It may be that to assume that a ratio exists at all between *E. coli* or thermotolerant coliforms and pathogens is erroneous. Nonetheless, if these assumptions are applied for the purposes quantifying pathogen densities, then the variability and uncertainty associated with the estimated ratio need to be incorporated into the calculations. In particular, the implications of significantly underestimating pathogen densitie should be thoroughly explored.

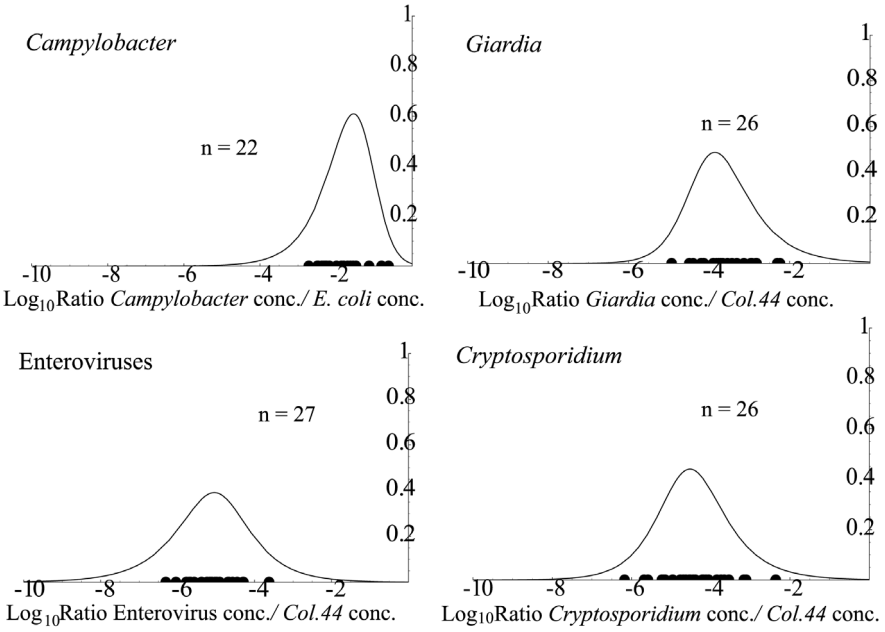


Figure 15. PDFs for the ratio between maximum likelihood gamma distributions for index organisms (*E. coli* and Coliforms at 44°C) and pathogen concentration, based on reported concentrations in surface waters (*n* is the number of paired concentrations). Data source: Medema *et al.* [2000]

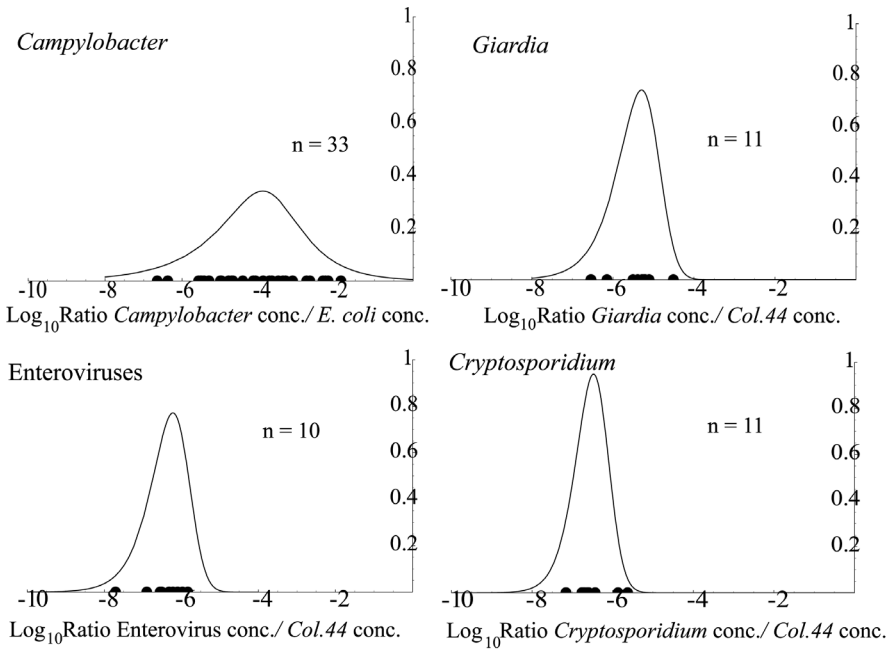


Figure 16. PDFs for the ratio between maximum likelihood gamma distributions for index organisms (*E. coli* and Coliforms at 44°C) and pathogen concentration, based on reported concentrations in sewage (*n* is the number of paired concentrations) Data source: *Campylobacter* [Hőller,1988], *Giardia, Cryptosporidium* and Enteroviruses [Medema *et al.*, 2000]

Quantifying Method Recovery

Analytical methods are imperfect, not all organisms present in the original sample may be recovered and enumerated in the laboratory. The result of a microbiological analysis is therefore a reflection of the number of identifiable micro-organisms present at the *conclusion* of the assay method. Interpretation of the original sample density from the assay results requires a quantitative understanding of the method recovery. Assuming that analytical results are precise representations of the original organism numbers may significantly underestimate the density. In addition, unaccounted for variability in method recovery may lead to apparent high variability in micro-organism density that is actually a product of the analytical process rather than of the original water quality.

Quantifying the magnitude and variability of recovery is important for interpreting analytical results. Experiments specifically designed to estimate recovery involve spiking a known number of micro-organisms into a sample volume that is subsequently analysed using the relevant protocol. Results of such experiments indicate that recovery varies between micro-organisms, between analytical methods and between laboratories. Recovery may also be expected to vary between subsequent samples even when analysis is undertaken using the same method at a single laboratory. There is also uncertainty that the spiked microbial preparation behaves the same as ‘native’ microorganisms or how difference in the water matrix between laboratory and natural samples influence the method recovery.

Microbial counts

In conventional seeding procedures for *Cryptosporidium* and *Giardia*, one sample is typically split into two for analysis. One sample is spiked with a known number of oo/cysts, while the other is unseeded. Statistical models for analysing this type of recovery data have been published [Teunis *et al.*, 1999a]. A modification to this approach has been developed [Francey *et al.*, 2004] involving the use of labelled oo/cysts known as ColorSeed™ (BTF Pty. Ltd., Sydney) that are spiked into a single sample, enabling the analyst to estimate the number of recovered seeded organisms (due to a unique colour), along with the native organisms from a single sample. Every sample analysed for *Cryptosporidium* and *Giardia* from the CTS 8 source water reservoir received a ColorSeed™ internal spike. The recovered ColorSeed™ counts (100 ±1 oocysts seeded from 110 samples analysed for *Cryptosporidium* are illustrated in Figure 17.

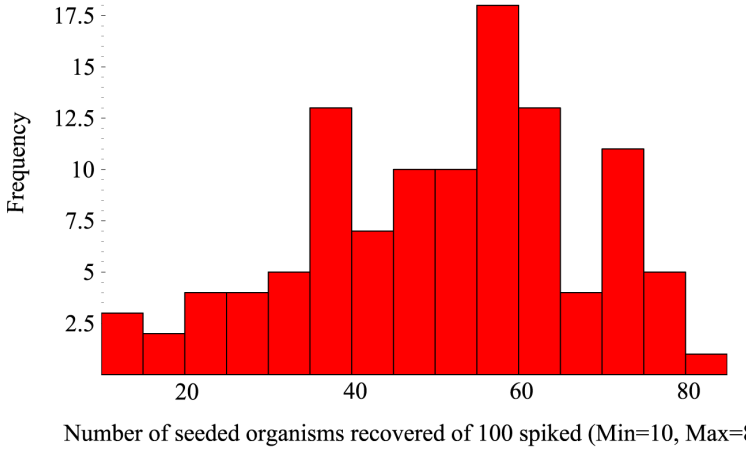


Figure 17 Histogram of the number of seeded oocysts recovered (from 100 spiked) for *Cryptosporidium* samples from raw river water at CTC 1

The variability in recovery between samples is clearly evident. These results are from the same source water, analysed at the same laboratory, using the same experimental protocol, and yet the variability is still high. The great advantage of the ColorSeed™ internal spike, is that a sample specific estimate of the recovery is obtained for each native count. Factors that drive the variability in recovery are still largely unknown, and therefore the internal spike reduces the uncertainty associated with the unknown influence of sample characteristics (including turbidity, temperature and pH) on the estimated recovery.

When an internal spike result is available for each native count, the recovery can be accounted for directly when estimating the PDF for source water concentration. If the recovery is assumed to be a binomial process (each organism may have one of two outcomes – it will be recovered or not recovered), where every organism has a certain probability (p) of being recovered. The number of spiked organisms recovered is an estimator of the probability of recovery (p). In this example, the probability of recovery on each sampling day was assumed to be independent of other sampling occasions.

Native counts were assumed to follow a negative binomial distribution and a likelihood function was constructed to account for the binomial probability of recovery (BOX 2).

BOX 2- Incorporating recovery into the negative binomial count distribution

If the counts are assumed to be generated from a Poisson process, with a probability of detection (*p)*, then the probability of counting *n* organisms given a sample concentration (μ) is:

$$P(n|\mu, p) = \frac{(\mu.p)^n e^{-(\mu.p)}}{n!}$$

Equation 16

When the mean source water concentration (μ) is assumed to follow a gamma distribution. The solution can be rearranged into the form of the negative binomial count distribution:

$$g(n|\lambda, \rho, p) = \frac{\Gamma(\rho + n)}{n!\Gamma(\rho)} \frac{\lambda^\rho p.^n}{(\lambda + p)^{\rho+n}}$$

Equation 17

If the number of recovered organisms k_i is assumed to be a precise estimate of the probability of recovery, then the likelihood function may be constructed as:

$$L(\lambda, \rho \mid n_{1-m}, k_{1-m}) = \prod_{i=1}^{n} g(n_i \mid \lambda, \rho, \frac{k_i}{100})$$

Equation 18

Allowing for uncertainty in estimation of binomial parameter (*p*)

The uncertainty associated with estimating *p* from the number of recovered organisms may be incorporated into the model using a hierarchical structure. Within a Bayesian framework, the posterior distribution of λ and ρ is proportional to the likelihood multiplied by the prior:

$$P(\lambda, \rho \mid n_{1-m}, k_{1-m}) \propto \underbrace{P(\lambda, \rho, p_{1-m})}_{\Pr ior}.\underbrace{\prod_{i=1}^{m} P(\lambda, \rho \mid n_i, p_i)}_{Negativebinomal}.\underbrace{\prod_{i=1}^{m} P(p_i \mid k_i)}_{Binomial}$$

Equation 19

Where the binomial likelihood is described by: $l(p;k) = \binom{100}{k} p^k (1-p)^{100-k}$

An MCMC procedure then allows the stationary posterior distribution of λ and ρ to be characterised.

The maximum likelihood gamma distribution (expected value = 2.02 oocysts.L^{-1} and upper 95 percentile (variability) = 8.59 oocysts.L^{-1}) and credible intervals from

MCMC (upper 95 percentile = 11.57 oocysts.L^{-1}) for the source water *Cryptosporidium* concentration CTS 8 is illustrated in Figure 18. The size of the dataset and the incorporation of recovery estimates for each day leads to a small credible interval surrounding the maximum likelihood gamma distribution.

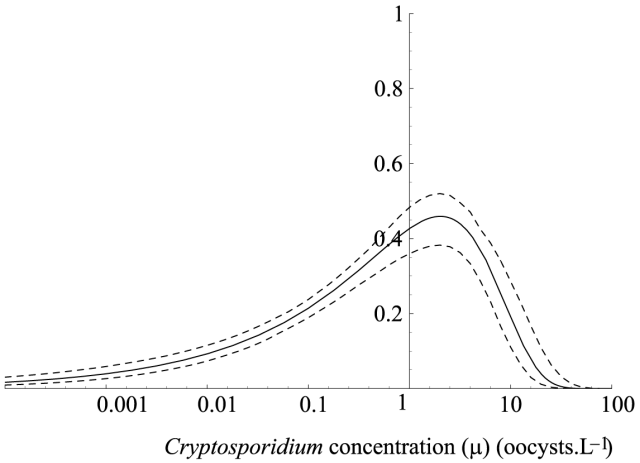


Figure 18. PDF for *Cryptosporidium* oocyst density in the raw river water for CTC 1: accounting for method recovery. Maximum likelihood gamma distribution (solid line), and 95% credible intervals (dashed lines) from MCMC modelling.

Internal spike material appears to be the most appropriate approach for estimating sample specific recovery for *Cryptosporidium* and *Giardia* oo/cysts. Unfortunately, this analysis was not provided with the dataset from any of the other CTSs studied. At best, laboratories provided a sample of recovery results believed to be representative for the entire dataset. In this case, recovery may be included as a variable in the risk model, described by a PDF [Teunis *et al.*, 1996]. In comparison to the 110 recovery results provided for CTS 8, three recovery experiments were undertaken for CTS 11. These experiments were also undertaken using the ColorSeed™ internal spike, however as they were not undertaken for every sample, the direct sample specific recoveries cannot be applied to the native results in the same manner as the previous example for CTS 8. Rather, by fitting a distribution to the recovery results, recovery may be included in the risk model as an independent variable. The results of these experiments are included in Table 7. Based on these results, the average recovery for *Cryptosporidium* oocysts may be expected to be 12%. This would be an appropriate point estimate assumption based on this data, however a point estimate does not allow for variability in the recovery between samples.

Table 7. *Cryptosporidium* oocyst recovery results CTS 11

***Cryptosporidium* oocysts**	
Number Spiked	Number Recovered
100	12
100	10
100	14

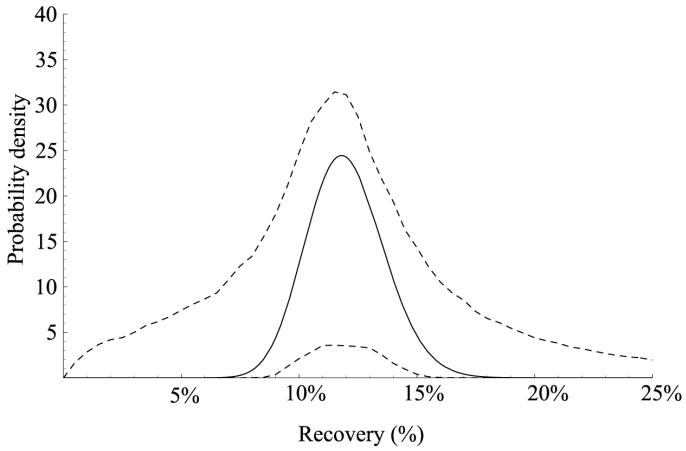


Figure 19. PDF for *Cryptosporidium* oocyst recovery with Maximum likelihood Beta Distribution(α= 47.1, β=345.43) (solid line), and 95% credible intervals from MCMC analysis.

To incorporate variability of recovery in the QMRA model, the probability distribution for recovery needs to be estimated. The variability in recovery has previously been described using a Beta distribution [Teunis *et al.*, 1999a, Teunis *et al.*, 1996] which is considered suitable as it is flexible and bound by 0 and 1. The Beta distribution was fitted to the data points (Table 7) using the method of maximum likelihood to obtain a best estimate of the PDF for recovery. Figure 19 show the shape of the beta distribution with the maximum likelihood estimates for α and β, and the 95% credible region constructed using an MCMC approach. Additional uncertainty is introduced when the recovery is incorporated into the model as an independent variable. Firstly, when the range of between sample variability is applied to every result, the variability and uncertainty associated with the source density may be expected to be increased. Secondly, there is danger that a small sample from a highly variable recovery may lead to an unrepresentative PDF. Comparison of the results in Table 7 with the variability in results illustrated in Figure 17, along with consideration of recoveries reported in the literature [Kuhn and Oshima, 2002], suggest that the three datapoints from CTS 11 may not be representative of the entire distribution of recovery. The reported recoveries are however low, leading to conservative estimates of source water density when the distribution in Figure 19 is applied to the native counts.

An important consideration of this approach for describing variability is the impact of fitting a continuous distribution to the experimental datapoints. The beta distribution

projects to very low values close to 0. The true recovery would not however be expected to approach zero, but rather there would be minimum result below which, samples would be disregarded (based on laboratory QA protocols). When running the Monte Carlo simulation for QMRA, these unrealistic, very low values lead to occasional excessively high pathogen densities. These high pathogen densities are not considered to be representative of the system, but rather a consequence of the model assumptions. The parametric distributions are a tool to enable the estimation of the true PDF, and when they project beyond the realistic range of values, it is pragmatic to truncate them. It is therefore advisable to select a lower value at which to truncate the PDF for recovery, such as 1%, to avoid the generation of unrealistic values.

Presence/Absence results

Quantifying method recovery is also important for analytical methods that rely on identifying the presence/absence of the target organism in a sample volume. The assumptions of the previous examples involving microbial counts where recovery was assumed to be a binomial process may be extended to the presence/absence scenario. A specific experiment was used to investigate the recovery of *E. coli* O157 at one CTS [Suez Environnement, 2005]. This investigation consisted of three separate spiked solutions of known density (BioBall^{TM}). Each of these three solutions was sub-sampled (1L) and analysed 10 times for the presence or absence of the *E. coli* O157:H7. The results are included in
Table 0.8.

Assuming that the spiked solution has a known concentration (μ), then the probability of the analysis yielding a positive ($n \geq 1$) or negative ($n = 0$) will follow a Poisson distribution with mean ($p\mu$, Equation 16) - where p is the binomial probability of recovery. A likelihood function was constructed based on the analytical outcomes (Table 0.8) to estimate the constant probability of recovery (p). The maximum likelihood estimator for the probability of recovery was 0.4, and the posterior sample for p based on MCMC sampling is illustrated in Figure 20.

Table 0.8. Results from *E. coli* 0157 detection limit experiment [Suez Environment, 2005]

Spike (org.L^{-1})	**No. Pos (+)**	**No. Neg (-)**
1	2	8
5	9	1
10	10	0

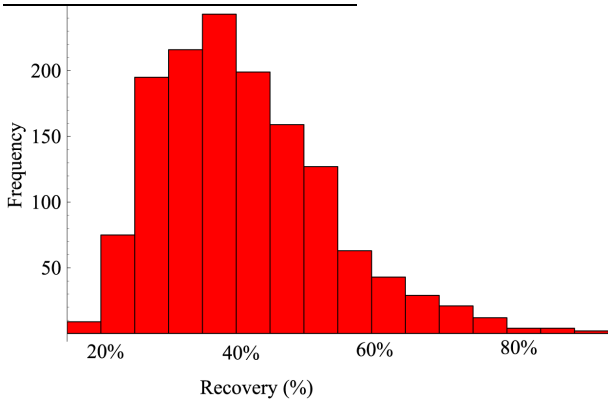


Figure 20. Posterior sample of probability of *E coli* O157 recovery parameter (p) from MCMC analysis

Figure 20 illustrates the uncertainty associated with the estimation of the probability of recovery. When these results are used in the QMRA the estimated recovery must be assumed to be representative of all *E. coli* 0157 analyses to which it is applied; therefore assuming that native organisms behave in the same manner as the spiked organisms regardless of water quality. The current model was chosen for simplicity and considered appropriate to the available data, however the recovery is only described as a point estimate and variation is not accounted for.

Quantifying Treatment Performance

A wide range of treatment processes exist for the physical, chemical and microbiological purification of drinking water. Each of these processes contributes to the removal or inactivation of pathogens from the water column. The effectiveness of each process in removing pathogens is variable: variable between different types of processes; between the same processes operated at different treatment facilities; and even variable over time for an individual process at a specified treatment plant.
The same treatment process may perform differently with respect to pathogen removal at different plants due to a number of factors including:

- process design – processes are optimized for the treatment of specific source waters, within the physical constraints of each specific sight;
- Source water- Different physical and chemical characteristics of source waters may be expected to affect the treatment performance; and
- Management – management protocol can vary between different managing agencies.

The performance of any given process may be expected to vary over time depending upon:

- Inlet water quality (including chemical, e.g pH, physical, e.g turbidity, and microbiological, e.g. algae count);
- Process conditions (e.g. chemical dosing, flow rate); and
- Maintenance (e.g. age of filter media).

Quantifying treatment removal performance for a drinking water CTS, accounting for the individual characteristics of the system being studied, and the expected temporal variability in performance for each process unit is a great challenge. Careful consideration of the available data is essential. Incorporation of different types of data (including literature data, online data and surrogates) for estimating removal performance has been discussed in Chapter 4. In this section, modelling techniques for quantifying variability (and the associated uncertainty) in treatment performance based on microbiological data are discussed.
Two general approaches to modelling treatment performance can be applied:

- **Mechanistic process model:** Treatment performance is predicted based on models describing the mechanisms of pathogen removal/inactivation. For example, disinfection is modelled based on hydraulic flow characteristics, disinfection dosage and individual pathogen inactivation kinetics (Chapter 4).
- **Empirical transformation model:** Without specific consideration of the individual process, the outflow pathogen concentration is compared with the inflow pathogen concentration. This approach has great value for many

processes, where quantitative information regarding the mechanisms of removal is very limited. For example, when modelling filtration performance, a mechanistic process model would give specific consideration to adsorption/desorption of microorganisms to the filter media, straining efficiency and inactivation rates within the filter. In contrast, a transformation model simply estimates total removal across the filter (or series of filters) based on outflow concentration as a fraction of the inflow.

The obvious limitation of simplifying a treatment process to a simple transformation is that any estimated removal is specific to the individual process and system studied, for the time frame represented by the available dataset. Since the underlying mechanisms and process characteristics are not accounted for (e.g. flow rate, source water quality, hydraulic design etc.) the impact of modifications to the system on pathogen removal cannot be quantitatively projected. Similarly, direct translation of results from one system to another is difficult, since the impact of even apparently minor differences in design or source water quality is unknown. In this section, approaches for modelling treatment processes by simple transformation are presented. The emphasis is placed on describing the variability and uncertainty associated with the transformation from the data available.

Transformation model assumptions

Pairing data points

Samples collected at the inlet and outlet may or may not be *paired*. Given a set of data, where samples have been collected and analysed from the inlet and the outlet of a water treatment process, should samples collected on the same day be assumed to be paired? For example, consider results from the enumeration of *Giardia* from water samples collected from the inlet and outlet of the water treatment process at CTS 10 (Table 9). For these results, is the reduction in *Giardia* cysts calculated on a daily basis actually representative of the variability in removal performance of the process? Or is it better to look at the overall removal as a point estimate of the process performance? When two samples are assumed to be paired, they are assumed to represent the "same" water as it enters and leaves the treatment process. To obtain a truly paired sample, the outlet sample must be delayed from the inlet sample by the hydraulic retention time of the process. In reality this is rarely the case. Results from CTS 10, while not technically "paired", are clearly correlated (Spearman rank correlation coefficient = 0.37)[8], with high inflow samples coinciding with high outflow samples (e.g. day 11) and vice versa (Day 5 and 6). It is often reasonable, for rapid processes, to assume that samples collected at approximately the same time, on a given sampling occasion are paired. Analysis can be undertaken with or without the assumption of pairing, however assuming that samples are paired allows for between sampling day variability to be characterized.

[8]Spearman rank correlation coefficient: $\tau = 1 - \frac{6\sum_{i=1}^{k}(r_{i,1} - r_{i,2})^2}{k(k^2-1)}$, Where k is the number of paired sample points and r_1 and r_2 are the within sample ranks for each inflow and outflow sample [Haas *et al.*, 1999, p343]

Table 9. *Giardia* counts enumerated from inflow and outflow samples collected from CTS10, with estimated concentrations and Log_{10} removal rates[9]

Sampling Day	Inflow			Outflow			REDUCTION
	Count	Volume	Estimated Conc. (cysts.L^{-1})	Count	Volume	Estimated Conc. (cysts.L^{-1})	Estimated Removal (Log_{10} reduction)
1	3	10	0.3	0	100	0	> 0.52
2	6	10	0.6	14	100	0.14	0.63
3	2	10	0.2	7	100	0.07	0.46
4	2	10	0.2	3	100	0.03	0.82
5	0	10	0	0	100	0	-
6	0	10	0	0	100	0	-
7	1	10	0.1	6	100	0.06	0.22
8	3	10	0.3	0	100	0	> 0.52
9	6	10	0.6	0	50	0	> 0.22
10	2	10	0.2	0	50	0	> 0.70
11	30	10	3	18	100	0.18	1.22
12	0	5	0	2	100	0.02	-
Average of all days			0.46			0.042	1.04

Variability in source water concentration

Constantly varying source water concentrations and random sampling effects can influence the representativeness of direct concentration comparisons, particularly for small microbial datasets. When outflow density estimates are "paired" with inflow samples collected on the same day, perplexing results can emerge including the apparent increase in microbial densities on some days (e.g. Table 9, Day 12). It is conceivable that for some processes, during a particular event, the microbiological concentration in the water column may be increased as a result of passage (for example regrowth of bacterial pathogens in filter media, or sloughing of a filter), however a far more likely and common explanation leans on the knowledge that source water concentration is constantly varying. By implication, if the number of organisms in the sample collected at the inlet was an instantaneous low density, while the outflow sample contained an instantaneous high density, the apparent change in density would be an increase, even though the underlying mean concentration may have decreased. Similarly, as a result of random sampling variability (see Section 0 p24) if the inflow sample contained a small number or organisms given the mean density, and the outflow sample contained a large number of organisms, an apparent increase in concentration may also result.

Accounting for random sampling variability and a varying source water concentration in the transformation model allows the underlying treatment removal performance to be characterised. In the same way as for source water characterisation, for the analysis

[9] Raw results are the counted number of organism ("Count") in sample volume ("Volume"). These results were used directly to find an estimate of the cyst density for that sample (Count/Volume), and then the estimated Log reduction was calculated from those densities ($Log_{10}Reduction = Log_{10}(Conc_{out}/Conc_{in})$). When there were no cysts found in the outflow concentration, the estimated $Log_{10}Reduction$ was reported as $>Log_{10}Conc_{in}$, when there were no organisms reported in the inflow and outflow, then no estimate of reduction could be made. Overall removal was calculated based on the average $Conc_{in}$ and average $Conc_{out}$ over all samples (including zeros).

of MicroRisk datasets, sampling variability was accounted for using the Poisson distribution, and mean sample concentration was assumed to vary according to a Gamma distribution (see Section 0).

Ratio of outflow concentration to inflow concentration

The probability that any individual organism will pass the treatment barrier (π, see Figure 1) can be estimated by the ratio of the outflow concentration to the inflow concentration. The Log_{10} of π is the Log_{10} removal of the process(es). For example, in Table 9, the Log10 removal is estimated for each day by calculated the ratio of the outflow concentration to the inflow concentration for each sampling day. This concept may be extended to account for variability in the inflow and outflow concentrations by finding the ratio of the outflow concentration PDF to the inflow concentration PDF. The resulting PDF for the ratio is therefore the PDF for π. An example of this approach is illustrated using Giardia results from CTS 10 (Table 9). This example is illustrative of using pathogen data to characterise treatment performance. In this situation, if the only aim was to calculate microbial risk, the outflow *Giardia* density PDF would be a suitable input to the QMRA model. By characterising the process performance, however, opportunities for management are possible (i.e. is the process working according as expected) and source water pathogen densities (where monitoring is more feasible) can be translated to expected outflow density.

A gamma distribution was fitted to the *Giardia* counts using the Poisson-gamma (negative binomial) mixture model (Equation 13), the maximum likelihood gamma distributions for the inflow and outflow are illustrated in Figure 21.

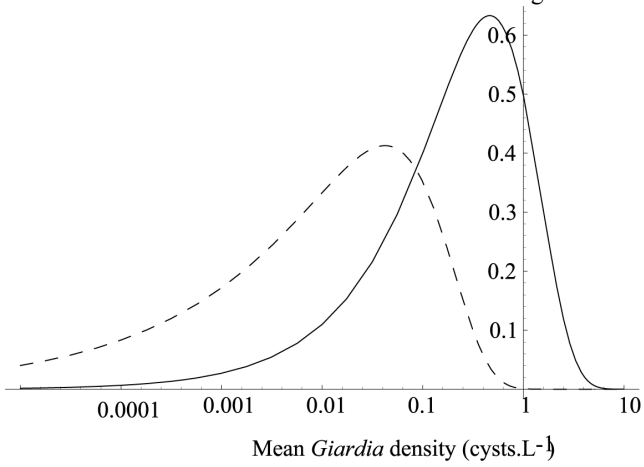


Figure 21. Maximum Likelihood Gamma distributions for mean *Giardia* density in the inflow (solid line) and outflow (dashed line) at CTS 10. Inflow: Gamma [$\hat{\rho}$ =0.61, $\hat{\lambda}$ =0.75], Outflow: Gamma [$\hat{\rho}$ =0.31, $\hat{\lambda}$ =0.13].

The distribution for the ratio of the outflow gamma distribution to the inflow gamma distribution was calculated using a random sampling procedure (Monte Carlo analysis) were random sample were drawn from the inflow and outflow distribution, and the ratio calculated for each random sample. A histogram of 10 000 random samples is illustrated in Figure 22a).

The great advantage of using the ratio to estimate treatment performance is simplicity; the estimate is a direct comparison between outflow and inflow. In addition the assumptions associated with the removal performance are limited, relying only on the

assumed pathogen density distribution. The ratio is not constrained to be less than one, and therefore the removal performance can take any value, and may even be

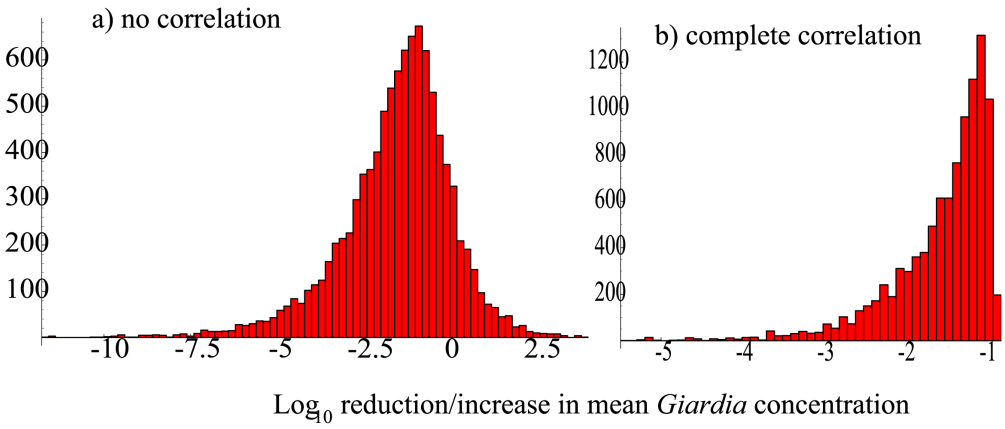


Figure 22. Histogram of Monte Carlo sample for the ratio between inflow and outflow Gamma distributions for *Giardia* density. a) no correlation - Lower 5%, 50% and Upper 95% quantiles of the sample were [- 4.51, -1.33, 0.70] b) complete correlation – random samples sorted before ratios calculated Lower 5%, 50% and Upper 95% quantiles of the sample were[-2.76, -1.34, -0.92]

positive (increase in pathogen density). Application of this ratio, to any *Giardia* inflow density sample, should provide an indicative estimate of the outflow density at CTS 10[10]. The ratio distribution illustrated in Figure 22a) was constructed assuming that the inflow and outflow concentrations were independent. This is however counter intuitive. Since the mean concentration would be expected to be reduced as a result of treatment, the random sample from the outflow distribution should always be less than the random sample from the inflow distribution. A simple modification to the Monte Carlo sampling procedure was undertaken assuming 100% correlation between the variables: random samples of inflow and outflow distributions were sorted (forcing complete rank correlation). The ratio was then calculated on these rank paired random samples, the resulting histogram of the Log_{10} reduction is illustrated in Figure 22b. The expected values of both analysis are similar (-1.33 versus -1.34), however the variability (and uncertainty, since they are not separated in this model) in the distribution for *Giardia* removal is greatly reduced in the correlated model in comparison to the independent model. The true result would be expected to lie somewhere between these two unrealistic extremes.

Binomial models

Models have been presented in the literature that describes organism passage through treatment as a binomial process [Teunis *et al.*, 1999a], where each microorganism faces one of two possible outcomes, passage or removal. Mean microorganism concentration in the inflow (μ) is assumed to follow a gamma distribution. Microorganism concentration in the finished water is then assumed to equal $\pi.\mu$, where π is the binomial probability of passage. A joint likelihood function can then be constructed that describes the inflow and outflow counts by a single gamma

[10] Parameter uncertainty associated with the fit of the Gamma distributions was not accounted for in this analysis.

distribution, scaled by the binomial probability of passage (π). Models for estimating the beta distributed probability of passage, under the assumptions of paired and unpaired data, as presented by [Teunis *et al.*, 1999a] were applied to the *Giardia* data from CTS 10 (Table 9). The results from likelihood analysis, including maximum likelihood parameter values are given in Table 10.

Table 10 Results from likelihood analysis of paired and unpaired binomial models fitted to *Giardia* counts from CTS 10 (Table 9)

	-2Loglik	$\hat{\rho}$	$\hat{\lambda}$	$\hat{\alpha}$	$\hat{\beta}$	$Log_{10}(\alpha/\alpha+\beta)$*
Unpaired	116.3	0.62	0.73	0.63	5.23	- 0.97
Paired	112.9	0.67	0.67	0.51	3.72	- 0.92

Log_{10} of the expected value of the beta distribution

Results from the likelihood analysis indicate that estimated reduction in *Giardia* was similar under either the paired or unpaired data assumptions. The paired model achieved a slightly better fit than the unpaired model (compare -2Loglik, deviance = 3.7) for the given dataset. The maximum likelihood beta distribution is illustrated with credible intervals in Figure 23. The expected removal of *Giardia* was 0.92 Log_{10} units with a 95% interval of variability ranging from 3.3 to 0.37 Log_{10} units.

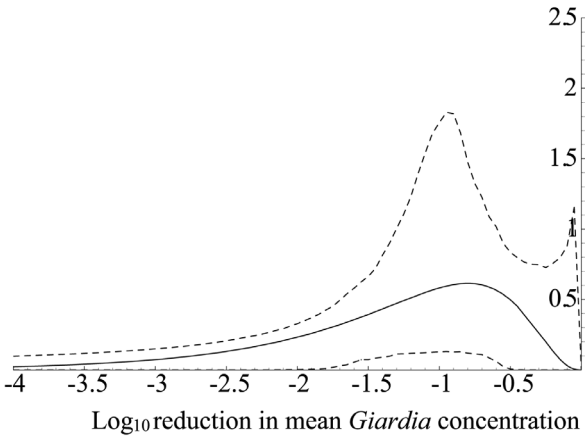


Figure 23 Maximum likelihood Beta distribution (solid line) and 95% credible intervals (dashed lines) for *Giardia* removal at CTS 10

The binomial modelling approach goes beyond the ratio calculations presented previously by making additional assumptions, in particular, that the passage of organisms through a treatment barrier is a binomial process. If each organism is assumed to have a certain probability π of passage, then the estimation of π is restricted to values between 0 and 1, eliminating the issue of negative removal. Within this approach, the uncertainty associated with the shape of the beta distribution can also be explored. The calculations involved in undertaking this analysis are relatively straight forward to implement in a mathematical software package (equations available in Teunis *et al.* [1999]), however the likelihood functions can become complex requiring not insignificant computational time. The implementation of a hierarchical modelling approach can simplify these computational issues (see

Gelman *et al.* [2004] for full explanation of methodology) while achieving equivalent results.

Surrogates

In order to apply a transformation model, and estimate the removal of organisms, it is necessary to have a sample of data points from the inlet and the outlet of the individual process or treatment chain to be studied. While some negative results can be easily managed through the application of a discrete distribution, the outlet sample must contain some positive results. Pathogens are usually present in finished waters at densities well below the limits of detection, and therefore this data requirement can be difficult to meet. The analysis and interpretation of surrogates (native or spiked) can then be necessary. While the same models can be used to evaluate treatment removal of microbiological surrogates, it is important to not ignore the uncertainty associated with assuming a given surrogate is representative of the pathogen of interest. In order to account for this quantitatively, data must be available to quantify the ratio or relationship between the surrogate and the pathogen for the given treatment process. When this ratio (including uncertainty) is applied to the surrogate removal PDF, a more representative estimate of the uncertainty associated with the pathogen removal PDF may be obtained. This type of data is rarely available, however the implications of uncertainty associated with the application of treatment performance surrogates must be accounted for. One proposal would be to use evaluate the assumption using sensitivity analysis (see Section 0).

Dose-Response

Many studies have been published that estimate dose-response relationships for human pathogens. Those studies relevant to the MicroRisk project are summarised in Table 11 and in the following sections. There are numerous sources of uncertainty regarding how adequately the cited dose-response models reflect the true impact of pathogen consumption on the population. Frequently, more than one model is relevant for consideration. Rather than proposing one single model as correct for each reference pathogen, some notes providing guidance on the different models considered and the data on which they are based are included in the following sections.

Campylobacter

Two studies have been conducted for estimating parameter values for the dose-response relationship of *Campylobacter jejuni.* The first fitted the Beta-Poisson model to data from a single human feeding trial, where administered doses were generally high [Black *et al.* 1988; Medema *et al.* 1996]. More recently, a second study has been presented that fits the dose-response relationship to both the first human feeding study and also two small outbreaks related to the consumption of raw milk [Teunis *et al.* 2005]. This second study gives consideration to low dose behaviour and indicates that health risks may be higher at lower doses than previously assumed from the first published parameter estimates. This second model is therefore more conservative and may be more representative of the entire population (including children) rather than simply healthy adults.

Table 11. Summary of key dose-response studies and associated results to be used for risk analysis

Reference Pathogen	DR Study Organism	Model	Parameters		Original Data Source	DR Analysis Source
Campylobacter	*Campylobacter jejuni*- Human feeding study	Beta-Poisson	α =0.145	β=7.59	Black *et al.* [1988]	Medema *et al.* [1996]
	Campylobacter jejuni – Outbreak data*	Beta-Poisson**	α =0.024	β=0.011	Van den Brandhof *et al.* [2003] Evans *et al.* [1996]	Teunis *et al.* [2005]
E. coli 0157	Enteropathogenic *E. coli* (EPEC) *Shigella disenteriae* (combined data)	Beta-Poisson	α=0.22	β=8.7 $\times 10^3$	Levine *et al.* [1973]	Powel *et al.* [2000]
	E. coli 0157 – Outbreak data	Beta-Poisson**			Shinagawa *et al.* [1997]	Teunis *et al.* [2004]
	Adults:		α =0.084	β=1.44		
	Children:		α =0.050	β=1.001		
	Shigella disenteriae	Beta-Poisson	α =0.157	β=9.16	Levine *et al.* [1973]	Teunis *et al.* [1996]
Norovirus	Rotavirus strain CJN(clinical isolate, not passed prior to administration) Human feeding trial	Beta - Poisson	α = 0.253	β= 0.422	Ward *et al.* [1986]	Teunis *et al.* [1996]
Enterovirus	Echovirus 12 clinical isolate Human feeding trial	Beta – Poisson	α = 0.401	β= 227.2	Schiff *et al*. [1984]	Teunis *et al.* [1996]
	Coxsackie A	Exponential	r = 0.014493			
Cryptosporidium	*Cryptosporidium parvum* (isolate from a calf) Human feeding trial	Exponential	r =4.005 $\times 10^{-3}$		DuPont *et al.* [1995]	Teunis *et al.* [1996]
	Combined dataset of three isolates	Beta-Poisson	α = 0.115	β = 0.176	Teunis *et al*., 2002a	Teunis *et al*., 2002a

	collected from neonatal calves.				
Giardia	*Giardia lamblia* Human feeding trial	Exponential	$r = 1.99\times10^{-2}$	Rendtorff (1954)	Teunis *et al.* 1996

*In this study, outbreak data was combined with the previous human feeding study to find overall dose-response parameter estimates.
**This study used the exact Beta-Poisson relationship rather than the approximation used in the other studies cited in the table.

Figure 24 illustrates the difference between the two published *Campylobacter* models. Teunis *et al.* [2005] assumes higher infectivity at low doses. If this model were used instead of the previous model Medema *et al.* [1996], predicted infection estimates would be more than an order of magnitude higher at low doses. Conversely, at high doses, the model from Teunis *et al.* [2005] would predict lower infection rates.

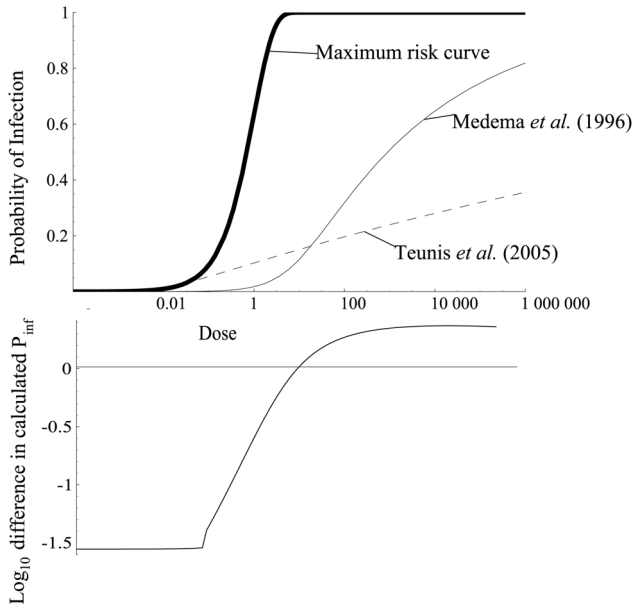


Figure 24. Dose-response relationships for *Campylobacter* and maximum risk curve, including the Log_{10} difference in calculated probability of infection with dose

E. coli 0157

Powell *et al.* [2000] combined human feeding study results from two surrogate organisms Enteropathogenic *E. coli* (EPEC) and *Shigella disenteriae* to provide a relationship for *E. coli* 0157:H7. Teunis *et al.* [2004] analysed actual outbreak data from school children and teachers who ate contaminated lunch [Shinagawa *et al.* 1997]. This later study compared their results with the former and identified the results from *Shigella disenteriae* (analysed by Teunis *et al.* [1996]) appeared to have the greatest agreement with the actual outbreak data [Teunis *et al.* 2004].

Viruses

While under development, the dose-response model for *Norovirus* is yet to be published. The dose-response relationships for Rotavirus, Echovirus and Coxsackie A are illustrated in Figure 25 with the maximum risk curve.

Figure 25 illustrates the relative infectivity assumed for each virus dose-response model. Echovirus 12 is the least infectious, followed by Coxsackie A and Rotavirus. The comparison of each individual model prediction with the maximum risk demonstrates that model choice has a large impact on estimated probability of infection at low doses. For viruses of relatively low infectivity, such as Echovirus 12, the choice of a conservative model such as Rotavirus or the maximum risk curve, may overestimate infection risk by more than 2 orders of magnitude. Conversely for pathogens that are known to be highly infectious, the assumption of the maximum risk curve may be considered, particularly in the absence of pathogen specific information.

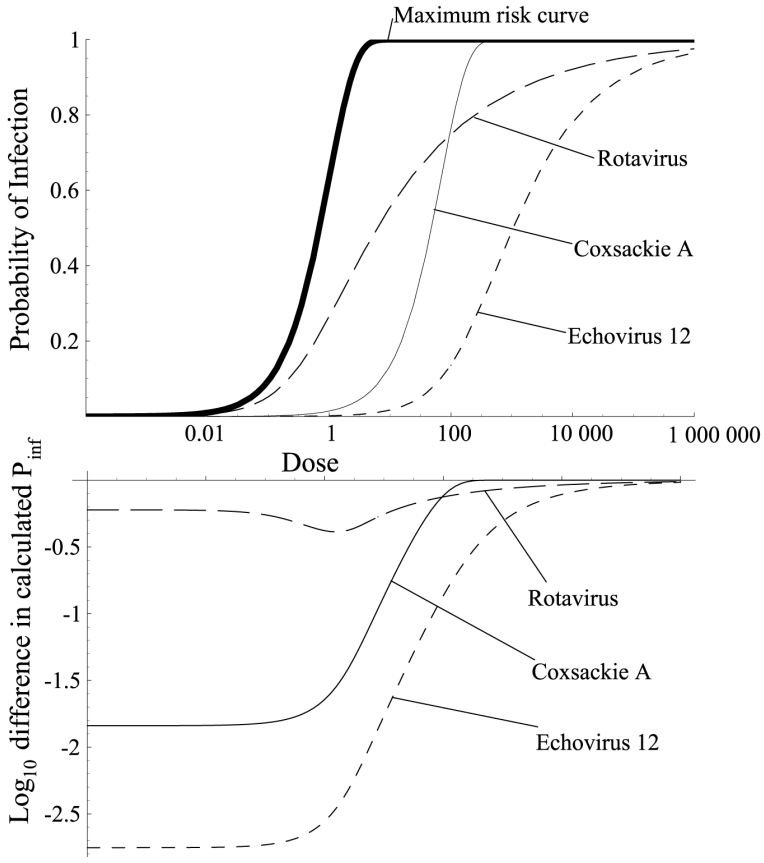


Figure 25 Dose-response relationships for Rotavirus, Coxsackie A, Echovirus 12 and the maximum risk curve, including the Log10 difference in calculated probability of infection between each individual model and the maximum risk.

Giardia

Results from human feeding trial with *Giardia lamblia* in were reported more than 50 years ago [Rendtorff, 1954]. The data was adequately fitted by an exponential distribution.

Cryptosporidium

Results from a human feeding trial with *Cryptosporidium parvum* were published by DuPont *et al.* [1995]. The data was adequately fit by the exponential model. More recently, work has been undertaken to investigate the variability in infectivity for *Cryptosporidium* between isolates [Teunis *et al.* 2002a] and between hosts [Teunis *et al.* 2002b]. The maximum likelihood estimates for parameter values when the Beta-Poisson model was fitted to the combined isolates dataset is included in Figure 26 The dose-response relationships for *Cryptosporidium* and *Giardia* are illustrated in Figure 26.

Figure 26 illustrates the comparison between the two published *Cryptosporidium* models. Selection of the more recent *Cryptosporidium* dose-response model instead of the previous model [Teunis *et al.*, 1996] may increase probability of infection estimates at low doses by more than 2 orders of magnitude.

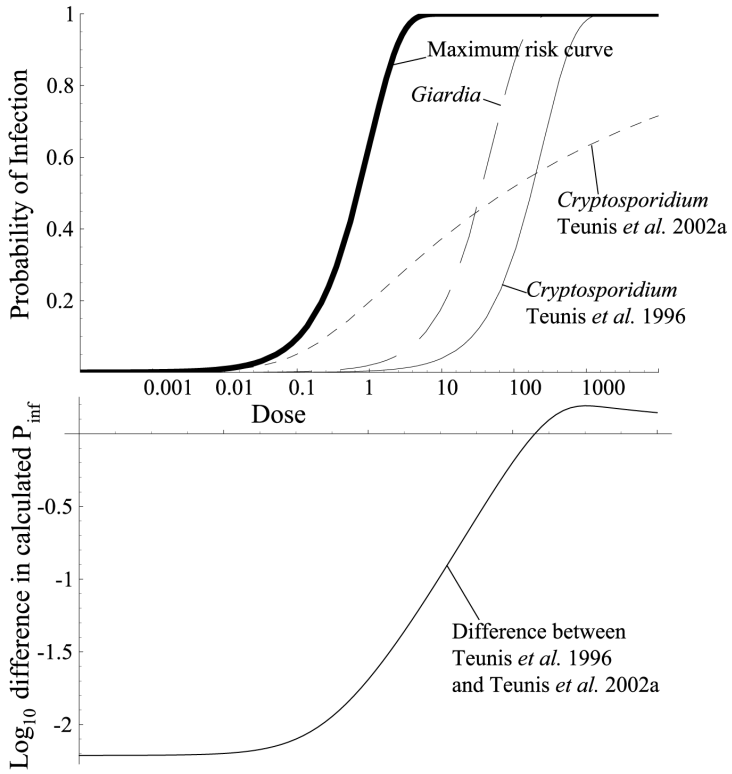


Figure 26 Two dose-response relationships for *Cryptosporidium, Giardia* and the maximum risk curve, including the Log_{10} difference in calculated probability of infection between the two *Cryptosporidium* curves

References

Ashbolt, N.J., Grabow, W.O.K., and Snozzi, M. (2001) Indicators of microbial water quality. *In* Fewtrell, L. and Bartram, J. (ed.), Water Quality: Guidelines, Standards and Health. Risk assessment and management for water-related infectious disease. IWA Publishing, London pp:289-316.

Barer, M.R., Harwood, C.R. (1999) Bacterial viability and culturability. *Advances in Microbial Physiology* **41**, 93-137.

Bates, B.C., and Campbell, E.P. (2001) A Markov chain Monte Carlo scheme for parameter estimation and inference in conceptual rainfall-runoff modeling. **37**, 937-947.

Bedford, T. and Cooke, R (2001) *Probabilistic Risk Analysis: Foundations and Methods* Cambridge University Press

Black, R.E., Levine, M.M., Clements, M.L., Hughes, T.P. and Blaser, M.J. (1988) Experimental *Campylobacter jejuni* infection in humans *J Infect Dis* **157**, 472-479.

Campbell, E.P., Fox, D.R. and Bates, B.C. (1999) A Bayesian approach to parameter estimation and pooling of nonlinear flood event models. *Water Resources Research* **35**, 211-220.

DeVries, T.A. and Hamilton, M.A. (1999) Estimating the antimicrobial log reduction: part 1. quantitative assays. *Quantitative Microbiology* 1,25-49.

DiGiorgio, D.J., Gonzalez , D.A. and Huitt, C.C. (2002) *Cryptosporidium* and *Giardia* recoveries in natural waters by using Environmental Protection Agency Method 1623 *Applied and Environmental Microbiology* **68**(12), 5952-5955.

DuPont, H.L, Chappell, C.L., Sterling, C.R., Okhuysen, P.C., Rose, J.B. and Jakubowski, W. (1995) The infectivity of *Cryptosporidium parvum* in healthy volunteers. *The New England Journal of Medicine* **332**(13),855-859.

Evans, M.R., Roberts, R.J., Ribeiro, C.D., Gardner, D and Kembrey, D. (1996) A milk-borne *Campylobacter* outbreak following an educational farm visit. *Epidemiol Infect* **117**,457-462.

Federighi., M., Tholozan, J.L., Cappelier, J.M., Tissier, J.P., Jouve, J.L. (1998). Evidence of non-coccoid viable but non-culurable Campylobacter jejuni cells in microcosm water by direct viable count, CTC-DAPI double staining, and scanning electron microscopy. Food Microbiology 15(5),539-550.

Francy, D.S., Simmons III, O.D., Ware, M.W., Granger, E.J., Sobsey, M.D. and Shaefer III, F.W. (2004) Effects of seeding procedures and water quality on recovery of *Cryptosporidium* oocysts from stream water by using U.S. Environmental Protection Agency Method 1623 *Applied and Environmental Microbiology* **70**(7), 4118-4128.

Frey, H.C. and Patil, S.R. (2002) Identification and review of sensitivity analysis methods *Risk Analysis* **22**(3), 553-578.

Furumoto, W.A. and Mickey, R. (1967) A mathematical model for the infectivity-dilution curve of tobacco mosaic virus: the theoretical considerations. *Virology* **32**,216-223.

Gelman, A., Carlin, J.B., Stern, H.A. and Rubin, D.B. (2004) *Bayesian Data Analysis* Second Edition Chapman & Hall.

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1996) (Eds.) *Markov Chain Monte Carlo in Practice* Chapman and Hall, London.

Haas, C.N., Rose, J.B. and Gerba, C.P. (1999) *Quantitative Microbial Risk Assessment.* Wiley, New York, USA.

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

Havelaar, A.H., Hollander de A.E.M., Teunis P.F.M., Evers E.G., Kranen van H.J., Versteegh F.M., Koten van J.E.M. and Slob W. (2000a) Balancing the risks and benefits of drinking-water disinfection: disability-adjusted life years on the scale. *Env. Health Pers.* **108**, 315-21.

Havelaar, A.H., Van Duynhoven Y.T.H.P., Nauta M.J., Bouwknegt M., Heuvelink A.E. De Wit G.A. Nieuwenhuizen M.G.M., Van de Kar N.C.A.J. (2003) Disease burden in the Netherlands due to infections with Shiga-toxin producing *Escherichia coli* O157. Bilthoven: National Institute of Public Health and the Environment Report no. 284550008.

Kuhn, R.C., and Oshima, K.H. (2002) Hollow-fiber ultrafiltration of *Cryptosporidium parvum* oocysts from a wide variety of 10-L surface water samples. *Can. J. Microbiol*. **48**, 542–549.

Levine, M.M., Kaper, J.B., Herrington, D., Losonsky, G., Morris, J.G., Clements, M.L., Black R.E.E., Tall, B. and Hall, R. (1973) Pathogenesis of *Shigella dysenteriae* 1 (shiga) dysentery. *The Journal of Infectious Diseases* **127**(3), 261-270.

McFeters, G.A. (1990) Enumeration, occurrence, and significance of injured indicator bacteria in drinking water. In *Drinking Water Microbiology.* (Ed. GA McFeters) pp. 478-492. Springer-Verlag, New York.

Medema, G.J., Teunis, P.F.M., Havelaar, A.H. and Haas, C.N. (1996) Assessment of the dose-response relationship of *Campylobacter jejuni. Int J Food Microbiol* **30**, 101-111.

Pipes, W.O., Ward, P. and Ahn. S.H. (1977) Frequency distributions for coliform bacteria in water. *Journal of the American Water Works Association* **69**(12), 664-668.

Powell, M.R., Ebel, E., Schlosser, W., Walderhaug, M., and Kause, J. (2000) Dose-response envelope for *Escherichia coli* O157:H7 *Quantitative Microbiology* **2**, 141-163.

Rendtorff, R. (1954) The experimental transmission of human intestinal protozoan parasites 1. *Giardia lamblia* cysts given in capsules. *American Journal of Hygiene* **59**, 209.

Rodgers, M.R., Flanigan, D.J., and Jakubowski, W. (1995). Identification of algae which interfere with the detection of Giardia cysts and Cryptosporidium oocysts and a method for alleviating this interference. *Applied and Environmental Microbiology* **61**(10), 3759-3763.

Schiff, G.M., Stefanovic, E.C., Young, J.R., Sander, D.S., Pennekamp, J.K. and Ward, R.L. (1984) Studies of Echovirus –12 in volunteers: determination of minimal infectious dose and the effect of previous infection on infectious dose. *Journal of Infectious Diseases* **150**, 858.

Shinagawa, K., Hu, D.Y. and Yoshida, S (1997) Correspondence and problem for enterohemorrhagic *E. coli* O157 outbreak in Morioka city, Iwate, *Koshu Eisei Kenkyu*, **46**, 104-112.

Signor, R.S., Ashbolt, N.J. and Roser, D.J. (2005) Quantitative Microbial Risk Assessment and management implications of rainfall induced runoff entering a source drinking-water reservoir Submitted to *Journal of Water Supply: Research & Technology* AQUA

Simmons, O.D., III, Sobsey, M.D., Schaefer III, F.W., Francy D.S., Nally, R.A. and Heany C.D.(2001) Evaluation of USEPA method 1622 for detection of *Cryptosporidium* oocysts in stream waters *J. Am. Water Works Assoc*. **93,** 78–87.

Smith, J.J., Gunasekera, T.S., Barardi, C.R.M., Veal, D.A., and Vesey, G. (2004). Determination of Cryptosporidium parvum oocyst viability by fluorescence in situ hybridisation using a ribosomal RNA directed probe. J*ournal of Applied Microbiology* **96**(2), 409-417.

Storey, M.V. and Ashbolt, N.J. (2003) Enteric virions and microbial biofilms - a secondary source of public health concern? *Wat. Sci. Tech.* **48**,97-104.

Suez Environnenment (2005) *Evaluation des Performances Détection des Escherichia coli 0157:H7 Par une méthode moléculaire semi-quantitative.* CIRSEE –PEA Réf. 0502SCr27_bio_perf 24/03/2005

Teunis, P., Takumi, K., and Shinagaw, K. (2004) Dose response for infection by *Escherichia coli* O157:H7 from outbreak data *Risk Analysis* **24**(2), 401-407.

Teunis, P., van den Brandhof, W., Nauta, M., Wagenaar, J., van den Kerkhof, H. and van Pelt, W. (2005) A reconsideration of *Campylobacter* dose-response relation *Epidemiol. Infect.* **133**, 583-592.

Teunis, P.F.M, Evers, E.G. and Slob, W. (1999a) Analysis of variable fractions resulting from microbial counts. *Quantitative Microbiology* **1**, 63-88.

Teunis, P.F.M. and Havelaar, A.H. (1999b) *Cryptosporidium* in drinking water: Evaluation of the ILSI/RSI quantitative risk assessment framework RIVM National Institute of Public Health and the Environment, Bilthoven, The Netherlands RIVM Report no. 284 550 006

Teunis, P.F.M., Chappell, C.L. and Okhuysen P.C. (2002a) *Cryptosporidium* dose response studies: variation between isolates. *Risk Analysis* **22**(1), 175-183.

Teunis, P.F.M., Chappell, C.L. and Okhuysen P.C. (2002b) *Cryptosporidium* dose response studies: Variation between hosts. *Risk Analysis* **22**(3), 475-485Teunis, P.F.M. and Havelaar, A.H. (2000) The Beta Poisson dose-response model is not a single hit model. *Risk Analysis* **20**, 513-520.

Teunis, P.F.M., Medema, G.J., Kruidenier, L. and Havelaar, A.H. (1997) Assessment of the risk of infection by *Cryptosporidium* or *Giardia* in drinking water from a surface water source. *Water Research* **31**, 1333-1346.

Teunis, P.F.M., van der Heijden, O.G., van der Giessen, J.W.B. and Havelaar, A.H. (1996) *The dose-response relation in human volunteers for gastro-intestinal pathogens* Technical Report 284550002, Rijksinstituut voor Volksgezondheid en Milieu RIVM.

Tillet, H.E. and Lightfoot, N.F. (1995) Quality control in environmental microbiology compared with chemistry: what is homogeneous and what is random? *Water Science and Technology* **31(5-6)**, 471-477.

US-EPA (1999). Method 1623 - Cryptosporidium and Giardia in Water by Filtration/IMS/IFA. Office of Water, United States Environment Protection Agency, EPA-821-R99-006, Washington D.C.

van den Brandhof, W., Wagenaar J. and van den Kerkhof, H. (2003) An outbreak of campylobacteriosis after drinking unpasteurized milk, 2002, The Netherlands. *Int J Med Microbial* **293**, 142.

Ward, R.L., Bernstein, D.I., Young, E.C., Sherwood, J.R., Knowlton, D.R. and Schiff, G.M. (1986) Human Rotavirus studies in volunteers: determination of infectious dose and serological response to infection. *Journal of Infectious Diseases* **154**, 871-880.

WHO (2004) Guidelines for Drinking-water Quality Third Edition. Volume 1. Geneva: World Health Organization.

Zwietering, M.H. and van Gerwen, S.J.C. (2000) Sensitivity analysis in quantitative microbial risk assessment. *International Journal of Food Microbiology* **58**, 213-221.