Incorporating method recovery uncertainties in stochastic estimates of raw water protozoan concentrations for QMRA

Susan R. Petterson, Ryan S. Signor and Nicholas J. Ashbolt

ABSTRACT

The impact of incorporating recovery data on protozoan concentration estimates was investigated for *Cryptosporidium* and *Giardia* using a large dataset ($n = 99$) of [oo]cyst assay results with paired recovery estimates. Stochastic [oo]cyst concentration was estimated using three approaches: I – no availability/consideration of recovery, II – limited recovery data, where sample recovery was considered as an independent random variable, and III – every [oo]cyst assay result was adjusted for a concurrently derived recovery estimate. Critically, Approach I underestimated [oo]cyst concentrations by about 100% compared to Approaches II and III, which were similar. The impact of dataset size on statistical uncertainty about the concentration estimate for Approach II was investigated; little improvement in parameter uncertainty was achieved beyond $n = 20$. It is suggested that recovery data be incorporated into source water concentration estimates, especially when used to infer health risks to consumers, so as not to underestimate the risk. Where none is available, conservatively low recoveries should be assumed. When designing monitoring programmes, recovery data should be collected as a pair with [oo]cyst count data for an initial period at least, so that site-specific relationships between those parameters may be ascertained and incorporated into source water concentration estimates.

Key words | *Cryptosporidium*, drinking water, *Giardia*, QMRA, recovery

Susan R. Petterson (corresponding author)
Ryan S. Signor
Nicholas J. Ashbolt
Centre for Water and Waste Technology,
University of New South Wales,
Sydney NSW 2052,
Australia
E-mail: *s.petterson@unsw.edu.au*

INTRODUCTION

The primary aim of the European Union commissioned *MicroRisk* project was to investigate the use of Quantitative Microbial Risk Assessment (QMRA) techniques (Haas *et al.* 1999) as a tool to aid development of water system management strategies, particularly with a view toward providing water that meets health-based quality targets for a variety of waterborne pathogens. Eleven medium to large-scale water supply systems from Europe and one from Australia, each with different source water and treatment method characteristics, provided case studies for the project. At the project's onset it was recognized that adequately estimating the prevalence of pathogens in a system's untreated source waters was central to performing a QMRA (see Teunis & Havelaar 2002), and by association, management actions borne from the assessment results.

doi: 10.2166/wh.2007.142

Cryptosporidium parvum and *Giardia lamblia* are two of the most implicated pathogens with respect to causes of waterborne diseases worldwide (Hunter *et al.* 2002; Craun *et al.* 2003) and are the focus of the current work; noting *C. homonis* followed by *C. parvum* are probably the main human infective species of *Cryptosporidium* (Leoni *et al.* 2006), but that oocyst speciation is not routinely undertaken. Environmental water samples are commonly assayed for *Cryptosporidium* oocysts and *Giardia* cysts ([oo]cysts) adhering to the United States Environmental Protection Agency's (USEPA's) Method 1623 (USEPA 1999). A complication to the interpretation of the assay results, however, is that enumeration methods are imperfect, involving many processes and steps, each of which may lead to loss/inactivation of, or simply not detecting, some

viable organisms that were originally present in the water sample being analysed. The primary reported result of an assay for parasitic protozoa on a sample volume of water using USEPA Method 1623, i.e. the count of identified [oo]cysts in the sample, is therefore a reflection of the number of enumerable organisms, not the total number present in the sample. "Recovery" is the term used to refer to the portion of microorganisms identified by a particular enumeration method with respect to the number that were actually initially present in the water sample. For protozoan assays the associated method recovery may be evaluated by spiking a known number of fluorescently labeled [oo]cysts into a sample and counting the fraction identified by that enumeration method, as described by Francy *et al.* (2004). Recovery values evaluated as such have been shown to be quite variable and dependent on both water quality characteristics as well as the skill level of the practitioner (Kuhn & Oshima 2002). Many have sought to quantify a dependence between more easily measured characteristics of the water sample and the expected recovery value. Some have identified a drop in expected recovery at very high turbidities (e.g. ~160 nephlometric turbidity units, NTU) (DiGiorgio *et al.* 2002; Kuhn & Oshima 2002; Feng *et al.* 2003; Francy *et al.* 2004). However, the findings of Feng *et al.* (2003) suggested that a moderate degree of turbidity (say, 5 NTU) actually enhanced recovery over less turbid waters. Nonetheless, DiGiorgio *et al.* (2002) noted that the nature of the turbidity and the background water/particles matrix is likely to be just as important as the absolute NTU measurement of the water sample.

There is currently no known general native surrogate available for estimating the recovery for *Cryptosporidium* and *Giardia* assays associated with a water sample. Further, recognising that recovery may vary in an unpredictable manner has led some groups to collecting recovery estimates with every environmental protozoan assay (e.g. Warnecke *et al.* 2003; Ferguson *et al.* 2004) and adjusting each reported raw count accordingly. Such an approach has not enjoyed widespread adoption. As part of the *MicroRisk* project's scope, data was collated from the twelve case-study system source waters with a view to quantify the distribution of enteric protozoa. No new data collection schemes were applied: rather, standard protocols were utilized by local personnel for each system, so as to reflect the varying degrees and types of available information a QMRA practitioner may encounter and be expected to work with. Where "source water" was defined as the water body at the immediate point of intake to the water treatment plant, Table 1 documents the nature of the supplied recovery data related to protozoa assays conducted on source water samples for each system. Only two (of the 12) systems provided a recovery estimate associated with every assay performed. Some form of recovery data or information was provided for an additional seven case systems, of which only two had recovery data collected prior to the commencement of the *MicroRisk* project. Further, how well recovery estimates related to environmental samples was often uncertain, since some recovery data was obtained in the laboratory using a variety of different sample types (e.g. a standard distilled water).

The mean recovery fraction estimates also varied markedly between systems (Table 1), suggesting that characteristics of the recovery estimates were unique to individual source water and laboratory locations. Effectively then, depending on the type of data available, three ways were considered to incorporate recoveries into the system-specific estimate of source water [oo]cyst concentrations. These three approaches were as follows:

Approach I: In the absence of any recovery data, the impact of recovery on the estimation of source water concentration was ignored. The raw counts from the enumeration method were assumed reflective of the number of [oo]cysts that were present in the sample.

Approach II: When some recovery data was available, but was unpaired with raw counts, [oo]cyst counts and method recovery were considered to be independent variables.

Approach III: When paired protozoan count and method recovery assay were provided, each count was individually adjusted. Intuitively, it is this third approach that would produce most representative concentration estimate as it best allows correlations between water sample characteristics and method recovery to be incorporated into the estimates.

Observed variability in a string of microbial water quality (and other environmental) data from a specific monitoring site is common and has typically been

Table 1 | Differences in protozoa assay recovery data characteristics for water at the water treatment plant intake for each *MicroRisk* system

					Cryptosporidium		*Giardia*	
Location	**Description of data collection**	**Recovery data collection scheme in place prior to *MicroRisk*?**	**Recovery estimate available for each source water assay supplied?**	***n***	**Mean**	**Std. dev.**	**Mean**	**Std. dev.**
Netherlands	Recovery evaluated on environmental source water samples	× **(= no)**	×	3	12%	16%	6%	5.4%
France (1)	Recovery routinely evaluated six times per year using ultra-pure water samples	✓ **(= yes)**	×	NA	30–40%	–	30–40%	–
Sweden	Recovery evaluated on environmental source water samples	×	×	4	12%	7%	8%	7%
Germany (1)	Recovery evaluated on ultra-pure water samples	×	×	3	19.2%	5.7%	14.9%	4.5%
Australia	Recovery routinely evaluated on environmental source water samples on each occasion that a sample was assayed for protozoa as part of in-house monitoring programme	✓	✓	28	50%	13%	47%	17%
France (2)	Recovery evaluated from environmental source water samples + ultra-pure water samples	×	✓	13	26%	21%	30%	29%
Germany (2)	Recovery evaluated on ultra-pure water samples with sediment added to simulate environmental source water	×	×	3	12%	3.1%	10.7%	7.3%

Source: adapted from Dechesne & Soyeux (2006). NB: Five other systems surveyed for the *MicroRisk* project supplied no recovery data at all.

adequately described by considering the environmental parameter as a random variable adhering to a lognormal or gamma type Probability Density Function (PDF) (e.g. Ott 1995). Those PDFs are characterised as being "skewed", i.e. non-symmetrical, where the overall mean value of the random variable is prone to be heavily influenced by the relatively rarer occurring periods of higher values. The added advantage of describing the parameter as a random variable is that it is in a form ready to be used in a probabilistic QMRA (e.g. Teunis *et al.* 1997) whereby the output is an estimated range, rather than simply a point estimate, of the consumer health risk. Further, the overall mean of the quantified stochastic risk to a consumer is typically sensitive to the higher percentile values of skewed source water concentration PDFs (Teunis *et al.* 2004). A related issue is that of *uncertainty* – which should also be assessed, reported and considered during the interpretation of QMRA outputs. As a general rule the level of uncertainty associated with statistical inferences made about a dataset is inversely proportional to the dataset size, and microbiological datasets are often "small" – as in Table 1 where recovery data sample sizes were $3 < n < 13$ for five of the seven case sites. Hence, describing the full extent of the source water quality variability and associated uncertainty, especially in the peak or higher percentile concentration estimates, is crucial to assessing the full extent of the risk of consumer exposure to waterborne pathogens.

When using protozoa count data to provide a source water concentration input for a probabilistic QMRA, consistently failing to account for the analysis method recovery may underestimate the source water pathogen concentration and lead to an underestimation of the consumer health risk. Stemming from the observed differences in the extents of recovery data available to complement environmental protozoa data for each *MicroRisk* water supply case system, the current work aimed primarily to explore the effect that it may have on the estimation of PDF parameters and other statistics used to describe variable source water *Cryptosporidium* and *Giardia* concentrations. This was undertaken illustratively using a paired protozoa count and assay method recovery dataset made available for study within the *MicroRisk* project from the Australian case system. The data had been collected over many years as part of routine monitoring by water utility personnel of a river that flows into the associated surface drinking-water reservoir. The dataset was ideal for the work undertaken given that it was relatively large ($n = 99$, which was the most comprehensive environmental protozoa prevalence dataset made available for the *MicroRisk* project) and contained significant proportions of assayed samples that had a reported [oo]cyst count > 0. Stochastic protozoa concentration estimates were made and associated uncertainties assessed then compared for several hypothetical scenarios. Those scenarios were based on the three possible approaches listed earlier, with which to incorporate the recovery into the concentration estimate (dependent on the nature of recovery data available). The implications for source water monitoring practices with a view toward the assessment and management of waterborne disease risk are also discussed.

MATERIALS AND METHODS

Data

Analytical results from 99 assays of 10 L environmental water samples for both *Cryptosporidium* and *Giardia* [oo]cysts collected over five years (2000–2004 inclusive) were available for the study. All sample analyses were undertaken by the Australian Water Quality Centre's (AWQC's) Analytical Laboratory (Bolivar, South Australia). Source water samples (10 L) were assessed for the presence of protozoa in water by flocculation concentration, immunomagnetic separation and immunofluorescence microscopy according to USEPA (1999) Method 1623, whereby the results were reported as an integer of counted [oo]cysts in the sample. Additionally, into each raw water sample were injected 100 fluorescently labeled 'ColorSeed™' (BTF Decisive Microbiology, Sydney) [oo]cysts. At the completion of the assay, the native [oo]cyst recovery was estimated by adjusting for the number of the 100 fluorescently labeled [oo]cysts identified. For 28 of the samples assessed, the corresponding recorded sample turbidity values were also available. Turbidity measurements of samples were made adhering to AWQC Standard Methods (18-01) and reported as nephlometric turbidity units (NTU). For the illustrative purposes of the current work, the dataset was assumed to be a totally random sample of the water quality conditions in the river over the sampling period.

Data analysis methods

General

Datasets for each protozoan were of the form $\{(c_1, r_1), (c_2, r_2), \ldots, (c_i, r_i), \ldots, (c_{99}, r_{99})\}$, where:

c_i = the number of [oo]cysts counted in the ith of a total of 99 10 L samples,

r_i = the number out of 100 labeled [oo]cysts counted in the ith of a total of 99 samples.

Statistical models were constructed to reflect the three ways with which to incorporate recovery information into protozoa concentration estimates as described in the introduction. Native and labeled [oo]cyst counts were discrete (rather than continuous) data, and therefore data analyses were based on counting statistic methods as described and applied by others for analyzing microbial count data (Pipes *et al.* 1977; Teunis *et al.* 1999, Teunis & Havelaar 1999, Teunis *et al.* 2004; Petterson *et al.* 2001). Maximum likelihood methods (e.g Montgomery & Runger 1999) were used to estimate model parameters and the uncertainty about the model parameters was quantitatively evaluated using Markov Chain Monte Carlo (MCMC) methods (Gelman *et al.* 2003). More details specific to each approach are given in turn. All described mathematical operations (i.e. numerical integrations, numerical optimization of likelihood functions and MCMC sampling applications) were performed using Mathematica 5.0 (Wolfram Research Inc., Champaign, USA) software.

Approach I: No recovery data available

As presented by Teunis *et al.* (1999), when [oo]cysts are assumed to be randomly dispersed in the source water, then the number counted c_i, in a sample volume v may be described by a Poisson distribution with parameter μ (representing the mean concentration). The mean concentration (μ) is, however, unlikely to be a constant value, and may be expected to vary. When that variability is described by a gamma distribution with scale and shape parameters λ and ρ (i.e. $\mu \sim \text{gamma}[\lambda, \rho]$), then the result is a contagious distribution[1], which can be arranged into the form of a negative binomial distribution, with the same parameters λ and ρ. The likelihood function for this model, given all sample volumes were 10 L, is given in Equation (1):

$$L(\lambda, \rho | c_1, c_2, \ldots) = \prod_{i=1}^{n} \frac{\Gamma(\rho + c_i)}{c_i! \Gamma(\rho)} \frac{\lambda^{\rho} 10_i^{c_i}}{(\lambda + 10)^{\rho + c_i}} \quad (1)$$

An MCMC procedure based on the Metropolis–Hastings algorithm (Gelman *et al.* 2003) using uninformative priors ($\text{Log}_{10}\alpha \sim \text{Uniform}[-10,10]$; $\text{Log}_{10}\beta \sim \text{Uniform}[-10,10]$) was used to generate a posterior sample of paired parameter estimates of λ and ρ from Equation (1). From this posterior sample, the 95% credible intervals (representing statistical uncertainty) about the maximum-likelihood gamma PDF were constructed.

Approach II: Accounting for method recovery when unpaired with counts

When paired recovery counts were assumed to be unavailable, the effect of recovery on the source water concentration was evaluated by treating recovery as an independent variable. The mean source water concentration μ^* was calculated by adjusting the countable concentration μ by the inverse of the true recovery fraction p:

$$\mu^* = \mu . \frac{1}{p} \quad (2)$$

It was again assumed that $\mu \sim \text{gamma}[\lambda, \rho]$: maximum likelihood estimates $\hat{\lambda}$ and $\hat{\rho}$ were derived from the previous likelihood function (Equation (1)) and MCMC-based uncertainty in parameter values were estimated as described above. Recovery may be considered as a binomial process, where every [oo]cyst in a sample has a probability p of being detected (and $1 - p$ of going undetected). The binomial parameter p can be estimated from the binomial distribution with 100 independent trials (100 seeded [oo]cysts), and r_i successes (enumerated). Like the source water concentration, the recovery fraction was observed to vary between experiments. The beta distribution with scale and shape parameters λ and ρ have been successfully applied by others (Teunis *et al.* 1999a) to describe the variability in p (i.e. $p \sim \text{beta}[\alpha, \beta]$) and was adopted here. The calculations were undertaken within a Bayesian hierarchical framework, which draws on a simulation approach to simplify the calculations of complex statistical models (Gelman *et al.*

[1] A probability distribution which is dependent on a parameter that itself has a probability distribution.

2003). The Bayesian structure is shown in Equation (3):

$$P(\alpha,\beta|r_1,r_2,\ldots) \propto \underbrace{P(\alpha,\beta)}_{\text{Prior}} \overbrace{\underbrace{\prod_{i=1}^{m} P(\alpha,\beta|p_i)}_{\text{Beta Distribution}} \underbrace{\prod_{i=1}^{m} P(p_i|r_i)}_{\text{Binomial}}}^{\text{LIKELIHOOD}} \quad (3)$$

where

$$\mathrm{P}(\alpha,\beta|\rho) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \quad \text{and}$$

$$P(p|r) = \binom{100}{r} p^{r}(1-p)^{100-r}$$

According to Bayes theorem, the Posterior distribution of the Beta parameters (α, β), given the number of spiked organisms recovered, is proportional to the Prior (α, β) multiplied by the likelihood. In this model, the likelihood consisted of two parts: firstly the binomial probability of recovery (p_i) given the number of spiked organisms recovered (r_i), and secondly the beta distribution describing the variability in probability of recovery (p). An MCMC procedure based on the Metropolis–Hastings algorithm was used to characterize the stationary posterior distribution of α and β (uninformative priors: $\mathrm{Log}_{10}\alpha \sim \mathrm{Uniform}[-10,10]$, $\mathrm{Log}_{10}\beta \sim \mathrm{Uniform}[-10,10]$). An estimate of the final variable source water concentration and its uncertainty was made by inputting $\mu = \mathrm{gamma}[\hat{\lambda},\hat{\rho}]$ (from Equation (1)) and $p = \mathrm{beta}[\hat{\alpha},\hat{\beta}]$ (from Equation (3)) into Equation (2) and using Monte Carlo sampling techniques (e.g. Vose 1996) to attain 500 estimates of μ^*. To summarize the variability in the Monte Carlo sampling outputs and to allow comparisons with gamma PDFs from other approaches it was assumed $\mu^* \sim \mathrm{gamma}[\lambda^*,\rho^*]$ and so those 500 values were then used to derive best estimates of the PDF parameters. Statistical uncertainty was evaluated as follows. Each posterior MCMC sample pair of α and β describing a beta distribution of recovery p was randomly combined (recovery samples less than 0.1% were rejected) with a posterior sample pair of λ and ρ describing source water concentration μ and input to Equation (2). Each time, 500 estimates of μ^* were made using Monte Carlo sampling. A new gamma distribution was fitted to this random sample to generate estimates of parameters λ^* and ρ^* describing actual source water concentration. From the resulting sample of gamma distributions the 95% uncertainty interval associated with predicting the actual source water concentration from the independent recovery sample and the native counts was estimated.

Approach III: Accounting for method recovery when paired with counts

Again the true mean [oo]cyst concentration was considered to be μ^*, and the recovery fraction to be p, then assuming random dispersion the number of organisms counted, c, in any water sample of volume v will be Poisson distributed with parameter $= \mu^* pv$. Assuming that $\mu^* \sim \mathrm{gamma}[\lambda^*, \rho^*]$, then derived similarly to Equation (1), the likelihood function to estimate those parameters is

$$L(\lambda^*,\rho^*|\{(c_1,p_1),(c_2,p_2),\ldots\}) = \prod_{i=1}^{n} \frac{\Gamma(\rho^*+c_i)}{c_i!\Gamma(\rho^*)} \frac{\lambda^{\rho^*} p_i 10^{c_i}}{(\lambda^*+p_i 10)^{\rho^*+c_i}}$$

Within a hierarchical framework, this equation was combined with the binomial probability of recovery (p) to estimate (by simulation) the stationary distribution of gamma parameters λ^* and ρ^* according to Equation (4)

$$P(\lambda^*,\rho^*|c_{1-n},r_{1-n}) \propto \underbrace{P(\lambda^*,\rho^*)}_{\text{Prior}} \overbrace{\underbrace{\prod_{i=1}^{n} P(\lambda^*,\rho^*|c_i,p_i)}_{\text{Negative binomial}} \underbrace{\prod_{i=1}^{n} P(p_i|r_i)}_{\text{Binomial}}}^{\text{LIKELIHOOD}} \quad (4)$$

From Bayes' theorem, the posterior distribution of gamma parameters (λ^* and ρ^*) given native and spiked count results (c, r), is proportional to the prior (in this case uninformative priors: $\mathrm{Log}_{10}\lambda^* \sim \mathrm{Uniform}[-10,10]$ and $\mathrm{Log}_{10}\rho^* \sim \mathrm{Uniform}[-10,10]$) multiplied by the likelihood. In this model, the likelihood function was comprised of two parts, the binomial probability of recovery (p), given the number of seeded organisms recovered (r), and the negative binomial count distribution describing the gamma parameters (λ^* and ρ^*) given the native counts and probability of recovery (p).

An MCMC procedure based on the Metropolis–Hastings algorithm (Gelman *et al.* 2003) was used to generate a posterior sample of paired parameter values of λ^* and ρ^* from Equation (4). The 95% credible intervals (representing statistical uncertainty) about each percentile of the

maxi mum-likelihood gamma PDF derived from our dataset were inferred from those MCMC samples.

Assessing recovery sample size impacts

For *Cryptosporidium*, the methods for assessing oocyst concentration variability and uncertainty from Approach II (i.e. assuming count/recovery data was unpaired and independent) were reapplied using smaller recovery datasets of $n = 50$, 20, 10 and 3, respectively, noting that some of the recovery datasets supplied from *MicroRisk* case systems were as small as $n = 3$. The smaller recovery datasets were artificially created by randomly sampling (with replacement) the required number of known seeded and recovered protozoa data pairs from the entire recovery dataset of $n = 99$. For each smaller dataset size assessed the results of this analysis would be expected to vary, depending on which particular values were selected during the random sampling, and therefore provided an example of only one possible outcome. To evaluate the extent of that variation, the random sampling of recovery data and subsequent analyses were repeated five times per smaller sized dataset, except for $n = 3$ which was repeated 15 times. As the emphasis was on the impacts of smaller recovery datasets only, all available protozoa count data was used on each reapplication so that it would remain a relatively representative sample of the protozoa counts from assays, and that any discrepancies in output results from the *status quo* (i.e. when using all recovery data) were attributable to differences in the nature of the available recovery information only.

RESULTS

General data properties

The mean recovery fraction for both assessed protozoa was 50%. For *Cryptosporidium* oocysts, the observed recovery ranged from 12–81% and for *Giardia* cysts it was between 10–96%. There were environmental [oo]cyst counts of > 0 in 73% (*Cryptosporidium*, max. count = 83) and 44% (*Giardia*, max. count = 41) of 10 L source water samples analysed, respectively.

Relationships among the sample [oo]cyst count, recovery and turbidity data with best-fitting linear trend lines are shown in Figure 1. The large scatter, near-horizontal trend lines and the low corresponding R^2 values on the environmental counts vs. recovery plots for both protozoa suggest a likely independence between those parameters. The assay method recovery value appeared inversely related to the sample turbidity for *Cryptosporidium*, though any such relationship was less evident for *Giardia*. Increased sample turbidities were likely to be associated with a higher count of enumerable environmental [oo]cysts present in a sample.

Protozoa concentration estimates and comparison of approaches

For visual comparison, the outputted maximum-likelihood PDFs with 95% credible intervals/uncertainty limits from each approach to describe variable and uncertain source water concentration estimates are displayed (Figure 2). Key numerical results are also documented (Table 2) – each of the listed values as estimated from Approach I were roughly half of the values from Approaches II and III, which in turn produced very similar results.

Recovery dataset size evaluation

Figure 3 illustrates the impacts of using 3–50 method–recovery data points in reducing the uncertainty in estimating the PDF for oocyst concentration. The smaller the recovery dataset the greater the associated uncertainty in the beta-distributed recovery variability and so concentration estimates. Table 3 documents the results of the analyses undertaken to assess recovery dataset size impacts on *Cryptosporidium* source water concentration estimates, whereby Approach II was employed using all available oocyst count data combined with an artificially generated smaller sized recovery dataset. Documented are the best estimates of the mean and 95th percentile of variability values, together with the upper bound 95% uncertainty limits of both those statistics, from each trial as indicated. Additionally, the change from the *status quo* estimates (i.e. those made when using all recovery data) is recorded as the decimal logarithm of the ratio ϕ of the revised estimate from the smaller dataset to the *status quo* value.

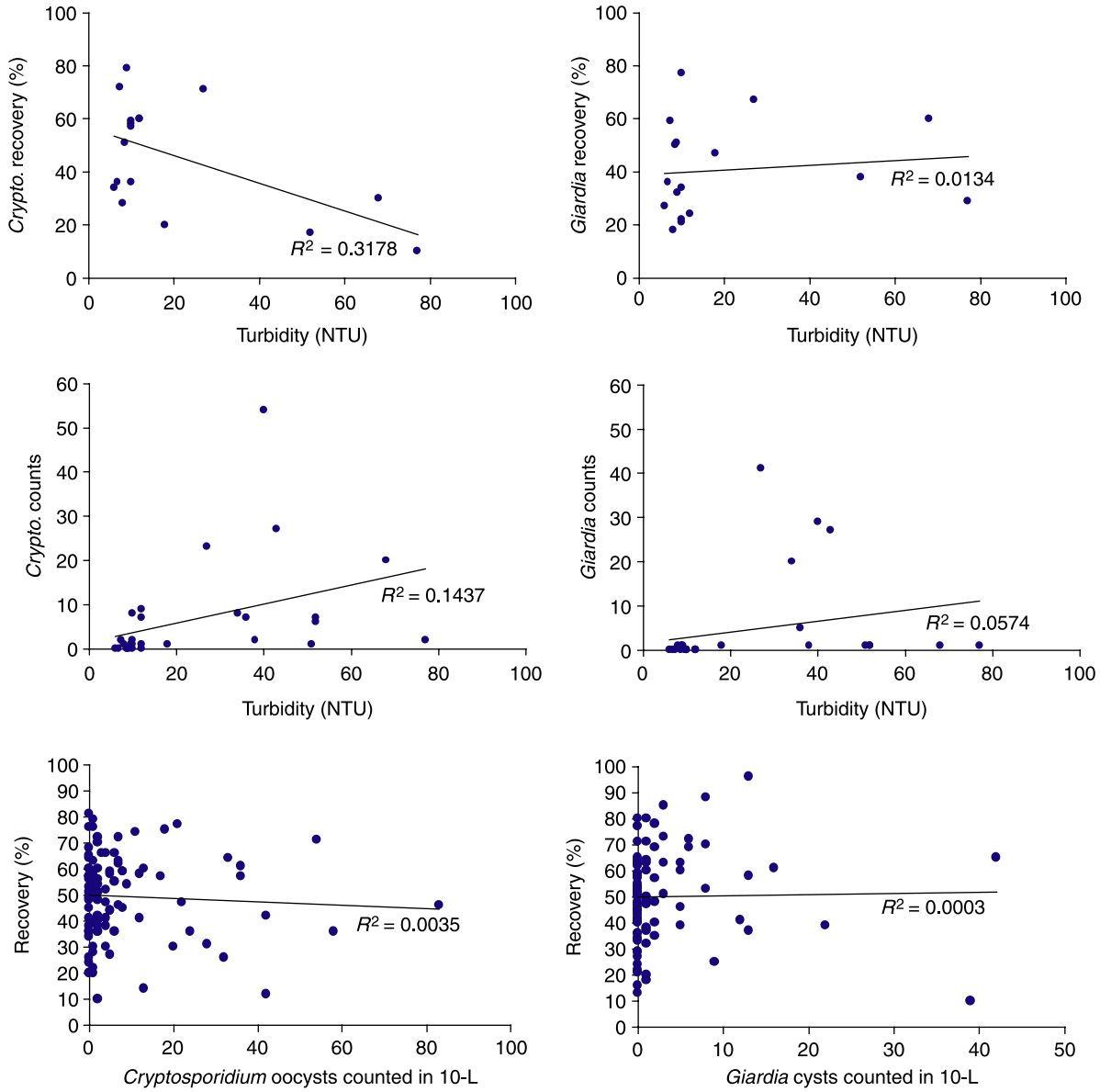

Figure 1 | [Oo]cyst counts, recovery and turbidity relationships.

The larger the recovery dataset size the more consistently close were the statistic estimates to those of the *status quo*. No $|\log_{10}(\phi)|$ values for any statistic were > 1 when the recovery dataset size was $n = 10$, i.e. the estimates did not deviate more than one order of magnitude from the *status quo* estimates. Conversely, when the recovery data was size $n = 3$, the new upper band of uncertainty statistic estimates were on average more than 10 times (and on occasion up to 100 times) greater than when using the full dataset.

The impacts on the best estimates of the statistics were relatively minor as compared to the impacts on their upper levels of uncertainty. For recovery datasets of $n = 3$ the absolute average $\log_{10}(\phi)$ values were 0.15 (mean) and 0.10 (95th percentile), i.e. the new estimates were generally of the same order of magnitude as those derived when using the full dataset. The corresponding values concerning the uncertainties in the statistics, however, were significantly greater at values of 1.04 and 1.12. The same trend was evident,

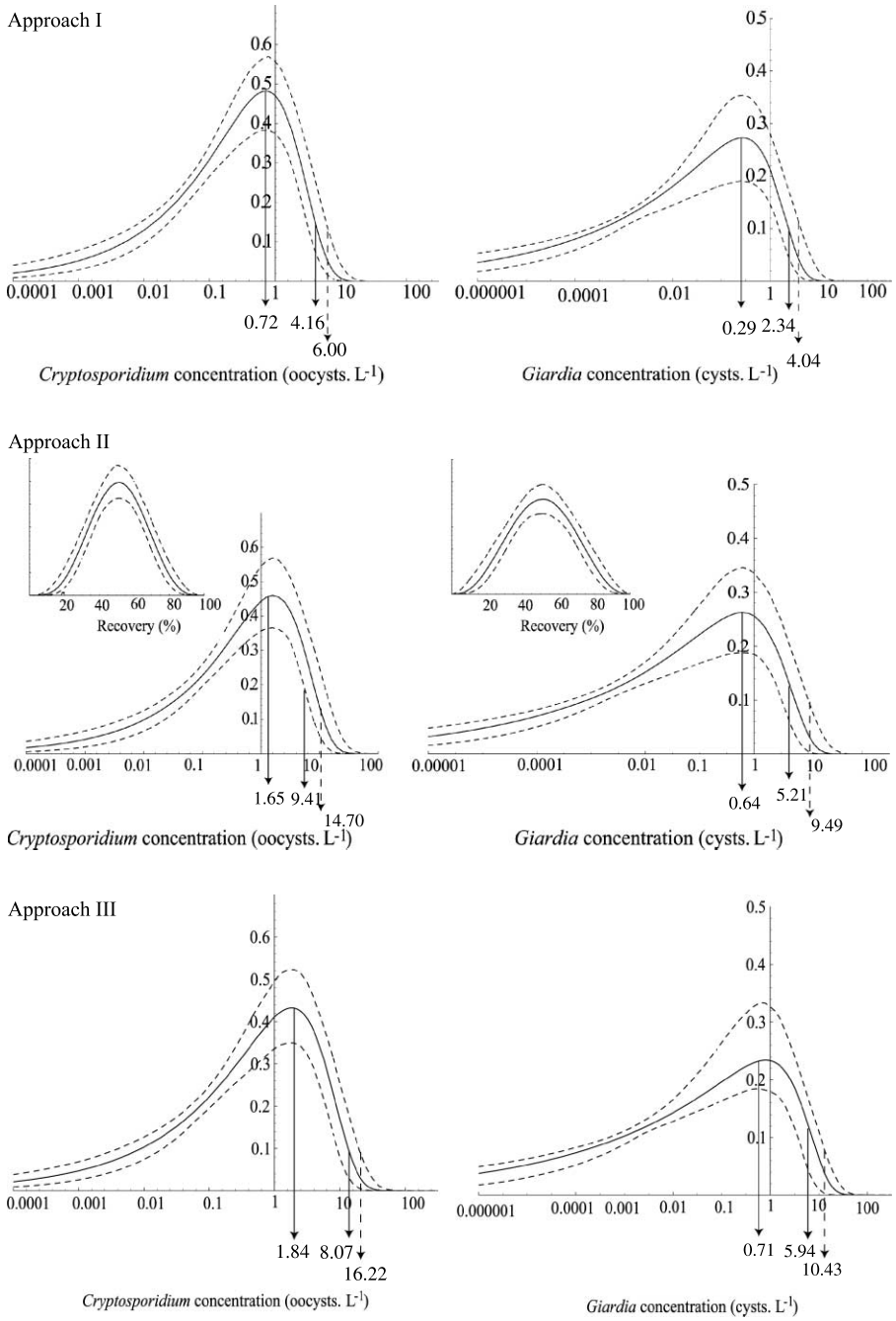

Figure 2 | Protozoan concentration PDFs with method recovery by Approaches I, II (with beta distribution inset) and III. **Approach I**: Maximum Likelihood Gamma PDF for *Cryptosporidium* and *Giardia* concentration based on native counts alone (no consideration of recovery) with 95% credible intervals (dotted lines) from MCMC analysis (mean concentration value, most likely upper 95% quantile and upper 95% credible limit of the upper quantile are indicated). **Approach II**: Maximum Likelihood Beta distribution for recovery with 95% credible intervals from MCMC analysis (inserts) and Maximum Likelihood Gamma PDF for *Cryptosporidium* and *Giardia* concentration with 95% credible intervals from MCMC analysis based on combination of Gamma PDF fitted to raw counts and Beta distributed recovery by Monte Carlo simulation. **Approach III**: Maximum Likelihood Gamma PDF for *Cryptosporidium* and *Giardia* concentrations based on paired native and spiked counts with 95% credible intervals (dotted lines) from MCMC analysis.

Table 2 | Comparison of key statistics in estimation of source water [oo]cyst concentrations

Method recovery*	*Cryptosporidium* (oocysts L^{-1})					*Giardia* (cysts L^{-1})				
	$\hat{\lambda}$	$\hat{\rho}$	Mean	95%ile	Upper 95%ile	$\hat{\lambda}$	$\hat{\rho}$	Mean	95%ile	Upper 95%ile
Approach I	0.40	1.81	0.72	4.16	6.00	0.18	1.53	0.29	2.34	4.04
Approach II	0.385	4.369	1.65	9.41	14.70	0.175	3.825	0.64	5.21	9.49
Approach III	0.34	5.41	1.84	8.07	16.22	0.15	5.54	0.71	5.94	10.43

*See methods section for description of recovery approaches I–III.

though to lesser degrees, as the sample size was increased, and was near negligible for recovery data points with $n = 20$.

DISCUSSION

The result of failing to account for recovery data when estimating protozoa concentrations was obvious from the outset. From the general Equation (2) it was apparent that as the fraction $p \to 0$, the estimate of the true mean concentration $\mu^* \to \infty$ regardless of the value of μ estimated from the raw counts. Hence, the consequences of ignoring imperfect detection increase as the recovery worsens. For the datasets examined, the estimates of source water concentration were of the same order of magnitude when

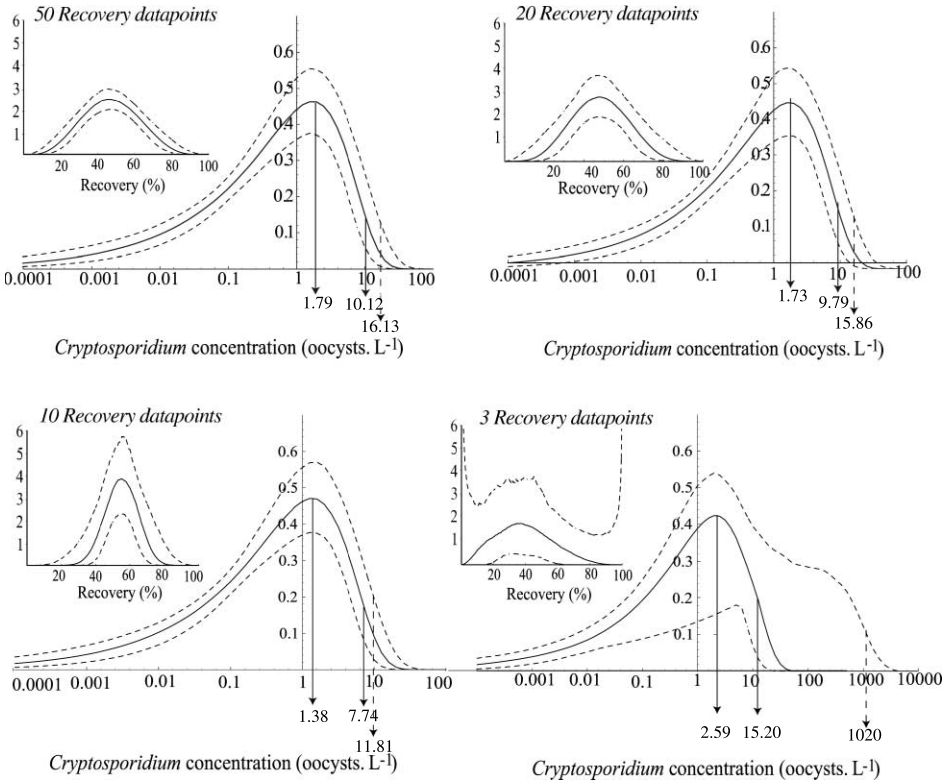

Figure 3 | Examples of fitted Beta distributions (inserts) and predicted source water *Cryptosporidium* oocyst concentration PDFs, including credible intervals, for various recovery dataset sizes using Approach II.

Table 3 | *Cryptosporidium* concentration estimates (oocysts.L^{-1}) using smaller recovery datasets of size *n* employing Approach II as indicated for a number of separate trials. The $|\log_{10}(\phi)|$ value is the log value of the ratio of the statistic indicated from the trial as compared to when all recovery data was used and reported in Table 2 (values >1 are bold, most extreme value is highlighted)

Trial	Mean Value	$\lvert\log_{10}(\phi)\rvert$	u* Mean value	$\lvert\log_{10}(\phi)\rvert$	95th % value	$\lvert\log_{10}(\phi)\rvert$	u* 95th % value	$\lvert\log_{10}(\phi)\rvert$
***n* = 50**								
1	1.45	0.05	2.14	0.04	8.18	0.05	12.57	0.05
2	1.69	0.01	2.52	0.03	9.64	0.02	14.30	0.01
3	1.55	0.02	2.31	0.01	8.66	0.03	13.46	0.02
4	1.75	0.03	2.66	0.05	10.03	0.04	15.42	0.04
5	1.65	0.01	2.39	0.01	9.28	< 0.01	14.03	< 0.01
Average		*0.02*		*0.03*		*0.03*		*0.02*
***n* = 20**								
1	1.64	0.00	2.55	0.04	9.30	< 0.01	15.62	0.05
2	1.70	0.02	2.72	0.07	9.64	0.02	16.14	0.06
3	1.88	0.06	3.96	0.23	10.85	0.07	24.90	0.25
4	1.66	0.01	2.57	0.04	9.40	0.01	15.27	0.04
5	1.89	0.07	3.16	0.13	10.84	0.07	18.78	0.13
Average		*0.03*		*0.10*		*0.03*		*0.10*
***n* = 10**								
1	1.38	0.07	2.26	0.02	7.87	0.07	13.65	0.01
2	1.79	0.04	3.44	0.17	10.31	0.05	20.40	0.16
3	1.53	0.03	2.47	0.02	8.65	0.03	14.27	0.01
4	2.24	0.14	5.66	0.38	12.93	0.15	36.05	0.41
5	2.75	0.23	15.85	0.83	16.42	0.25	114.05	0.91
Average		*0.10*		*0.28*		*0.11*		*0.30*
***n* = 3**								
1	1.76	0.03	**48.89**	**1.32**	7.87	0.07	**401.49**	**1.46**
2	1.91	0.07	4.18	0.25	10.28	0.05	24.24	0.24
3	1.78	0.04	6.05	0.41	9.22	0.00	38.56	0.44
4	3.25	0.30	**190.92**	**1.91**	11.84	0.11	**1470.67**	**2.02**
5	1.79	0.04	7.50	0.51	9.27	0.00	51.06	0.56
6	5.32	0.51	**150.96**	**1.81**	24.09	0.42	**1183.45**	**1.93**
7	2.26	0.14	**35.13**	**1.18**	10.21	0.04	**263.35**	**1.27**
8	1.56	0.02	21.86	0.97	7.44	0.09	**166.21**	**1.07**
9	1.84	0.05	21.22	0.96	3.36	0.44	**163.12**	**1.07**
10	1.98	0.08	19.54	0.92	9.89	0.03	133.14	0.98
11	2.05	0.10	**23.36**	**1.00**	8.73	0.02	**181.88**	**1.11**
12	2.63	0.21	11.09	0.68	10.25	0.05	75.49	0.73
13	3.86	0.37	**114.87**	**1.69**	14.67	0.20	**907.79**	**1.81**
14	3.87	0.38	**44.63**	**1.28**	11.02	0.08	**328.25**	**1.37**
15	1.79	0.04	22.50	0.98	8.89	0.02	**169.38**	**1.08**
16	1.90	0.07	**100.32**	**1.63**	10.66	0.06	**784.94**	**1.75**
17	2.97	0.26	15.22	0.81	12.31	0.12	101.13	0.86
18	1.18	0.14	2.13	0.04	7.21	0.11	12.54	0.05
19	2.01	0.09	**53.25**	**1.36**	9.79	0.03	**421.90**	**1.48**
20	1.50	0.04	**29.70**	**1.10**	6.85	0.13	**233.64**	**1.22**
Average		*0.15*		***1.04***		*0.10*		***1.12***

*u is a measure of "uncertainty" defined as the upper 95% credible interval of the posterior sample of the particular statistic from MCMC.

ignoring recovery as when it was incorporated, though the concentrations were underestimated, as expected. That the differences were not more marked was a result of the "good" recovery values reported from the laboratory with $E(p) = 50\%$, i.e. about half of all [oo]cysts present in a sample could be expected to be counted. Consequently the concentration estimate statistics from Approaches II and III were more or less double those from Approach I. Nonetheless, the summarised recovery estimates provided in Table 1 illustrate that they were as low as 12% from some laboratories. Teunis *et al.* (1999) have demonstratively estimated the mean recovery of assays from another laboratory in the Netherlands to be as low as 2% and large variabilities in [oo]cyst recoveries have been reported in trials of Method 1624 (Bukhari *et al.* 1998; Clancy *et al.* 1999; McCuin *et al.* 2000, McCuin *et al.* 2001). Neglecting the recovery impacts on count data from those labs would see an underestimation of concentrations by about 10 and 100 times, respectively. Others (Teunis *et al.* 1997) have already prescribed that an inclusion of the impacts of recovery into source water concentration estimates as a routine part of good QMRA practice and these results serve to support that call.

Those who have previously recommended that internal recovery estimates be made for every protozoa assay conducted on environmental samples (e.g. Warnecke *et al.* 2003; Ferguson *et al.* 2004) have done so based on sound logic borne from scientific analyses. In short, the reasoning was that the recovery was observed to be variable and dependent on the water quality conditions (e.g. turbidity) of each sample. Raw [oo]cyst counts within environmental samples have also displayed correlations with some of the same water quality parameters (e.g. higher turbidity values correlated with higher observed counts), so providing internal quantitative controls, providing the most direct and informative quality assurance mechanism with which to assess each datum collected. Undoubtedly some recovery data should be collected by any practitioner who provides microbial data to test methods and for quality assurance purposes. Internal spiking of labeled [oo]cysts into environmental samples overcomes many limitations experienced by isolated laboratory recovery trials, in particular the need to replicate the specific water matrix of the environmental samples. An emphasis of the current work was on assessing the impacts of the type and quantity of recovery data available on the estimates of variability and uncertainty in raw water protozoan concentrations used for quantitative health risk assessments.

From the case-study data presented here, rising sample turbidity levels were generally associated with lower recoveries and higher [oo]cyst counts (more so for *Cryptosporidium* than *Giardia*). However, that did not translate here to any identifiable relationship between the counts and recoveries themselves (Figure 1). A relationship between counts and recovery should be intuitively expected; given a constant underlying source water concentration, as the recovery increases, so the number of organisms counted would also increase. The fact that this relationship was not observed in these datasets suggests that the influence of variation in recovery was small relative to the variation in underlying [oo]cyst concentration. For our large dataset ($n = 99$) the best estimates and associated uncertainties of parameters from Approaches II (assumed independence between counts and recoveries) and III (keeping paired recovery and count data together) were very similar (Figure 2, Table 2). Hence, where independence between counts and recovery can be reasonably assumed Approach II is as suitable a methodology as Approach III for estimating protozoa concentrations in water. Though more computationally rigorous and requiring an assumption about the underlying distribution of the variable recovery, the primary advantage of Approach II over III was that it could be utilized with unpaired datasets of different sizes, which is the current norm. Furthermore, a recovery estimate need not be collected for every assay to adequately estimate the concentration stochastically. That would appeal to a water utility manager designing a long-term monitoring program, due to the reduced cost. In such circumstances however, the experimental design for undertaking the recovery assays should be formulated carefully to ensure that the full range of environmental conditions are represented.

Importantly, the assumption of a beta-distributed recovery (Approach II) appears to be an adequate assumption about the variability of the recovery fraction, but should be further tested with different datasets. A limitation, however, of assuming that recovery may be described as a continuous random variable is that, at low values of recovery, the beta distribution projects

toward zero and Monte Carlo simulation in this region leads to unrealistically high random concentration samples. A pragmatic solution employed in this study to avoid the generation of extreme values was to truncate the beta distribution at a recovery of 0.1%. Further consideration of a practical truncation point or alternative algorithm would be useful for QMRA simulation.

In this study, no quantitative consideration was given to how well labeled [oo]cysts represented the behavior of native [oo]cysts during analysis. Warnecke *et al.* (2003) reported that over the analysis of 494 water samples, on average ColorSeed™ *Cryptosporidium* recoveries were 3.3% lower than unlabeled *Cryptosporidium*, and ColorSeed™ *Giardia* recoveries were 4% lower than unlabeled *Giardia*. Reliance on labeled organisms may therefore lead to a slight underestimation of recovery, and therefore a conservative concentration estimate. While the mean discrepancy in recovery between native and labeled [oo]cysts appears to be relatively small, Warnecke *et al.* (2003) did not report variability in these discrepancies between samples. If the variability was large, the influence on the uncertainty in the concentration estimates may not be negligible.

With emphasis on a dataset with a large number of assay results available to characterize the variability in oocysts counts (in this case $n = 99$), a series of trials was conducted to assess recovery dataset size impacts on variability and uncertainty as compared to the *status quo*. From the results (Table 3) it was apparent that the smaller the recovery dataset the less adequately the overall concentration was described. Even for datasets of $n = 3$ though the best estimates of the mean and 95th variability percentile were quite similar to when $n = 99$, consistently varying by less than an order of magnitude. For $n = 20$ the differences were negligible. It was in the statistical uncertainties that the small dataset effects were most profound. It was the assumption of beta-distributed recovery fractions that drove this outcome – as the PDF properties meant that the smaller the recovery dataset the less certain that the real value of $E(p)$ was not approaching zero (Figure 3), leading to very high concentration estimates. For $n = 3$ the upper mean and 95th percentile were up to 100 times greater than the *status quo*. For $n = 20$ the average absolute deviations for the same statistics from the primary results using all recovery data were in the range of just 1.0–1.8 times. For $n = 50$ the changes were negligible. In summary, based on the data analysed, a recovery dataset of size $20 < n < 50$ would have generally been as informative as $n = 99$ for both best estimates and uncertainties in statistics. Recovery datasets used in QMRA should then ideally be at least about of size $n > 30$ (which is coincidentally in agreement with traditional statistical theory texts that refer to dataset sizes $n < 30$ as "small") to reduce statistical uncertainty about concentration estimates to low practical levels.

The inclusion of quantified parameter variability and uncertainty into QMRA models is an established concept, and should be incorporated whereever possible by conducting either a two-dimensional (variability and uncertainty) risk assessment or via sensitivity analysis that especially examines impacts of uncertainties. From a risk management perspective, the high level of uncertainty in source water concentrations reported for when recovery datasets were size $n < 10$ could pose problems for decision-making, e.g. the high uncertainty bands would make it difficult to assuredly compare QMRA outputs to health targets. Where a QMRA result is presented that does not incorporate uncertainty about a QMRA model parameter and the assessment result comfortably meets some health target, the result should not necessarily provide one with real confidence that the water product was safe. Reconducting the QMRA with the incorporation of parameter uncertainty may see some aspect of the risk outputs exceed target values (particularly when datasets are small), resulting in a likely need to collect more (in this case recovery) data to verify that the system meets health requirements or otherwise. Given that estimates of the source water concentration mean had upper band uncertaintiy values up to 100 times greater (when the recovery dataset was $n = 3$ as compared to 99) suggests a real problem for interpreting QMRA outputs with small recovery datasets.

CONCLUSIONS

For the purposes of undertaking a QMRA, system-specific recovery data should be collected and incorporated into the source water concentration estimates, so that risks are not underestimated. While others have recommended recovery estimates for every source water assay for quality assurance, this level of additional cost is not necessarily required for a

satisfactory QMRA. That said, doing so would prove advantageous for the source water concentration estimation, as altering every sample eliminates the need to make any assumptions about the nature of the variability of the recovery data or its relationship with [oo]cyst count data. Where doing so is not logistically possible or desirable, adequate source water concentration estimates can be made from the unpaired modeling approach utilized here – specifically where recovery and count data can be assumed independent and when at least some 20–30 recovery data points (representative of the range of site conditions) are available. Particularly so when only fewer site-specific recovery data points are available for analysis, the statistical uncertainty about source water concentration estimates should be reported along with data variability statistics. Where no system-specific recovery data is available, it is recommended that the sensitivity of QMRA outputs to a range of conservative estimates about the mean value of the recovery fraction (say, for values from 0.01–0.1) be assessed rather than just ignoring the effects or applying data from another site. Doing so will provide an assessment of the impacts of method recovery on quantitative health risk assessments and ensure better informed decisions about the adequacy of the water supply system to supply safe water.

ACKNOWLEDGEMENTS

The *MicroRisk* project (see Medema *et al.* 2006) was co-funded by the European Commission under the Fifth Framework Programme, Theme 4: "Energy, environment and sustainable development" (contract EVK1-CT-2002-00123). *MicroRisk* was conducted primarily by Kiwa Water Research, the Swedish Institute for Infectious Disease Control, Anjou Recherche, Veolia Water Partnership, Water Research Centre-NSF, University of Bonn, Suez Environnement, University of East Anglia, Technical University of Delft, Technologiezentrum Wasser and the University of New South Wales.

REFERENCES

Bukhari, Z., McCuin, R. M., Fricker, C. R. & Clancy, J. L. 1998 Immunomagnetic separation of *Cryptosporidium parvum* from source water samples of various turbidities. *Applied and Environmental Microbiology* **64**, 4495–4499.

Clancy, J. L., Bukhari, Z., McCuin, R. M., Matheson, Z. & Fricker, C. R. 1999 USEPA method 1622. *J. AWWA* **91**, 60–68.

Craun, G. F., Calderon, R. L. & Nwachuku, N. 2003 *Causes of Waterborne Outbreaks Reported in the United States, 1991–1998*. CRC Press, Boca Raton, FL.

Dechesne, M. & Soyeux, E. 2006 Source water quality. In *Quantitative Microbial Risk Assessment in the Water Safety Plan* (Medema, G., Loret, J., Stenström, T.-A. & Ashbolt, N (Eds.)). Report for the European Commission under the Fifth Framework Programme, Theme 4: "Energy, environment and sustainable development" (contract EVK1-CT-2002-00123). Kiwa Water Research, Nieuwegien, The Netherlands, ch 3.

DiGiorgio, C., Gonzalez, D. & Huitt, C. 2002 *Cryptosporidium* and *Giardia* recoveries in natural waters by using Environmental Protection Agency Method 1623. *Appl. Environ. Microbiol.* **68**, 5952–5955.

Feng, Y. Y., Ong, S. L., Hu, J. Y., Song, L. F., Tan, X. L. & Jern, N. W. 2003 Effect of particles on the recovery of *Cryptosporidium* oocysts from source water samples of various turbidities. *Appl. Environ. Microbiol.* **69**, 1898–1903.

Ferguson, C., Kaucner, C., Krogh, M., Deere, D. & Warnecke, M. 2004 Comparison of methods for the concentration of *Cryptosporidium* oocysts and *Giardia cysts* from raw waters. *Can. J. Microbiol.* **50**, 675–682.

Francy, D. S., Simmons, O. D. 3rd, Ware, M. W., Granger, E. J., Sobsey, M. D. & Schaefer, F. W. 3rd 2004 Effects of seeding procedures and water quality on recovery of *Cryptosporidium* oocysts from stream water by using US Environmental Protection Agency Method 1623. *Appl. Environ. Microbiol.* **70**(7), 4118–4128.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. 2003 *Bayesian Data Analysis*, 2nd edn. Chapman and Hall, Boca Raton, FL.

Haas, C. N., Rose, J. B. & Gerba, C. P. 1999 *Quantitative Microbial Risk Assessment*. John Wiley & Sons, Inc, New York.

Hunter, P. R., Waite, M. & Ronchi, E. 2002 *Drinking Water and Infectious Disease: Establishing the Links*. CRC Press, Boca Raton, FL.

Kuhn, R. C. & Oshima, K. H. 2002 Hollow-fiber ultrafiltration of *Cryptosporidium parvum* oocysts from a wide variety of 10-L surface water samples. *Can. J. Microbiol.* **31**, 417–423.

Leoni, F., Amar, C., Nichols, G., Pedraza-Diaz, S. & McLauchlin, J. 2006 Genetic analysis of *Cryptosporidium* from 2414 humans with diarrhoea in England between 1985 and 2000. *J. Med. Microbiol.* **55**(6), 703–707.

McCuin, R. M., Bukhari, Z. & Clancy, J. L. 2000 Recovery and viability of *Cryptosporidium parvum* oocysts and *Giardia intestinalis* cysts using the membrane dissolution procedure. *Canadian Journal of Microbiology* **46**, 700–707.

McCuin, R. M., Bukhari, Z., Sobrinho, J. & Clancy, J. L. 2001 Recovery of *Cryptosporidium* oocysts and *Giardia* cysts from source water concentrates using immunomagnetic separation. *Journal of Microbiological Methods* **45**, 69–76.

Medema, G., Loret, J., Stenström, T.-A. & Ashbolt, N (eds) 2006 *Quantitative Microbial Risk Assessment in the Water Safety Plan*. Report for the European Commission under the Fifth Framework Programme, Theme 4: "Energy, environment and sustainable development" (contract EVK1-CT-2002-00123). Kiwa Water Research, Nieuwegien, The Netherlands.

Montgomery, D. C. & Runger, G. C. 1999 *Applied Statistics and Probability for Engineers*. Wiley, New York.

Ott, W. 1995 *Environmental Statistics and Data Analysis*. Lewis Publishers, Boca Raton, FL.

Petterson, S. R., Teunis, P. F. M. & Ashbolt, N. 2001 Modeling virus inactivation on salad crops using microbial count data. *Risk Analysis* **21**, 1097–1107.

Pipes, W. O., Ward, P. & Ahn, S. H. 1977 Frequency distributions for coliform bacteria in water. *J. AWWA* **69**, 664–668.

Teunis, P., Davison, A. & Deere, D. 2004 *Short-term Fluctuations in Drinking Water Quality and their Significance for Public Health. Report*. World Health Organisation, Geneva, Switzerland.

Teunis, P., Medema, G., Kruidenier, L. & Havelaar, A. 1997 Assessment of the risk of infection by *Cryptosporidium* or *Giardia* in drinking water from a surface drinking water source. *Wat. Res.* **31**(8), 1333–1346.

Teunis, P. F. M., Evers, E. G. & Slob, W. 1999 Analysis of variable fractions resulting from microbial counts. *Quant. Microbiol.* **1**, 63–88.

Teunis, P.F.M. & Havelaar, A.H. 1999 *Cryptosporidium in Drinking Water: Evaluation of the ILSI/RSI Quantitative Risk Assessment Framework*. Report no. 284 550 006. National Institute of Public Health and the Environment, Bilthoven, The Netherlands.

Teunis, P. F. M. & Havelaar, A. H. 2002 Risk assessment for protozoan parasites. *International Biodeterioration and Biodegradation* **50**, 185–193.

USEPA 1999 *Method 1623 -Cryptosporidium and Giardia in Water by Filtration/IMS/IFA. Office of Water*. United States Environment Protection Agency, Washington DC.

Vose, D. 1996 *Risk Analysis: A Quantitative Guide*, 2nd edn. John Wiley & Sons, Chichester.

Warnecke, M., Weir, C. & Vesey, G. 2003 Evaluation of an internal positive control for *Cryptosporidium* and *Giardia* testing in water samples. *Lett. Appl. Microbiol.* **37**, 244–248.