International Journal of Food Microbiology 195 (2015) 48–57

Contents lists available at ScienceDirect

International Journal of Food Microbiology

journal homepage: www.elsevier.com/locate/ijfoodmicro

Impact of microbial count distributions on human health risk estimates

A.S.R. Duarte *, M.J. Nauta

Technical University of Denmark — National Food institute, Mørkhøj Bygade, 19, Building H, DK-2860 Søborg, Denmark

ARTICLE INFO

Article history:
Received 8 May 2013
Received in revised form 7 August 2014
Accepted 22 November 2014
Available online 2 December 2014

Keyword:
Probability distribution
Zero-inflation
Microbial concentration
Prevalence
Limit of quantification
Quantitative microbiological risk assessment

ABSTRACT

Quantitative microbiological risk assessment (QMRA) is influenced by the choice of the probability distribution used to describe pathogen concentrations, as this may eventually have a large effect on the distribution of doses at exposure. When fitting a probability distribution to microbial enumeration data, several factors may have an impact on the accuracy of that fit. Analysis of the best statistical fits of different distributions alone does not provide a clear indication of the impact in terms of risk estimates.

Thus, in this study we focus on the impact of fitting microbial distributions on risk estimates, at two different concentration scenarios and at a range of prevalence levels. By using five different parametric distributions, we investigate whether different characteristics of a good fit are crucial for an accurate risk estimate. Among the factors studied are the importance of accounting for the Poisson randomness in counts, the difference between treating "true" zeroes as such or as censored below a limit of quantification (LOQ) and the importance of making the correct assumption about the underlying distribution of concentrations.

By running a simulation experiment with zero-inflated Poisson-lognormal distributed data and an existing QMRA model from retail to consumer level, it was possible to assess the difference between expected risk and the risk estimated with using a lognormal, a zero-inflated lognormal, a Poisson-gamma, a zero-inflated Poisson-gamma and a zero-inflated Poisson-lognormal distribution.

We show that the impact of the choice of different probability distributions to describe concentrations at retail on risk estimates is dependent both on concentration and prevalence levels. We also show that the use of an LOQ should be done consciously, especially when zero-inflation is not used. In general, zero-inflation does not necessarily improve the absolute risk estimation, but performance of zero-inflated distributions in QMRA tends to be more robust to changes in prevalence and concentration levels, and to the use of an LOQ to interpret zero values, compared to that of their non-zero-inflated counterparts.

1. Introduction

Quantitative microbiological risk assessment (QMRA) depends on consistent descriptions of the distributions of pathogen concentrations in food products (Bassett et al., 2012; Busschaert et al., 2010, 2011; Nauta et al., 2007; Straver et al., 2007), since these distributions eventually have a large effect on the distribution of doses at exposure. Furthermore, the concentrations in the high value tail of the distribution often have most impact on the risk (Bassett et al., 2010; Straver et al., 2007). Usually, concentrations are represented by probability distributions fit to enumeration data obtained from a set of samples. One of the most frequently used distributions for this purpose is the lognormal (LN), due to its generally good fit to enumeration data and its attractiveness for statistical testing (Gonzales-Barron and Butler, 2011; Kilsby and Pugh, 1981)

In this study, the impact on risk estimates of the choice of different probability distributions to describe *Campylobacter* concentrations in chicken at retail is considered.

When fitting a distribution to enumeration data from a sample of food products, several factors have an influence on the accuracy of the fit. First, enumeration of low contaminated sample units can give zero counts ("artificial" zeroes) that add to the number of "true" zeroes resulting from non-contaminated units, thereby inflating the total number of zeroes in a sample of microbial counts. This results in a typically zero-inflated distribution of counts. In this study, we investigate what the impact is of an inappropriate treatment of zero counts on the estimated risk, given that the distribution of counts is zero-inflated. To deal with zeroes, two approaches can be adopted: to treat the total number of zeroes as left-censored data (results below a limit of quantification (LOQ)) (Busschaert et al., 2010, 2011; Delignette-Muller et al., 2010; Helsel, 2005a, 2005b; Lorimer and Kiermeier, 2007; Shorten et al., 2006), or alternatively to use zero-modified distributions, such as zero-inflated models (Bassett et al., 2010; Gonzales-Barron et al., 2010; Ridout et al., 1998) to model prevalence and concentration.

* Corresponding author. Tel.: +45 23673256.
E-mail addresses: asrd@food.dtu.dk (A.S.R. Duarte), maana@food.dtu.dk (M.J. Nauta).

http://dx.doi.org/10.1016/j.ijfoodmicro.2014.11.024
0168-1605/

Zero-inflated distributions specify the probability of obtaining a non-contaminated unit and the concentration distribution for the contaminated ones, thus allowing for a separation between "true" and "artificial" zeroes.

Second, the test portion taken from each sample unit is homogenized prior to serial dilution and enumeration, and the enumerated colony forming units (cfu) can be assumed to be generated by a Poisson process. The resulting distribution of cfu counts has been referred to as the "measurement distribution" (Gonzales-Barron and Butler, 2011) and is theoretically close to a Poisson. Continuous distributions, like the LN or the gamma, are often considered appropriate to model the distribution of concentrations at retail, but they do not account for the discrete Poisson process at the measurement level. They exclusively describe the heterogeneity in concentrations (a continuous measure) between food products. Generalized Poisson distributions, like the Poisson-gamma (PGM) or the Poisson-lognormal (PLN) are considered more mechanistic as they describe both realities (Reinders et al., 2003) — distribution of continuous concentrations and the discrete "measurement distribution."

Third, microbial concentrations in food are often considered to be lognormally distributed (Busschaert et al., 2010; Crepet et al., 2007; Kilsby and Pugh, 1981). However, the frequency distribution of pathogens in food is commonly characterized by a low prevalence (hence a high probability of "true" zeroes) and low concentrations (potentially yielding "artificial" zeroes). This complicates the fulfillment of lognormality, as the LN distribution does not allow zero as an outcome. Recently, several alternatives to the LN distribution have been proposed to represent microbial contamination data with low prevalence and low concentrations more appropriately (Bassett et al., 2010; Gonzales-Barron et al., 2010; Gonzales-Barron and Butler, 2011). Among these are the discrete generalizations of the Poisson distribution, PGM and PLN and the zero-inflated distributions for their ability to model data with a substantial amount of zeroes.

Fourth, data sets usually consist of concentration values that were estimated from enumeration data obtained from a sample of food products. The uncertainty associated to sampling (Zhao and Frey, 2004) and the measurement uncertainty (Marks and Coleman, 1998) influence the difference between a "true" distribution in a population (for example of concentrations in food products) and the fitted distribution(s). Only the selection of a large representative sample and a perfect enumeration procedure (without measurement errors), could minimize those uncertainties, but both are unlikely to achieve in reality.

Recently, many authors have explored different solutions to these challenges by studying choices of the probability distribution to fit through enumeration data (Bassett et al., 2010; Commeau et al., 2012; Gonzales-Barron et al., 2010; Gonzales-Barron and Butler, 2011; Pouillot et al., 2013; Reinders et al., 2003; Williams and Ebel, 2012a, 2012b.). Most research on fitting distributions to microbial data has focused on the best statistical fits of different models to existing data sets. Bassett et al. (2010) additionally investigated the impact of microbial distributions on public health, without considering different levels of prevalence of contamination. From the perspective of risk assessment, the importance of applying the best statistical fit when selecting a distribution to describe microbial data is however not clear. The impact on the risk estimate of choosing a specific probability distribution of concentrations is not known. If this impact would be small, selection of the best fitting distribution need not be a priority in QMRA.

In this study, we therefore took this research question one step further and investigated the importance of fitting microbial distributions on QMRA estimates. This is done at different concentration and prevalence levels, using a hypothetical example of *Campylobacter* on broiler meat. We assessed the impact of the use of different parametric distributions and of different ways of dealing with zeroes when fitting those distributions, either by separating "true" from "artificial" zeroes (i.e. by estimating prevalence) or by treating all zeroes as left-censored results below a certain LOQ (i.e. by assuming 100% prevalence). By running a simulation experiment under different prevalence scenarios while keeping the same original concentration distribution, we assessed the advantage of using zero-inflated distributions to model prevalence separately from concentrations, depending on the probability of occurrence of "true" zeroes.

Furthermore, we investigated whether the impact of fitting different distributions on the risk changed depending on the contamination level at retail. For that purpose, risk estimates obtained for two realistic levels of contamination were compared: the lowest and highest contamination levels of broiler meat with *Campylobacter* at Danish retail.

2. Materials and methods

2.1. Simulating different scenarios of retail concentration

Two scenarios were analyzed, representing "true" distributions of *Campylobacter* concentrations (C) in broiler meat at retail. For each scenario, we defined an LN distribution:

$$\log_{10}(C) \sim \text{Normal}(\mu, \sigma) \tag{1}$$

with geometric mean μ and geometric standard deviation σ.

Values for μ and σ were based on the analysis of Danish retail data (data not shown). The high contamination scenario ($\mu = 0.75$ $\log_{10}$ cfu/g meat) was based on the chilled meat data for high prevalence seasons in years 2004 to 2007; the low contamination scenario ($\mu = 0$ $\log_{10}$ cfu/g meat) was based on the frozen meat data for the low prevalence seasons in the same period (Boysen et al., 2011). The adopted standard deviation ($\sigma = 1$) was based on the analysis of the same retail data and did not differ substantially between the low and the high season.

From each distribution, we randomly sampled 500 values, representing a set of "true" concentrations in 500 broiler meat retail products.

Next, we simulated the experimental enumeration procedure for that sample. We assumed four serial dilutions with three replicates each, and a standard portion weight of 10 g taken for analysis. In the first dilution step ($j = 1$), we simulated the homogenization of the 10 g portion in a 90 ml volume, using a stomacher. Consequently, the number of cfu (N_{ij}) for each sample unit i, at each dilution step j resulted from a Poisson process with mean λ, where $\lambda = C_{\text{ml\_}ij} \times 3$.

$$N_{ij} \sim \text{Poisson}\left(C_{\text{ml\_}ij} \times 3\right) \tag{2}$$

and

$$C_{\text{ml\_}ij} = \frac{C_i}{10} \times d_j \tag{3}$$

where C_i represents the "true" concentration in cfu/g of a sample unit i ($i = 1,2,\ldots,500$), $C_{\text{ml\_}ij}$ represents the concentration in cfu/ml of that unit at dilution step j ($j = 1,2,3,4$) and d_j represents the dilution factor of the concentration in cfu/ml ($10^{(1-j)}$) at dilution step j. It was assumed that 1 ml is the analytical portion used for enumeration.

The estimate of the concentration in cfu/ml ($C_{\text{ml\_est\_}i}$) for each sample unit i was calculated as a weighted average of the results

$$C_{\text{ml\_est\_}i} = \frac{\Sigma_{j=1}^{4} N_{ij}}{\Sigma_{j=1}^{4}\left(d_j \times 3\right)} \tag{4}$$

and the estimate of the concentration in cfu/g ($C_{\text{est\_}i}$) for the same unit corresponded to $C_{\text{ml\_est\_}i} \times 10$.

According to this simulated protocol, the resulting limit of quantification (LOQ) of the enumeration procedure corresponds to the count of 1 cfu in one single replicate plate of any single dilution j. According to Eq. (4), the LOQ in cfu/ml is equal to 1 cfu/3.333 ml, i.e. 0.3 cfu/ml.