Quantitative Microbial Risk Assessment; the weight of variability in the assessment of risks.

Pierre Renault

Tutorials M2 EUR Implanteus - 16 December 2021

Quantitative Microbial Risk Assessment; the weight of variability in the assessment of risks.

Pierre Renault (INRAE, pierre.renault@inrae.fr)

I. Introduction

This tutorial on **Quantitative Microbial Risk Assessment (QMRA)** aims to consolidate students' knowledge, by proposing a case study where the assessment is carried out successively by a deterministic approach then by a stochastic (or probabilistic) approach.

The bases on Quantitative Microbial Risk Assessment were explained in the Lecture on this morning (Thusday 16 December). For the case study discussed in this tutorial, 3 pages recall the basics of the process; these 3 pages are attached at the end of this document. **It is important that these bases are acquired!** In fact, the 1st part of this tutorial (section III) will aim to remember them by carrying out a first risk assessment that does not consider variabilities in various parameters and variables.

The objective of the 2nd part of the tutorial (section IV) is to learn how to use in EXCEL tools allowing to carry out random distributions to simulate different types of distributions (uniform, normal and triangular distributions). Whatever the tool used (tools under EXCEL, R software, or functions in different programming languages), you should know that:

- random number generators all have their shortcomings / limitations. It may be useful to do some preliminary tests to verify that they are apparently suitable for your simulations: test on the conformity of the distribution obtained in relation to the target distribution, absence of correlations between different variables simulated by the same random number generator …;
- in the specific case of the use of generators for the quantitative assessment of microbial risks that are marginal (i.e. very low) (for example ensuring a DALY (Disability-adjusted life years) below an acceptable threshold of the order of 10^{-4} at 10^{-6} pppy (per person per year)), the number of random draws necessary to properly account for very rare risks can be extremely high (100,000 or more) and much easier to manage under the R software or by a scientific programming language than in a spreadsheet (note however that it is possible to do Visual Basic programming in EXCEL).

The objective of the last part of the tutorial (section V) is to resume the risk analysis by taking into account a dispersion of the values associated with certain variables. This section will begin by considering that the actual contamination of wastewater with salmonella follows a triangular distribution law.

II. Tutorial context

The tutorial was designed from the following article:

Amha Y.M., Kumaraswamy R., Ahmad F. 2015. A probabilistic QMRA of Salmonella in direct agricultural reuse of treated municipal wastewater. *Water Science and Technology*, 71(8), 1203-1211.

This article is accompanied by a *"Supplementary Information"* file downloadable from the journal's website, providing important additional information.

Certain tables, certain formulas and certain equations or values proposed in this article or in its *"Supplementary Information"* file are reproduced in this document.

In Abu Dhabi (United Arab Emirates), treated wastewater leaves the wastewater treatment plant after treatment with activated sludge, followed by disinfection with chlorine. Its nominal capacity is 260,000 $m^3.day^{-1}$. Currently, effluents (treated wastewater) are only used for landscape irrigation. A previous screening study carried out at the same treatment plant had identified *Salmonella enterica* as the most abundant of the pathogens. It was considered a potential danger in this new study.

Salmonella were monitored by molecular biology (qPCR). The authors assumed that any DNA sequence detected from *Salmonella* matches that of an infectious *Salmonella*. The extraction and quantification of the nucleic acids were carried out on 1 L water samples. In total, 24 samples of post-chlorinated treated effluents were taken on 8 dates (3 repetitions per date) between January and July 2014.

Table 1: Measured *Salmonella* concentrations ($number.L^{-1}$)

Date (in 2014)	Replicate 1	Replicate 2	Replicate 3	**Mean**
		$number.L^{-1}$		
9 January	10280	8220	9460	**9320**
24 February	1850	910	1530	**1430**
16 March	4390	3020	3030	**3480**
21 April	10150	2590	9130	**7290**
5 May	15890	9820	14670	**13460**
7 May	9770	4150	7080	**7000**
18 May	14800	6540	11240	**10860**
17 July	3790	1780	2290	**2620**

It is assumed that the individual analyses correspond to correct real measurements, and that there are no temporal trends such as seasonal fluctuations. Thus the 24 individual values of the preceding table can be considered as 24 random draws from a law of unknown random distribution.
The general average of the concentrations is **6,932.5 Salmonella.L^{-1}** of water.

On the sampling dates, a full physicochemical analysis was performed on the wastewater samples by the wastewater treatment plant, using samples dialled every hour over 24 hours. In addition, the concentration of *Escherichia coli* was estimated for each day of sampling.

Consumption of raw vegetables (lettuce, cabbage and cucumber) irrigated with treated wastewater was chosen as the exposure route for the QMRA model. The model considered two types of populations: a population that washes products before consumption (w = 1) and a population that does not wash before consumption (w = 0). The daily dose of Salmonella from consuming raw vegetables was determined using the following equation:

$$\lambda_w = M_{body} \times M_i \times C_{iw} \times V_{prod} \times 10^{-w} \times e^{-kt} \quad (1)$$

where λ_{washed} is the daily dose of salmonella ingested by people washing their products ($number.persone^{-1}.day^{-1}$), M_{body} is the human body mass (kg. $person^{-1}$), M_i is the per capita daily consumption per kg body mass ($g.(kg\ person.day)^{-1}$), C_{iw} is the concentration of salmonella in the water ($number.ml^{-1}$), V_{prod} is the volume of irrigation water trapped by the products ($ml.g^{-1}$), w is the removal due to washing ($\log_{10}$) of the concentration of

(or the amount of) pathogens, *k* is a kinetic decay constant of Salmonella in the field (day^{-1}) and *t* is the time interval between the last irrigation and harvest (days).

The authors of the study retained the following values:

Table 2: Parameter values retained by Amha et al. (2015)

Parameter	Mean	Distribution	Units
M_{body}	71		kg
M_i	Lettuce: 0.42 Cabbage: 0.064 Cucumber: 0.339		$g.day^{-1}.kg^{-1}$ inhabitant
C_{iw}	6932.5	Triangular (a= ; c= ; b=)**	$Salmonella.L^{-1}$
V_{prod}	Lettuce: 0.0178 Cabbage: 0.0889 Cucumber: 0.00360		$mL.g^{-1}$
w	washed: 1 Unwashed: 0		Log_{10}
k	0.8107		day^{-1}
t	1		Day

**: all the distributions have been well informed (not reported in this table for the sake of simplicity), with the exception of those on the concentrations of salmonella. The authors give us many indications on this without giving us the values of the constants a, c and b.

For the scenario of consumption of unwashed vegetables ($\lambda_{w=0}$), the term 10^{-w} is equal to 1 (i.e. w=0; there is no reduction due to washing ... since washing does not exist (!)).

The probability of infection $P_1(\lambda)$ for 1 exposure is assumed to follow a Beta-Poisson model:

$$P_1(\lambda) = 1 - \left[1 + \left(\frac{\lambda}{ID_{50}}\right)\left(2^{1/\alpha} - 1\right)\right]^{-\alpha} \quad (2)$$

where α is an "infectivity constant" and ID_{50} the median infectivity dose, i.e. the dose at which the probability, $P_1(\lambda)$ of being infected is equal to 0.5.
The annual risk $P_{iA}(\lambda)$ of being infected is then equal to:

$$P_{iA}(\lambda) = 1 - [1 - P_1(\lambda)]^n \quad (3)$$

where *n* is the number of contamination exposure events per year.
The probability of illness can then be estimated by considering the proportion of infected people developing the disease:

$$P_{ill/y.} = P_{iA}(\lambda) \times (ill:inf) \quad (4)$$

where $(ill:inf)$ is the proportion of infected people developing the disease.

We can then deduce a number of years of life lost (DALYs) per person and per year (pppy) with the following equation:

$$DALYs = P_{ill/y.} \times DALYs_{per\ case} \quad (5)$$

The parameters retained by the authors of the study are:

Table 3: Parameter values proposed by Amha et al. (2015)

Parameter	value	unit
α	0.3126	-
ID_{50}	23.6	salmonella
n	365	Event number
$(ill:inf)$	1	-
$DALYs_{per\ case}$	0.034	year

III. QMRA without considering the randomness of certain variables

First assess the risk of disease:

- per exposure event,
  - You must first calculate the average dose of pathogens ingested for one exposure (one meal) using equation (1), and this for each vegetable and distinguishing between the 2 situations "washed" and "unwashed";
  - You must then calculate the risk of infection using the beta-Poisson model as a Dose-Response model (equation (2));
- Then per year,
  - You must consider the number of exposures per year (i.e. the number of meals with the food considered) (equation (3));
  - You have to consider the ratio between sick people and infected people (equation (4));
- and finally, the "cost" of possible infections in number of years of life lost per person per year (DALYs), using equation (5).

IV. Choosing a probability distribution and using it to simulate a distribution of random values

In this tutorial, we propose to see how to simulate uniform, normal and triangular random distributions.

IV.1. The simulation of a uniform distribution between a and b

The basic tool is often a generator of random numbers between 0 and 1. Under EXCEL, if we write in a cell =ALEA(), this cell will take a random value between 0 and 1; this value will change "spontaneously" as soon as you modify your file. You can very easily switch from this value to a uniform random selection between 2 values a and b by the operation =a+(ALEA()*(b-a)), but there is another EXCEL function which can directly offer you a random draw between a and b: then write =ALEA.ENTRE.BORNES(a;b).

We suggest that you use here only the function ALEA(), because a generator of numbers uniformly distributed between 0 and 1 makes it possible to generate varied distributions as soon as we have access to their distribution function.

We ask you to test on 100 values if the ALEA function allows you to generate a uniform random distribution of values between 0 and 1.
To do this, proceed in 4 steps:

1. repeat the operation =ALEA() on 100 successive cells;
2. calculate numbers of values less than 0, 0.1, 0.2, 0.3 …, 0.9 et 1.0 using the function =NB.SI(plage;critère) *(you get an approximation of the distribution function of the simulated distribution to the previously mentioned values)* ;
3. change from these numbers to the numbers of values between 0 and 0.1, between 0.1 and 0.2…, and between 0.9 and 1.0 by differences;
4. then draw under EXCEL a histogram of the numbers of values obtained for each of the previous intervals.

IV.2. The simulation of a normal distribution of mean m and variance σ^2

In general, if we consider a random variable X which follows a uniform distribution between 0 and 1, and if we consider that a random selection x_0 of X corresponds to the value y_0 of the distribution function of a random variable Y ($x_0 = \int_{-\infty}^{y_0} p(y)dy$), then the value y_0 thus obtained follows the law of distribution of the random variable Y.

Concretely, if we consider a random variable following a normal law $N(m,\sigma^2)$, of mean *m* and standard deviation σ, its density function p(y) can be obtained by the EXCEL function =LOI.NORMALE.N(y;m;σ;FAUX) and its distribution function by the EXCEL function =LOI.NORMALE.N(y;m;σ;VRAI). But it is above all the inverse function of the distribution function which is useful here!
Thus, if you write in an EXCEL cell =LOI.NORMALE.INVERSE.N(ALEA();m;σ), you generate a random value which follows a normal distribution of mean *m* and standard deviation σ.

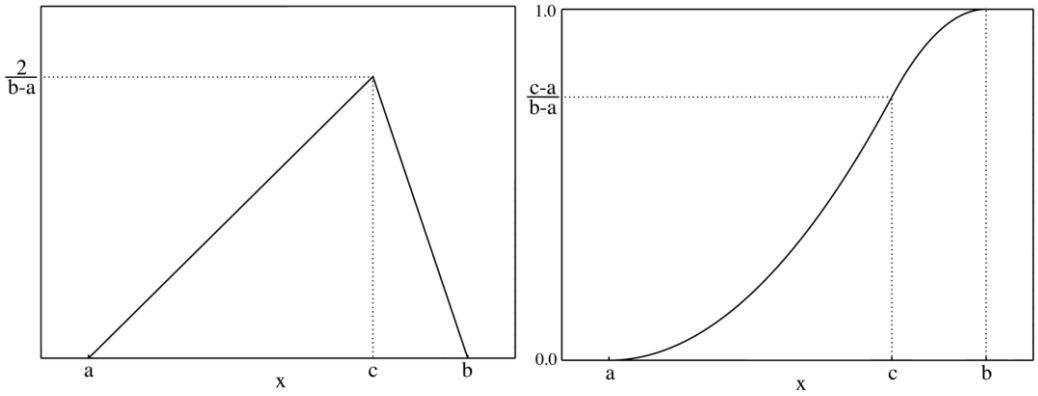
We ask you to test on 100 values if the function LOI.NORMALE.INVERSE.N(ALEA();m;σ) allows you to generate a normal distribution of expectation *m* and standard deviation σ.
To do this, proceed in 4 steps with m=2 et σ=1:

1. repeat the operation =LOI.NORMALE.INVERSE.N(ALEA();1;0.5) on 100 successive cells;
2. calculate the numbers of values less than -1, -0.5, 0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5 and 4 using the function =NB.SI(cell set;criteria) *(you get an approximation of the distribution function of the simulated distribution to the previously mentioned values)*;
3. go from these numbers to the numbers of values lower than -1, between -1 and -0.5, … between 2.5 and 3, and finally greater than 3;
4. then draw under EXCEL a histogram of the numbers of values obtained for each of the previous intervals.

IV.3. Simulating a triangular distribution

Maybe you've never heard of this probability distribution. It is simple to represent graphically, and its density and distribution functions easy to calculate. Beyond that, it is sometimes the best distribution function adjustable to the experimental data as is the case for certain measurements made in the study mentioned above, in particular concerning the distribution of the values of Salmonella concentrations in the treated wastewater.

A triangular distribution has density and distribution functions that can be represented by the graphs below:



For each of the 4 intervals to be considered (]-∞;a], [a;c], [c;b] and [b;+∞[), the distribution function of the triangular distribution can be written:

- over the interval]-∞; a], $F(x) = 0$;
- over the interval [a; c], $F(x) = \frac{(x-a)^2}{(b-a)(c-a)}$;
- over the interval [c; b], $F(x) = 1 - \frac{(b-x)^2}{(b-a)(b-c)}$;
- over the interval [b; +∞[, $F(x) = 1$.

Unfortunately, EXCEL has not programmed a function for this distribution, let alone for the inverse function to its distribution function.

We suggest that you set up what is necessary to simulate a triangular distribution corresponding to the numbers of salmonella per L of treated wastewater. Concretely:

1. we need to find the triangular function that the authors used, because Amha et al. (2015) did not tell us the values taken by a, c and b in their study. We can, however, assume that a≤900 salmonella.L^{-1} et c≥16000 salmonella.L^{-1}, values corresponding respectively to their minimum and maximum measured values. We have the following graph in the *"Supplementary Information"*.

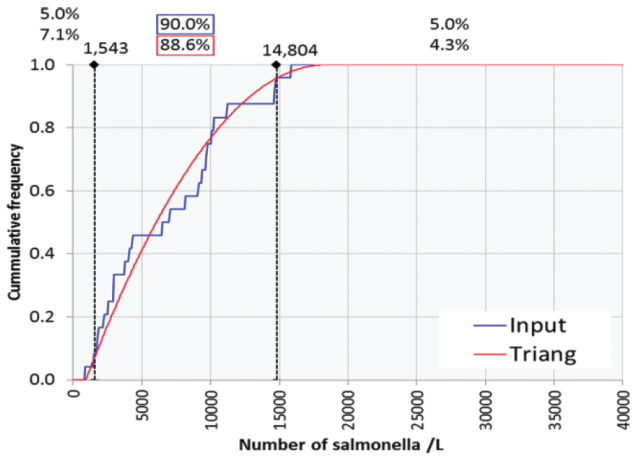


Figure S4 | Fit comparison using cumulative frequency between input data for detected *Salmonella* spp. in treated wastewater samples and fitted curves, using triangular distribution. The blue curve indicates input data and the red curve indicates triangular distribution. The delimiters on the graph signify the confidence intervals for the input data and for the triangular distribution curves.

But above all and using Table 1, we have an estimate of the distribution function at the 24 measured concentration values.

C_{iw}	F_{exp} (C_{iw})
910	0.04
1530	0.08
1780	0.13
1850	0.17
2290	0.21
2590	0.25
3020	0.29
3030	0.33
3790	0.38
4150	0.42
4390	0.46
6540	0.50
7080	0.54
8220	0.58
9130	0.63
9460	0.67
9770	0.71
9820	0.75
10150	0.79
10280	0.83
11240	0.88
14670	0.92
14800	0.96
15890	1.00

We suggest you to optimize the values of a, c and b so that the simulated function $F_{sim}(C_{iw})$ is closest to the experimental function by minimizing the sum of the squares of the deviations $\sum\left(F_{sim}(Ciw) - F_{exp}(Ciw)\right)^2$.

2. You can then generate random values according to a triangular law by using the reciprocal of the distribution function either for any random draw of a value of x uniformly distributed between 0 and 1:
   - If $x \geq F_{sim}(c)$, calculate y such that $y = b - \sqrt{(1-x) \times (b-a) \times (b-c)}$;
   - If $x \leq F_{sim}(c)$, calculate y such that $y = \sqrt{x \times (b-a)(c-a)} - a$;
3. For 100 random values then generated, check as before that you get the expected triangular distribution.

V. Risk assessment considering a variability in the *Salmonella* concentrations

In the following, only infections associated with eating lettuce that has been washed before are considered (M_i=0.42 g.j^{-1}.kg inhabitant^{-1} ; w=1). Only adults with 71 kg body weight are considered. **By contrast, it is assumed that the virus concentration of the water c_{iw} follows the triangular distribution previously adjusted.**

By setting up a table with 6 columns (ALEA() ; C_{iw} ; λ_{washed} ; $P_1(\lambda_{washed})$; $P_{iA}(\lambda_{washed})$; DALYs), simulate on 1000 lines 1000 random draws of the concentration of salmonella in the water, and assess its consequences on the values of λ_{washed}, de $P_1(\lambda_{washed})$, de $P_{iA}(\lambda_{washed})$ et de DALYs.

Calculate for these 1000 realizations the average values of C_{iw}, λ_{washed}, $P_1(\lambda_{washed})$, $P_{iA}(\lambda_{washed})$ and DALYs.

Compare these averages with the results obtained in section III. What thoughts do these comparisons suggest to you?

Annex: general information on quantitative microbial risk assessment (QMRA)

The **Quantitative Microbial Risk Assessment** (**QMRA**) makes it possible to reason the management of a practice according to:

- the risks it generates (with regard to other practices); and
- the costs and/or benefits associated with this practice (with regard to other practices) and the optimal nature of expenditure on this type of item with regard to the possible priority interest of expenditure on other items also relating to health

The tool is in a certain way complementary to the HACCP technique (Hazard Analysis Critical Control Point) commonly used by manufacturers and easier to implement without good scientific background. QMRAs are used in a wide variety of contexts: drinking water, food, air…. The tutorial is dedicated to the reuse of wastewater for agricultural irrigation.

The ***DALY*** (for ***Disability Adjusted Life Years***) is sometimes used as an index measuring the weight (burden) of a disease on a person. This index expresses the lost quantity of quality life due to disability or illness-related death, compared to the remaining life without disability in the absence of illness. It is very often expressed in lost year (s) either per case of disease (in pcd for per case of disease) or per person and per year (in pppy for per person per year). For this 2nd type of expression (pppy), it is calculated by adding the years of life lost following a premature death to the years lived with a disability:

$$DALY = YLL + YLD \tag{A.1}$$

where ***YLL*** and ***YLD*** are the **Years of Life Lost** and the **Years Lost due to Disability**, respectively.
YLLs result from premature death. They are calculated from the age-specific death rates (number of deaths per person (!) and per year) multiplied by the remaining life expectancy of the population at the age at which death occurs. The basic formula for YLL is as follows for a given causal factor and a specific age and sex population (regardless of social specifics):

$$YLL = N \times L \tag{A.2}$$

where N is the death rate of the given population per person per year, and L is its remaining life expectancy. Equation (A.2) could be generalized to a heterogeneous population (age, sex, social condition, etc.) by replacing it with a sum and/or an integral.
YLDs result from an incapacity (disease, handicap, etc.). They are calculated from the number of cases of disease multiplied by the average duration of the disease and by a severity factor ranging from 1 (very disabling situation) to 0 (perfect health):

$$YLD = I \times DW \times L = P \times DW \tag{A.3}$$

where I is the number of disease cases per person per year, DW the disability weight or severity factor varying between 0 and 1, L the average duration of disease until remission or death, and P the cumulative duration of the illness (number of years per year and per person)

The DALY index considers the acute effects (during the disease phase) but also the delayed and chronic effects (including morbidity and mortality). It allows you to compare the effects of different diseases, for example cancer and gastroenteritis. It can consider the social context (more or less marked effects/consequences depending on the age, sex and health context of the people). It authorizes objective management of health risks based on acceptability thresholds.

In the 1st step, we can then define a DALY threshold value expressed per person and per year (pppy) **beyond which the risk is no longer tolerable**. This threshold may depend on the context: in developed countries, the requirements may be high, while imposing too high constraints in developing countries may be financially unrealistic and/or result in suboptimal use of money for health.
We often use a threshold of 10^{-6} pppy (requirement often retained for drinking water) sometimes considered too severe and then replaced by a threshold of 10^{-4} pppy (requirement sometimes retained for food):

$$Tolerable\ DALYs\ \ pppy = 10^{-6} \tag{A.4}$$

We can then deduce a tolerable risk of disease, in practice a probability of disease per person per year:

$$Tolerable\, disease\ risk\ \ pppy = \frac{Tolerable\, DALYs\ \ pppy}{DALYs\ \ per\ case\ of\ \ disease} \tag{A.5}$$

Any infection that does not give rise to disease (part of the population may be immune, some people may be temporarily healthy carriers of pathogens), we deduce from the tolerable risk of disease a tolerable risk of infection which is greater than it:

$$Tolerable\ P_a\ \ \ pppy = \left(Tolerable\ disease\ \ risk\ \ pppy\right) \times R_{Infection/Disease} \tag{A.6}$$

where $R_{Infection/Disease}$ is the ratio (≥1) between the number of infections and the number of diseases. The tolerable risk of infection per person per year can be converted into a tolerable risk of infection per person and per event of exposure to the pathogen:

$$Tolerable\ P_a^{\ *}\ \ \ pppee = Tolerable\ nfection\ \ risk\ \ \ pppy = 1 - \left(1 - Tolerable\ P_a\right)^{1/n} \tag{A.7}$$

where n is the number of exposures to the pathogen events per person per year.

Table 1: example of DALY values per case of disease (WHO, 2006)

Pathogen	DALY pcd (per case of disease)	Ratio disease/infection ($R_{Disease/Infection}$)
Rotavirus		
(1) IC[a]	1.4×10^{-2}	0.05[b]
(2) DC[a]	2.6×10^{-2}	0.05[b]
Campylobacter	4.6×10^{-3}	0.7
Cryptosporidium	1.5×10^{-3}	0.3

[a]: IC: industrialized countries; DC: developing countries.

[b]: the disease/infection ratio is low for Rotavirus because people are immune after the age of about 3.

<u>In a 2nd step, we have to assess the risk of infection of people from their exposure</u> to pathogens. This exposure can be of 3 types:

- exposure through food: number of pathogens ingested per exposure;
- exposure through water: number of pathogens ingested per quantity of water drunk;
- exposure through air: number of inhaled pathogens then passing from the respiratory tract to the digestive tract, number of pathogens causing skin infections, etc

We limit ourselves here to the number of pathogens ingested with food, in practice market garden products irrigated by wastewater and eaten raw.

Several laws have been proposed to describe the relationships between exposure and infection (one often speaks of "dose-response law"). The most used models are the exponential model (Eq. (8a)) and the Beta-Poisson model (Eq. (8b)) :

$$P_a^{\ *}(d) = 1 - e^{-r \times d} \tag{A.8a}$$

$$P_a^{\ *}(d) = 1 - \left[1 + \left(\frac{d}{N_{50}}\right) \times \left(2^{1/\alpha} - 1\right)\right]^{-\alpha} \tag{A.8b}$$

where d is the dose (or number) of pathogens ingested during exposure, and r, α and N_{50} three constants. N_{50} corresponds to the quantity of pathogens to ingest so that $P_a^{\ *}(d)=0.5$. Equation (8a) can still be written:

$$P_a^{\ *}(d) = 1 - (0.5)^{d/N_{50}} \quad \text{with} \quad r = -\frac{\ln(0.5)}{N_{50}} \tag{A.9}$$

Other models can be used like the beta-binomial model.

In a 3rd step, we have to define a maximum exposure dose per exposure episode so that the ingested dose never exceeds the tolerable risk of infection. If we use the exponential model, we obtain:

$$Tolerable\,d = N_{50} \times \frac{\ln\left(1 - Tolerable\,P_a^{*}\right)}{\ln(0.5)} \quad \text{(A.10)}$$

If we use the Beta-Poisson model, we obtain:

$$Tolerable\,d = \left(\left(1 - Tolerable\,P_a^{*}\right)^{-1/\alpha} - 1\right) \times \left(\frac{N_{50}}{\left(2^{1/\alpha} - 1\right)}\right) \quad \text{(A.11)}$$

In practice, the tolerable quantities of pathogens ingested are very low: "on average" very much below 1 so that there is effectively only a very small proportion of the infected population and, above all, that this proportion remains negligible compared to what would be "natural" infections without wastewater reuse (the idea is that the risk due to wastewater reuse is negligible invisible in relation to the epidemiology of populations in absence of this practice).

In the 4th (and last) step, we have to go from the tolerable quantity of pathogens that can be ingested to a quality of water used in irrigation, and this possibly considering the fact that the removal in $\log_{10}$ of the quantities of pathogens present can occur not only during wastewater treatment but also afterwards. In the case of the reuse of wastewater (treated or not) in irrigation, several assumptions are often used:

A1: Assumptions relating to the efficiency of wastewater treatment before recycling in irrigation. It depends on the nature of the treatments, in particular tertiary treatments. It is sometimes admitted (wrongly!) that there are correlations between pathogen and bio-indicator contents (especially *E.* coli or thermo-tolerant coliforms). In the same way, it is often assumed that the $\log_{10}$ removal of pathogens is similar to the $\log_{10}$ removal of bio-indicators, while it is false, but the acceptability of this hypothesis depends on the pathogens and selected bio-indicators;

A2: Assumptions relating to the fate of pathogens during irrigations: it is often assumed that the pathogens present in the water retained by the consumed parts of the crops remain on these parts (even after drying the excess water);

A3: Assumptions relating to the fate of pathogens between irrigation and harvest: depending on the climate and the pathogens, an exponential decrease in the number of pathogens (or their infectivity) over time is often proposed;

A4: Assumptions relating to the effects of post-harvest operations with:

- the impact of washing consumed parts;
- the effect of peeling, or cutting with possible transport of pathogens to the consumed parts;
- the effect of cooking for foods not eaten raw.

The approach can be more complete (and complicated) by considering heterogeneities within the population (sex, age, height, weight, cultural/social background and eating habits, etc.).

Many points remain poorly understood and lead to assumptions and/or values that are very uncertain, or even already open to criticism.