Modelling the effect of chilling on the occurrence of *Salmonella* on pig carcasses at study, abattoir and batch levels by meta-analysis

CrossMark

Ursula Gonzales-Barron [a,b,*], Vasco Cadavez [b], James J. Sheridan [a], Francis Butler [a]

[a] *UCD School of Biosystems Engineering, University College Dublin, UCD Belfield, Dublin 4, Ireland*
[b] *Food Safety and Technology Group, CIMO Mountain Research Centre, Polytechnic Institute of Braganza, Portugal*

ABSTRACT

The effect of chilling on the occurrence of *Salmonella* on pig carcasses was investigated at study, abattoir and batch level by meta-analysis. Both the fixed-effects and random-effects model confirmed ($p < 0.05$) the significant effect of chilling in decreasing *Salmonella* occurrence on pig carcasses; although the random-effects solution was preferred to account for the significant variability in effect size ($p < 0.001$) estimated from the 13 primary studies considered, the 32 abattoirs surveyed, and the 51 sampled batches. Conservatively, it can be said that chilling reduces the *Salmonella* incidence on pig carcasses by a mean ratio of ~1.6 (95% CI: 1.0–2.6). Multilevel meta-analysis models investigating study characteristics that could explain the heterogeneity (τ^2) in the true effect size among primary studies ($\tau^2 = 0.578$), among surveyed abattoirs ($\tau^2 = 0.431$), and among sampled batches ($\tau^2 = 0.373$), revealed that study size (represented by the moderating variables of 'total sample size' and 'number of batches sampled in an abattoir') and 'carcass swabbed area' have a significant impact ($p < 0.05$) on the measured effect size of chilling. The fact that swabbed area explained between 56 and 62% and total sample size between 23 and 38% of the total heterogeneity in the chilling true effect size, indicates that differences in experimental design greatly affect our substantive conclusion about the effect of chilling on *Salmonella* occurrence. This inconsistency to elucidate the effect of chilling arises because of the many factors influencing both the performance of the chilling operation and the measurement itself. Meta-analysis was not only instrumental to show that small-size studies (i.e., only one batch sampled per abattoir, total number of sampled carcasses per batch < 50) and small swabbed areas (<100 cm^2) lead to imprecise and even conflicting conclusions, but most importantly, enabled definition of the characteristics of a well-designed study having a minimum statistical power to produce precise results. A sound experimental design derived by multilevel meta-analysis consists of swabbing carcass areas of at least 500 cm^2 from 25 pre-chill and 25 post-chill carcasses from a single production batch, with a minimum of two batches sampled per surveyed abattoir. If the survey were to be conducted in more than one abattoir, the total sample size should not be less than 400. Two methods to test for publication bias, a common problem in meta-analysis, suggested that whilst the presence of unpublished small-size studies is probable, it is not likely that this would significantly bias the overall chilling effect estimated in this study.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

Meta-analysis concerns the statistical summarisation of the results of a large collection of independently conducted primary studies on one specific research question (Glass, 1976). The primary aim of meta-analysis is to produce a more precise estimate of the effect of a particular intervention, with increased statistical power, than is possible using only a single study (Sutton et al., 2001). With meta-analysis, it is also possible to explain differences in the study outcomes by coding study characteristics such as research design features, data collection procedures or type of subjects sampled (Hox and de Leeuw, 2003). In *classical*

meta-analysis (Glass, 1976), if data is binary, the effect size of a treatment or intervention is estimated for each primary study by means of a standardised measure or parameterisation such as odds ratio or relative risk; if data is continuous, it is estimated by the difference between treatment and control. In a fixed-effects approach, combining studies is simple as they can be regarded as direct replications of each other and one can assume that the possible differences between study outcomes are due to sampling error. However, heterogeneity in primary study outcomes is expected as different studies employ different sampling methods, different experimental manipulations or measure the effects with different methodologies. To address this heterogeneity, a random-effects model is the best choice as it assumes that study outcomes vary not only because of random sampling effects (within-study variations), but also because of real differences between studies. If heterogeneity among primary studies is present, the next goal of meta-analysis is to attempt to identify the study characteristics or moderators that explain the

* Corresponding author at: Food Safety and Technology Group, CIMO Mountain Research Centre, Polytechnic Institute of Braganza, Portugal. Campus de Santa Apolónia, Apartado 1172, 5301-854 Bragança Portugal. Tel.: +351 273 303 325.
E-mail address: ubarron@ipb.pt (U. Gonzales-Barron).

differences between study outcomes. For this, an original new analytic direction was taken by Raudenbush and Bryk (1985), who argued that a meta-analysis could be regarded as a special case of *multilevel analysis* using hierarchical linear models, with subjects between studies at the first level and studies at the second level. The major advantage of using multilevel analysis instead of classical meta-analysis methods is its flexibility as it is simple to include moderators as explanatory variables in the regression model (Hox and De Leeuw, 2003).

Whilst there was a rapid increase of meta-analyses conducted in evidence-based medicine on the past two decades (Whitehead, 2002), realisation of the potential of meta-analysis in agriculture public health and food safety only began in the past few years (Den Besten and Zwietering, 2012; Gonzales-Barron et al., 2008; Gonzales-Barron and Butler, 2011; Grieg et al., 2012; McQuestin et al., 2009; Patil et al., 2004; Sánchez et al., 2007; Sargeant et al., 2006; Vialette et al., 2005; Wilhelm et al., 2011). In food safety research, meta-analysis may be conducted to address a broad range of research questions such as disease incidence, consumer practices, prevalence of microorganisms, effect of interventions pre- and post-harvest, etc. Within the risk analysis context, before any risk management option can be established, research is necessary to identify both (i) the control points in the food chain where interventions or treatments would be effective in reducing contamination, exposure to food-borne pathogens or those that reduce growth, and (ii) the interventions that are likely to be efficacious. Nevertheless, the results of a potential intervention strategy might be conflicting among primary studies, and different studies normally fail to provide the same level of confidence for effectiveness because of differences in study design, statistical power or sample size. Thus, meta-analysis becomes useful in the field of food safety for the identification, appraisal and summarisation of results from large quantities of research. Apart from the possibility to quantify the effect of the intervention under study with increased confidence, meta-analysis can also reveal a clearer picture of the state of knowledge from the body of information contained in all primary studies (Gonzales-Barron and Butler, 2011), and can provide a statistical distribution of the effect of an intervention that can be used in exposure assessment models (Gonzales-Barron et al., 2009).

Performing a meta-analysis, Gonzales-Barron et al. (2008) confirmed that the chilling operation has a significant effect on the reduction of pathogenic *Salmonella* occurrence in pig carcasses. However, the effect size of chilling was estimated using only pooled numbers of positive and tested carcasses per primary study; being a case of a classical meta-analysis integrating only summarised outcomes. As other primary studies became available, it was realised that, for most of the studies, the *Salmonella* occurrence data of pigs before and after chilling could be partitioned by abattoir and, within abattoir by sampling visit or batch; and moreover, that some study characteristics could be coded to attempt to explain the different outcomes from the primary studies. Thus, the objectives of this research are: (i) to compile all publicly accessible published and non-published findings on the effect of chilling on *Salmonella* occurrence on pig carcasses, and quantitatively summarise these outcomes on a study level, abattoir level and batch level; (ii) to assess the presence of heterogeneity in the observed effect sizes of chilling at each of these three levels, and if present, to evaluate its causes by means of multilevel meta-analyses using coded study characteristics such as number of abattoirs surveyed, number of total sampled batches (visits), total sample size and swabbed area; and (iii) to illustrate some methods to appraise likely publication bias, a common artefact in meta-analysis studies. Interpretation of results and applications in the experimental design of future food safety research are discussed.

2. Methodology

As indicated by Sargeant et al. (2005), to perform a meta-analysis, three important facets are to be considered: population, intervention or treatment and measured outcome. The problem statement in this study was the estimation of the overall effect of chilling on the prevalence of *Salmonella* on pig carcasses during pork production. The *population* was specified as eviscerated pig carcasses after meat inspection in slaughterhouses. The *treatment* is represented by the chilling stage during pork processing, which includes cooling and cold storage (18 to 24 h) at ~5 °C. The *measured outcome* is the detection of *Salmonella* on the pig carcass surface. Following the systematic review protocol presented by Sargeant et al. (2005), electronic searches were carried out to identify published and unpublished primary studies additional to those already found for the previous meta-analysis conducted by Gonzales-Barron et al. (2008). After assessing all the information presented in every study, thirteen primary studies were considered appropriate for inclusion in the meta-analysis models. This extends our earlier study (Gonzales-Barron et al., 2008) by four additional data sets. After problem statement formulation and data collection, a parameterisation or measure unit of the intervention's effect size needs to be determined. The effect size (θ) refers to the degree to which the hypothetical phenomenon (i.e., decrease in *Salmonella* prevalence due to chilling) is present in the population (i.e., pig carcasses during processing at slaughterhouses). The parameter measuring the effect size of an intervention is a common metric that permits direct comparison and summation of primary studies (Noble, 2006). Table 1 compiles some of the most widely-used effect size measures used for meta-analysing continuous and binary outcome data. If the sample size of the individual studies is at least higher than 20, it is usually reasonable to assume that the sampling distribution of the outcomes is normal (Bryk and Raudenbush, 1992). The variance of the sampling distribution is often known from statistical theory, although in some cases, a transformation is needed to achieve normality and known sampling variance (Table 1). Because the data generated by occurrence studies is binary (i.e., a pig carcass tests either positive or negative for *Salmonella*), the possible parameters to measure effect size are relative risk, odds ratio or probability difference. Relative risk (*RR*) was chosen for being less susceptible to differences in study protocols and suitable for risk assessment models (Gonzales-Barron et al., 2008). As shown in Table 1, the normality assumption is theoretically met by the natural logarithm transformation (log *RR*), and the sampling variance is known.

2.1. Description of data sets

The outcome data from the thirteen primary studies were available on n_T pig carcasses in the post-chill group (treated group) and n_C pig

Table 1
Measures of effect size and their sampling variance used in meta-analysis. Sub-indexes T and C stand for treated group and control group, respectively.

Measure	Estimator	Transformation	Sampling variance
Mean	$\bar{x}$	–	$\dfrac{\sigma^2}{n}$
Difference in means	$g = \dfrac{\bar{x}_T - \bar{x}_C}{\sigma}$	–	$\dfrac{n_T + n_C}{n_T n_C} + \dfrac{g^2}{2n_T + n_C}$
Standard deviation	σ	$\log(s) + \dfrac{1}{2} df$	$\dfrac{1}{2} df$
Correlation	r	$\dfrac{1}{2} \log\left[\dfrac{1+r}{1-r}\right]$	$\dfrac{1}{n-3}$
Proportion	p	$\log\left[\dfrac{p}{1-p}\right]$	$\dfrac{1}{np(1-p)}$
Relative risk	$RR = \dfrac{p_T}{p_C} = \dfrac{s_T/n_T}{s_C/n_C}$	$\log RR$	$\dfrac{n_T - s_T}{n_T s_T} + \dfrac{n_C - s_C}{n_C s_C}$
Odds ratio	$OR = \dfrac{p_T(1-p_T)}{p_C(1-p_C)}$ $= \dfrac{s_T(n_T - s_T)}{s_C(n_C - s_C)}$	$\log OR$	$\dfrac{1}{s_T} + \dfrac{1}{s_C} + \dfrac{1}{n_T - s_T} + \dfrac{1}{n_C - s_C}$
Probability difference	$PD = p_T - p_C$ $= \dfrac{s_T}{n_T} - \dfrac{s_C}{n_C}$	–	$\dfrac{s_T(n_T - s_T)}{n_T^3} + \dfrac{s_C(n_C - s_C)}{n_C^3}$

carcasses in the pre-chill group (control group). The number of successes (*Salmonella*-positive carcasses) in the post-chill and pre-chill group is represented by s_T and s_C, respectively. In most of the primary studies, results were not only presented as pooled (total) numbers but also broken down by abattoir surveyed and by batch sampled. Meta-analyses were conducted at three levels: study-level (pooled results), abattoir-level and batch-level.

Table 2 compiles the occurrence data for the study-level meta-analyses. The study characteristics of 'carcass swabbed area' (*A*), 'number of abattoirs surveyed' (*Na*) and 'number of total sampled batches' (*Nb*) were extracted from each of the primary studies, and included in the data set as moderating variables, although the complete information was not available in three primary studies (Quirke et al., 2001; Epling et al., 1993; and Oosterom et al., 1985; Table 2). An additional moderating variable was the 'total sample size' (*N*) defined as $N = n_C + n_T$. The meta-analyses models conducted at study-level comprised fixed-effects, random-effects and multilevel-effects with the moderating variables *A*, *Na*, *Nb* and *N*. Table 3 shows the input data for the abattoir-level meta-analyses; this is, the occurrence data partitioned by abattoir. In this case, the study characteristics extracted were 'total sample size' (*N*), 'carcass swabbed area' (*A*), and 'number of batches sampled per abattoir' (*Nba*). Information on *Nba* was not available in two primary studies (Epling et al., 1993; and Oosterom et al., 1985; Table 3). Also, the studies of Algino et al. (2009), Quirke et al. (2001) and Saide-Albornoz et al. (1995) could not be used for the abattoir-level meta-analyses, as they did not present results separately for each of the sampled abattoirs. A fixed-effects model, random-effects model and multilevel models with *N*, *A* and *Nba* as covariates were fitted to the remaining suitable data compiled in Table 3. In some of the studies, the abattoir-level incidence data appeared further partitioned by sampling visits (sampled batches). At this level, two primary studies had to be removed (Epling et al., 1993; and Oosterom et al., 1985) as outcomes were not available broken down by batch. Meta-analyses conducted on the remaining suitable data (Table 4) were the fixed-effects, random-effects, and multilevel models with the moderating variables *A* and *N*.

2.2. Fixed-effects meta-analysis

In its simplest form, a fixed-effect meta-analysis can be carried out to make a *conditional inference* only about the *J* primary studies included in the meta-analysis (Hedges and Vevea, 1998). For example, a fixed-effects model without moderators provides an answer to the question: How large is the average true effect in the set of *J* studies? On the other hand, a fixed-effects meta-analysis can also be

conducted when there is an assumption that the possible differences between study results are due to sampling variance. In any case,

$$\theta_j = \Theta + \varepsilon_j \tag{1}$$

with θ_j the observed effect size in the primary study j, Θ the population effect size, and ε_j the residual error due to sampling variance. It is assumed that the ε_j have a normal distribution with mean zero and a true variance ξ^2. So, it follows that for a fixed-effects meta-analysis

Table 3

Abattoir-level occurrence of *Salmonella*-positive pig carcasses before and after chilling as detected in primary studies, and extracted study characteristics.

Coded study	Coded abattoir	Area of swab (cm²) (*A*)	Number of batches (*Nba*)	Pre-chill group (control)		Post-chill group (treated)	
				s_C	n_C	s_T	n_T
1	1	1000	2	14	55	9	55
	2	1000	1	2	30	3	20
2	3	300	2	6	60	3	60
	4	300	2	1	62	3	62
	5	300	2	1	60	0	60
3	6	100	1	0	23	4	23
	7	100	1	0	30	2	30
	8	100	1	2	45	2	45
	9	100	1	0	30	0	40
	10	100	1	1	30	0	15
	11	100	1	1	15	0	15
	12	100	1	0	15	0	15
4	13	1000	1	7	25	3	25
5	14	400	4	1	66	1	61
	15	400	3	12	50	2	47
	16	400	2	3	29	1	23
	17	400	3	2	30	1	30
6	18	100	2	4	60	0	60
	19	100	2	8	60	4	60
	20	100	2	4	60	16	60
7	21	1000	NS[a]	19	64	9	64
	22	1000	NS	4	71	2	71
	23	1000	NS	4	75	1	75
8	24	100	NS	61	188	72	188
10	25	600	2	0	39	0	39
	26	600	2	2	43	0	43
	27	600	2	23	51	4	51
	28	600	2	5	44	1	44
	29	600	2	1	49	0	49
11	30	1350	3	118	311	31	310
	31	1350	3	61	135	17	135

[a] Number of batches within each sampled abattoir not specified (NS) in the primary studies.

Table 2

Total occurrence of *Salmonella*-positive pig carcasses before and after chilling as reported in primary studies, and extracted study characteristics.

Coded study	Area of swab (cm²) (*A*)	Number of abattoirs (*Na*)	Number of total batches (*Nb*)	Pre-chill group (control)		Post-chill group (treated)		Reference
				s_C	n_C	s_T	n_T	
1	1000	2	3	16	85	12	75	Botteldoorn et al. (2003)
2	300	3	6	8	182	6	182	Bouvet et al. (2003)
3	100	7	7	4	188	8	183	Cutter (2003)
4	1000	1	1	7	25	3	25	Davies et al. (1999)
5	400	4	12	18	175	5	161	Duggan et al. (2010)
6	100	3	6	16	180	20	180	Minvielle (personal comm.)
7	1000	3	NS[a]	27	210	12	210	Oosterom et al. (1985)
8	100	1	NS	61	188	72	188	Epling et al. (1993)
9	100	10	10	10	112	25	112	Algino et al. (2009)
10	600	5	10	31	226	5	226	De Busser et al. (2011)
11	1350	2	6	179	446	48	445	Arguello et al. (2012)
12	1000	3	NS	6	419	1	419	Quirke et al. (2001)
13	100	3	9	3	270	1	270	Saide-Albornoz et al. (1995)

[a] Number of total batches not specified (NS) in the respective primary studies.

Table 4
Batch-level occurrence of *Salmonella*-positive pig carcasses before and after chilling as detected in primary studies, and extracted study characteristics.

Coded study	Coded abattoir	Coded batch	Area of swab (cm²) (A)	Pre-chill group (control)		Post-chill group (treated)	
				s_C	n_C	s_T	n_T
1	1	1	1000	13	25	9	25
		2	1000	1	30	0	30
	2	3	1000	2	30	3	20
2	3	4	300	2	30	1	30
		5	300	4	30	2	30
	4	6	300	0	29	1	29
		7	300	1	33	2	33
	5	8	300	1	30	0	30
		9	300	0	30	0	30
3	6	10	100	0	23	4	23
	7	11	100	0	30	2	30
	8	12	100	2	45	2	45
	9	13	100	0	30	0	40
	10	14	100	1	30	0	15
	11	15	100	1	15	0	15
	12	16	100	0	15	0	15
4	13	17	1000	7	25	3	25
5	14	18	400	0	21	0	16
		19	400	0	13	0	13
		20	400	0	16	1	16
		21	400	1	16	0	16
	15	22	400	7	19	0	16
		23	400	0	16	0	16
		24	400	5	15	2	15
	16	25	400	2	15	1	15
		26	400	1	14	0	8
	17	27	400	1	10	1	10
		28	400	1	10	0	10
		29	400	0	10	0	10
6	18	30	100	0	30	0	30
		31	100	4	30	0	30
	19	32	100	1	30	4	30
		33	100	7	30	0	30
	20	34	100	2	30	3	30
		35	100	2	30	13	30
10	24	36	600	0	20	0	20
		37	600	0	19	0	19
	25	38	600	1	22	0	22
		39	600	1	21	0	21
	26	40	600	22	28	4	28
		41	600	1	23	0	23
	27	42	600	4	21	0	21
		43	600	1	23	1	23
	28	44	600	1	24	0	24
		45	600	0	25	0	25
11	29	46	1350	27	100	12	100
		47	1350	57	99	14	98
		48	1350	34	112	5	112
	30	49	1350	9	45	0	44
		50	1350	36	50	11	49
		51	1350	16	40	6	44

model, $\theta_i \sim Normal\left(\Theta, \xi^2\right)$. Notice that the nomenclature θ_j and Θ is a general notation that refers to any size effect measure for the observed and the true effect size, respectively. In our particular case, it refers to the natural logarithm of the relative risk (log RR). Apart from the values of θ_j from each primary study, the standard error of the effect size $\sigma(\theta_j)$ must be calculated,

$$\theta_j = logRR = \log\frac{s_T/n_T}{s_C/n_C} \tag{2}$$

$$\sigma\left(\theta_j\right) = \sigma(logRR) = \left(\frac{n_T - s_T}{n_T s_T} + \frac{n_C - s_C}{n_C s_C}\right)^{0.5}. \tag{3}$$

To estimate the population effect size Θ, the observed effect sizes θ_j should be averaged. Since studies usually differ from each other in the reliability of estimating the true effect size, for instance, due to differences in study sizes; a weighted average is preferred with weights w_j equal to the precision in estimating the population effect size. Because this simple fixed-effects model assumes that the deviation of the observed effect sizes from the population effect size is due to sampling error alone, the precision can be defined as the inverse of the (estimated) sampling variance. The estimated population effect size and its standard error would be,

$$\hat{\Theta} = \frac{\sum_j w_j \theta_j}{\sum_j w_j} \tag{4}$$

$$\hat{\sigma}\left(\hat{\Theta}\right) = \frac{1}{\left(\sum_j w_j\right)^{0.5}} \tag{5}$$

with $w_j = 1/\sigma^2(\theta_j)$.

To evaluate whether the effect size is larger than zero, often a Wald test assuming normality for the dependent variable is used comparing the estimated weighted average divided by its standard error with a standard normal distribution $\left(U = \hat{\Theta}/\hat{\sigma}\left(\hat{\Theta}\right)\right)$. The U statistic is compared with a chi-square distribution with 1 df.

Because meta-analyses are performed retrospectively, in many situations studies may differ from each other due to differences in measuring protocols, in the population from which the sample is drawn, and in the magnitude and kind of treatment that is imposed; all of these giving rise to heterogeneity in study outcomes. A popular homogeneity test (Q test) is described by Cochran (1954), again assuming normality,

$$Q = \sum_j \frac{\left(\theta_j - \hat{\Theta}\right)^2}{\sigma^2\left(\theta_j\right)}. \tag{6}$$

When effect sizes across studies are homogeneous, Q follows a chi-square distribution with $(j - 1)$ df. If the hypothesis is rejected, there is evidence that there are additional sources of variability other than within-study sampling error. It is then common practice either to examine moderating variables; to divide the studies in homogeneous groups to perform separate meta-analysis; or to use a random-effects model. Contrarily to what is often stated in the literature (DerSimonian, and Laird, 1986; Whitehead, 2002), Viechtbauer (2010) argues that a fixed-effects model does not assume that the true effects are homogeneous, but instead that they can provide perfectly valid inferences on the size of the average effect to the set of studies included in the meta-analysis.

2.3. Random-effects meta-analysis

Most meta-analyses are based on sets of studies that are not exactly identical in their methods and the characteristics of their samples, which may introduce variability (i.e., heterogeneity) among the true effects. One way to model the heterogeneity is to treat it as purely random (Viechtbauer, 2010). In contrast to the fixed effects model, random models provide an *unconditional inference* about a larger set of studies from which the J studies included in the meta-analysis are assumed to be a random sample (Hedges and Vevea, 1998). It envisions a hypothetical population of studies that comprises studies that have been conducted, that could have been conducted or that

may be conducted in the future. The random-effects model addresses the question: How large is the average true effect in this larger population of studies? In a random effects model, each study investigates its own true effect size Θ_j,

$$\theta_j = \Theta_j + \varepsilon_j = \overline{\Theta} + v_j + \varepsilon_j \qquad (7)$$

with $\overline{\Theta}$ the mean true effect size and v_j the deviation of the true study effect size Θ_j from the mean true effect size. The values of v_j are normally distributed random effects with a mean of zero and a variance of τ^2. It follows that for a random-effects meta-analysis model, $\theta_j \sim Normal\left(\Theta, \tau^2 + \xi^2\right)$. In this approach, two sources of variation are distinguished: sampling variation (ξ^2) and variation between true effect sizes (τ^2). By including this additional component (τ^2), the standard error in the effect size estimates represents random variability at both the subject level and the study level. As in the fixed-effects approach, a weighted method is also used to estimate the mean true effect size and its standard error. The inverse variance weight or precision of the primary studies should then be corrected (w_j^*) by addition of the between-study variability term τ^2,

$$w_j^* = \frac{1}{\sigma^2\left(\theta_j\right) + \tau^2} \qquad (8)$$

with the variance τ^2 estimated from the Q-statistic (DerSimonian and Laird, 1986)

$$\hat{\tau}^2 = \frac{Q - (j-1)}{\sum\limits_j w_j - \left(\sum\limits_j w_j^2 / \sum\limits_j w_j\right)}. \qquad (9)$$

The mean true effect size Θ and its standard error $\sigma(\Theta)$ are now estimated with Eqs. (4) and (5) using instead the corrected weights w_j^*. To test the mean effect size for significance, the U statistic can be used, as described above for the fixed-effects model.

2.4. Multilevel meta-analysis

A meta-analysis can be considered a special case of *multilevel analysis* using hierarchical linear models, with subjects between studies at the first level and studies at the second level. In a multilevel meta-analysis, as in any other multilevel analyses, one usually starts from the random-effects model. If the between-study variance is shown to be noteworthy, study characteristics or moderators can be added to the model to account for at least part of the heterogeneity in the true effects. This leads to the mixed-effects model given by,

$$\theta_j = \Theta_j + \varepsilon_j = \beta_0 + \sum_{s=1}^{S} \beta_s X_{sj} + v_j + \varepsilon_j \qquad (10)$$

with X_1 to X_S S study characteristics. This model treats the moderator effects β_s as fixed and the v_j as random effects that distribute normally with a mean zero and a variance of τ^2. Yet, τ^2 now denotes the amount of residual heterogeneity among the true effects, or the variability among the true effects that is not accounted for by the S moderators included in the model. The goal of the analysis is then to examine to what extent the moderators influence the size of the average true effect size Θ. The resulting model is thus more general than the ones commonly used in classical meta-analysis. If no predictors are included, the model of Eq. (10) simplifies to the random-effects model (Eq. (7)), or if the variance in true effects is zero, to the fixed-effects model (Eq. (1)). The use of a regression equation $\left(\beta_0 + \sum\limits_{s=1}^{S} \beta_s X_{sj}\right)$ for the study characteristics is appealing for several reasons. First, different predictors can easily be

investigated together; meaning that even possible inter-correlations can be taken into account, which is not the case when separate meta-analyses are performed to investigate the moderating effects of study characteristics. Secondly, regression is a general approach that can be used for continuous as well as for categorical moderator variables (Van den Noortgate and Onghena, 2003).

2.5. Fitting of models

To estimate the parameters, maximum likelihood estimation (MLE) procedures are most frequently used. In MLE, residuals on both levels (v_j and ε_j of Eq. (10)) are assumed to be independently distributed. To test the fixed parameters of the model (β_s), the Wald test is used like in the traditional approaches, comparing the parameter estimate by the standard error with a standard normal distribution. Apart from the Q statistic, other measures can be computed to facilitate the interpretation of the estimated amount of between-study heterogeneity (τ^2). The I^2 statistics or intra-class correlation estimates the proportion of between-study variance from the total variance. This is analogous to using the proportion of explained variance in standard regression models to indicate the importance of specific predictor variables. Hunter and Schmidt (1990) pointed out that with a large number of studies, the power of the significance test is high, and small variances will become significant. However, when the number of studies is small, lack of significance for τ^2 does not imply that the outcomes are homogeneous. So, they propose a 25% rule of thumb; this is, if the intra-class variance is higher than 25% of the total variance, the variance between studies can be deemed as large enough to attempt to model it using available study characteristics. The H^2 statistic is the ratio of the total amount of variability in the observed outcome to the amount of sampling variability (If $\tau^2 = 0$, then $H^2 = 1$). The I^2 statistics and H^2 ratio are just monotonic transformations of τ^2, and for equations, refer to Higgins and Thompson (2002).

For testing the variance between true effect sizes (τ^2), the likelihood ratio test can be used. Because the model without the variance component is nested with the model with the variance component, the difference between the deviance scores, defined as -2 times the log likelihood, follows a chi-square distribution with 1 degree of freedom. The likelihood ratio along with the Bayesian Information Criterion (BIC) derived thereof, are goodness-of-fit criteria. The BIC measure however allows for direct comparison of models with different numbers of parameters. For models including moderators, an omnibus or moderator test (QM test) of all model coefficients is conducted that excludes the intercept. By default, the test statistics of the individual coefficients in the model are based on the normal distribution, whilst the moderators test is based on a chi-square distribution with S degrees of freedom (S being the number of moderators tested). Finally, after attempting to explain the heterogeneity among studies using the study characteristics, the QE test can be performed to test the non-explained (residual) variance using the statistic (Raudenbush and Bryk, 1985)

$$QE = \sum_j \frac{\left(\theta_j - \hat{\beta}_0 - \sum\limits_{s=1}^{S} \hat{\beta}_s w_{sj}\right)^2}{\sigma^2\left(\theta_j\right)} \qquad (11)$$

which follows a chi-square distribution with $J - S - 1$ degrees of freedom.

The general statistical notation of the multilevel model (Eq. (10)) for the *study-level* meta-analyses became, $\theta_j = \beta_0 + \beta_1 A_j + v_j + \varepsilon_j$, $\theta_j = \beta_0 + \beta_1 Na_j + v_j + \varepsilon_j$ and $\theta_j = \beta_0 + \beta_1 N_j + v_j + \varepsilon_j$, where the carcass swabbed area (A), the total number of abattoirs surveyed (Nv) and the total sample size (Na) were independently assessed as moderating variables. The multilevel models for the *abattoir-level* data, $\theta_j = \beta_0 + \beta_1 Nva_j + v_j + \varepsilon_j$, $\theta_j = \beta_0 + \beta_1 A_j + v_j + \varepsilon_j$ and $\theta_j =$

Table 5
Results of the study-level meta-analysis models for the natural logarithm of relative risk of *Salmonella* presence on pig carcasses after chilling in relation to before chilling.

Model	Fixed-effects	Random-effects	With moderators		
			Swabbed area	Total sample size	Number of abattoirs
# Entries (J)	13	13	13	13	13
Parameters					
Intercept	$-0.561\ (0.077)^{***}$	$-0.468\ (0.251)$	$0.232\ (0.291)^{ns}$	$0.332\ (0.458)^{ns}$	$-0.966\ (0.420)^{*}$
Swabbed area			$-0.001\ (0.0004)^{**}$		
Sample size				$-0.002\ (0.001)^{*}$	
# abattoirs					$0.139\ (0.096)^{ns}$
Heterogeneity					
Q test	$93.9\ (df = 12)^{***}$				
τ^2		0.578	0.250	0.418	0.529
I^2		85.0%			
H^2		6.66			
QM test			$8.92\ (df = 1)^{**}$	$3.87\ (df = 1)^{*}$	$2.12\ (df = 1)^{ns}$
QE residual heterogeneity			$25.01\ (df = 11)^{**}$	$38.66\ (df = 11)^{***}$	$87.83\ (df = 11)^{***}$
Goodness-of-fit					
Log-likelih.	-47.81	-17.16	-20.84	-21.40	-17.53
BIC	98.17	39.29	48.88	49.99	42.26

Significance codes: 0 '***', 0.001 '**', 0.01 '*', 0.05 '·' and 0.1 'ns' Non-significant.

$\beta_0 + \beta_1 N_j + v_j + \varepsilon_j$, assessed whether the number of batches sampled in an abattoir (*Nba*), the area swabbed (*A*) and the total sample size (*N*) could explain the variability in effect size among abattoirs. At *batch level*, the multilevel model of Eq. (10), specifically became $\theta_j = \beta_0 + \beta_1 N_j + v_j + \varepsilon_j$ and $\theta_j = \beta_0 + \beta_1 A_j + v_j + \varepsilon_j$. Notice that, in this work, the subscript *j* used in the above equations is general and indistinctively refers to the meta-analysed entry unit. So it refers to studies, abattoirs and batches, when meta-analyses were conducted at study level, abattoir level and batch level, respectively. Meta-analysis models were fitted in R version 2.14.2 (R Development Core Team) using the 'metafor' package (Viechtbauer, 2010), which provides functions for fitting the various models described above.

3. Results and discussion

The study-level (Table 5), abattoir-level (Table 6) and batch-level (Table 7) meta-analyses, by both the fixed-effect and random-effects approaches, produced comparable significant results that led to a general agreement that, overall, the chilling stage in pork processing decreases the recovery of *Salmonella* on pig carcasses, despite some contradictory outcomes produced by a few primary studies.

3.1. Study-level meta-analyses

The fixed-effects solution of the study-level meta-analysis suggested that chilling reduces the occurrence of *Salmonella* in pig carcasses by a factor of ~1.8 (95% CI: 1.5–2.0) (taking the inverse of the exponential of the estimated overall effect size $\Theta = -0.561$ in Table 5). Although there is evidence of heterogeneity in effect sizes among studies by the significant Q statistic, this fixed-effects model still provides valid inferences as long as they are restricted to the set of studies included in the meta-analysis (Viechtbauer, 2010).

The forest plot shown in Fig. 1 highlights the variability in effect size estimates and precision between studies; and the marker size illustrates the contribution of each study (weight) to the overall effect estimate for the fixed-effects solution. It should be noticed that weights are not directly related to sample size, but instead to the number of successes or failures in proportion to the sample size. Analysing the definition of the standard error of the Ln RR parameterization (Eq. (3)), inverse-variance weights will be small when the number of *Salmonella*-positive carcasses (successes) in either group (before or after chilling) is close to zero. For instance, study #13, with very low numbers of successes pre-chill and post-chill ($s_C = 3$, $s_T = 1$) in relation to its sample size ($n_C = n_T = 270$) produced a very low weight (Fig. 1). In contrast,

Table 6
Results of the abattoir-level meta-analysis models for the natural logarithm of relative risk of *Salmonella* presence on pig carcasses after chilling in relation to before chilling.

Model	Fixed-effects	Random-effects	With moderator	
			Swabbed area	Batches in an abattoir
# Entries (J)	31	31	31	27
Parameters				
Intercept	$-0.646\ (0.080)^{***}$	$-0.582\ (0.193)^{**}$	$0.0479\ (0.260)^{ns}$	$0.607\ (0.612)^{ns}$
Swabbed area			$-0.001\ (0.0003)^{**}$	
# batches in abattoir				$-0.589\ (0.273)^{*}$
Heterogeneity				
Q test	$91.57\ (df = 30)^{***}$			
τ^2		0.431	0.165	0.331
I^2		61.8%		
H^2		2.62		
QM moderators			$9.68\ (df = 1)^{**}$	$4.66\ (df = 1)^{*}$
QE residual heterog.			$35.35\ (df = 29)^{ns}$	$34.05\ (df = 25)^{ns}$
Goodness-of-fit				
Log-likelihood	-70.42	-47.27	-51.18	-41.2337
BIC	144.28	101.34	112.46	92.1240

Significance codes: 0 '***', 0.001 '**', 0.01 '*', 0.05 '·' and 0.1 'ns' Non-significant.

Table 7
Results of the batch-level meta-analysis models for the natural logarithm of relative risk of *Salmonella* presence on pig carcasses after chilling in relation to before chilling.

Model	Fixed-effects	Random-effects	With moderator	
			Swabbed area	Total carcasses
# Entries (*J*)	51	51	51	51
Parameters				
Intercept	−0.948 (0.103)***	−0.648 (0.172)***	0.111 (0.283)^{ns}	−0.181 (0.267)^{ns}
Swabbed area			−0.001 (0.0003)***	
Total carcasses				−0.006 (0.0025)*
Heterogeneity				
Q test	64.16 (df = 50)·			
τ^2		0.373	0.145	0.230
I^2		37.5%		
H^2		1.60		
QM moderators			11.92 (df = 1)***	5.54 (df = 1)*
QE residual heterogeneity			46.34 (df = 49)^{ns}	52.65 (df = 49)^{ns}
Goodness-of-fit				
Log-likelihood	−87.59	−84.37	−84.48	−86.96
BIC	179.11	176.57	184.63	185.59

Significance codes: 0 '***', 0.001 '**', 0.01 '*', 0.05 '·' and 0.1 '^{ns}' Non-significant.

weights will be large when the numbers of successes in both groups are high. In an extreme case, if there were no failure in both groups, the weights would be equal to infinity. That explains why study #8, not necessarily having the highest sample size ($n_C = n_T = 188$) but a high proportion of successes ($s_C = 61$, $s_T = 72$), was assigned the highest weight (Fig. 1). Yet, comparatively, smaller studies can still be given a larger weight when they have relatively more successes. For instance, study #7 with a much smaller sample size ($n_C = n_T = 210$) than study #12 ($n_C = n_T = 419$) was assigned a higher weight (Fig. 1). Thus, the inverse-variance weighting permits the consideration of small but well-designed studies, that otherwise would have been disregarded because of their lack of statistical power to show a significant different for the outcome of interest. As stated by Patil et al. (2004) weighting based on the standard error reflects the complete

uncertainty or confidence in a measure and is preferable to weighting based on sample size alone.

A visual examination of the forest plot (Fig. 1) gives an idea of the discrepancy among outcomes, with four (#3, 6, 8, 9) out of the thirteen studies reporting increase in *Salmonella* occurrence during chilling. This is not surprising given the several sources of variability among studies and abattoirs such as sampling site, extent of swab, chilling equipment, cross contamination of carcasses, level of *Salmonella* infection at slaughterhouses, differences in the microbiological protocol, season, year and country, among others. Accounting for the heterogeneity in the true log relative risks between abattoirs ($\tau^2 = 0.578$; Table 5) led to a meta-analysis of significantly better fit than the fixed-effects model, as indicated by the lower log-likelihood and Bayesian Information Criterion (BIC). Other measures can be computed to facilitate the
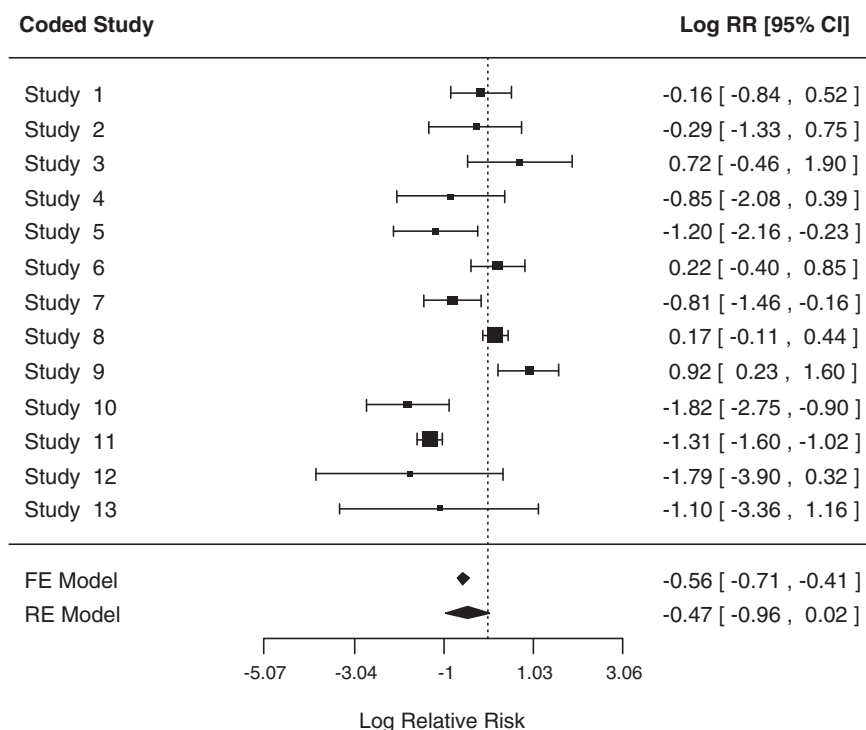
Fig. 1. Forest plot of the risk of *Salmonella* occurrence on pig carcasses after chilling relative to before chilling. Size of square markers reflects the weight assigned to the primary studies. Primary study estimates and overall fixed and random effects are shown with 95% confidence intervals.

interpretation of the estimated amount of between-study heterogeneity. In our case, the intra-class correlation (I^2) strongly underscored the presence of between-study variance ($I^2 = 85.0\%$ higher than the rule of thumb of 25%), and consequently it was a hint that some study characteristics should be coded to attempt to explain or understand such heterogeneity among the true effects. Similarly, the H^2 ratio was 6.66 ($\gg 1$), also indicating the high heterogeneity.

In the random-effect model, the overall effect size can be regarded as still significant although at a p-value lower than 0.10. This occurred because the addition of the relatively high variability at study level to the random variability at carcass level, produced an increase of the standard error of the overall effect size from 0.077 (fixed-effects) to 0.251 (random-effects; Table 5). As is common when significant between-study variability is present, the overall effect size approached by the random-effects was lower than the one estimated by the fixed-effects model. Contrarily to the fixed-effects solution, the overall effect size (or incidence reduction ratio due to chilling), obtained by the random-effects approach, can be generalised over a total population of studies, stating that after chilling the *Salmonella* occurrence is expected to decrease on average by a factor of ~1.6 ($\Theta = -0.468$ in Table 5) with a 95% CI: 1.0–2.6. In a meta-analysis conducted by Grieg et al. (2012), it was similarly found that dry chilling is an effective intervention for reducing generic *Escherichia coli* contamination of finished beef carcasses. Furthermore, the before-to-after incidence reduction factor for *E. coli* due to dry chilling, as recalculated from the original risk reduction ratio estimated by Grieg et al. (2012), was approximately 1.4 (95% CI: 1.2–1.6), which is a value very close to our reduction rate estimate for *Salmonella* on pig carcasses.

The power of multilevel meta-analysis becomes apparent when attempting to model the heterogeneity in the study outcomes. It was hypothesised that with larger studies in terms of more abattoirs involved (*Na*), more sampled batches (*Nb*), and overall higher total sample size (*N*), the observed effect size of chilling on *Salmonella* occurrence would become more precise; and that at least part of the heterogeneity found between studies could be explained by the differences in those experimental design moderating variables. In addition, in relation to the microbiological determination, it was hypothesised that larger swabbed areas (*A*), bringing about a reduction in measurement uncertainties, also cause effect sizes to be more precise. The multilevel meta-analyses showed that neither the number of abattoirs surveyed nor the number of total batches in a primary study explained some significant proportion of the between-study heterogeneity. However, opposite results were found for the study characteristics of swabbed area and total sample size only when the variables were entered separately in the models (Table 5). When both moderators were introduced into a single multilevel model, none of them were significant (results not shown).

The estimates of the effect size of chilling depend significantly on the extent of the carcass swabs, as indicated by the significant coefficient β_1 and the significant QM test for the moderating variable. The value of the coefficient β_1 indicates that an increase in one cm^2 in swabbed area corresponds to a reduction of -0.001 units in terms of the average log relative risk (Table 5). The estimated amount of residual heterogeneity is $\tau^2 = 0.250$, suggesting that (0.578–0.250)/0.578 = 56.7% of the total amount of heterogeneity between studies could be accounted for by including the 'swabbed area' in the model. However, the QE statistic indicates that the residual heterogeneity is still significant, and therefore other non-coded study characteristics may further explain the between-study variability. The total sample size, although a significant moderating variable, had less predictability than the swabbed area, as its inclusion in the multilevel model explained only half of the between-study heterogeneity ((0.578–0.418)/0.418 = 27.7%) that the swabbed area moderator explained. To facilitate the interpretation of the moderators, predicted average log relative risks as a function of the swabbed area and the total sample size are shown (Fig. 2). The observed log RR values are drawn proportional to the weights, and predictions are shown with
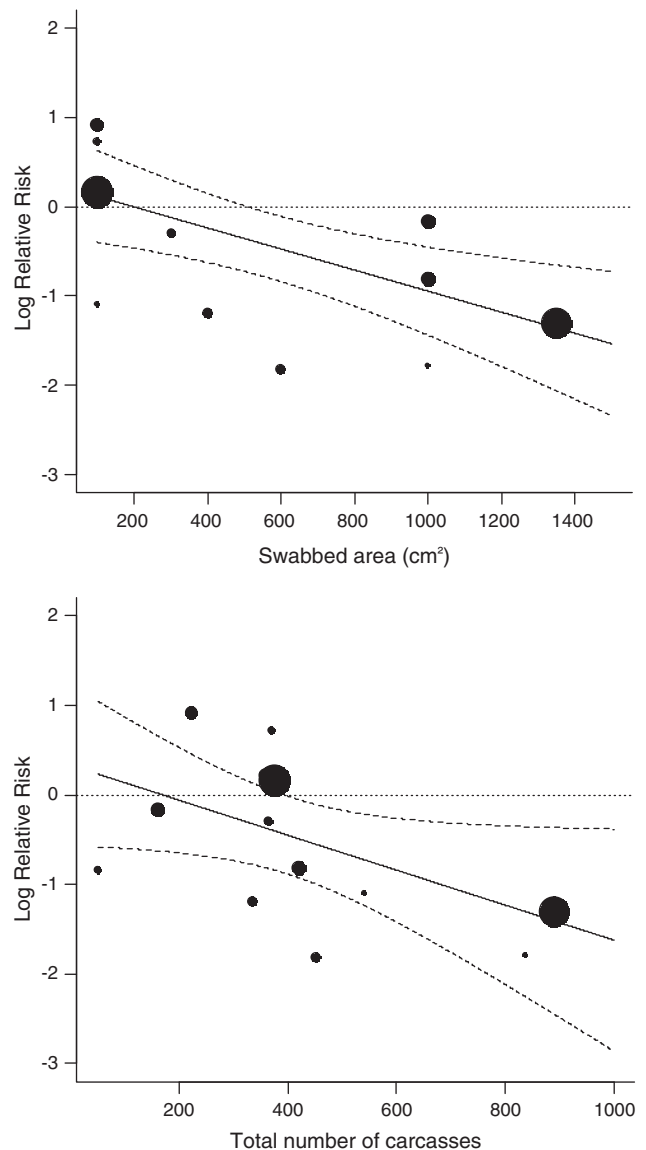

Fig. 2. Effect of the pig carcass swabbed area and the total sample size from the studies' experimental designs on the estimation of the risk of *Salmonella* occurrence on pig carcasses after chilling relative to before chilling, as modelled by separate multilevel meta-analyses at study level. Marker size reflects the weight assigned to the primary studies.

corresponding 95% confidence interval bounds. These plots illustrate how as swabbed area and total sample size increases in the experimental design of a primary study, the observed effect size tends to be lower in terms of log relative risk (i.e., higher observed reduction ratio of *Salmonella* occurrence due to chilling).

An important problem in meta-analysis is the so-called 'file drawer' problem or publication bias. The data for a meta-analysis are the results from previously published studies. Studies that find significant results may have a larger probability to be published. As a result, a sample of published studies can be biassed in the direction of reporting large effects. An approach to test for publication bias is to investigate the effect of the study size directly by including the total sample size of a study as explanatory variable in a multilevel meta-analysis. This allows a formal statistical test, and other study characteristics can be controlled simply by adding these to the explanatory variables (Hox and De Leeuw, 2003). The results of the multilevel model with total sample size (Table 5) and its corresponding plot (Fig. 2, bottom) indicated that differences in *study size* (i.e., sample size) pose threats to our substantive conclusion about the effect of chilling on *Salmonella* on pig carcasses. Because

there is a relation between the sample size and the reported effect size outcome, the existence of a file drawer problem is likely. A general discussion of the influence of the study size on the measured chilling effect is presented in the abattoir-level meta-analyses section. Likewise, another approach for testing the presence of publication bias is shown in the next section by means of a funnel plot.

3.2. Abattoir-level meta-analyses

The abattoir-level meta-analyses also highlighted the significant heterogeneity in the outcomes of, in this case, the surveyed abattoirs. Interestingly, the variability in the chilling effect size $\tau^2 = 0.431$ between abattoirs (Table 6) is numerically close to the variability between studies ($\tau^2 = 0.578$), allowing for the fact that three primary studies had to be removed for the abattoir-level meta-analyses due to format incompatibility. This implies that the variability in effect size among primary studies stems largely from the variability among sampled abattoirs. For this reason, models were fitted with the outcome of an abattoir as the entry unit to be meta-analysed (Table 6). The discrepancy among abattoir outcomes displayed by the forest plot (Fig. 3), as well as the goodness-of-fit measures, and I^2 and H^2 values (Table 6), were all evidence of the presence of between-abattoir heterogeneity; and hence the random-effects model was selected. This model allows us to make the generalisation that, for a population of abattoirs, the reduction ratio of *Salmonella* occurrence due to chilling is ~1.8 (95% CI: 1.2–2.6). This estimate is within the range of that estimated by the study-level random-effects meta-analysis (~1.6; 95% CI: 1.0–2.6 in Table 5).

When moderating variables were included to explain the differences in the abattoir outcomes, some were found significant only when introduced alone in the multilevel models. In this case, the 'total sample size' (N) was not a significant moderator (results not shown) whilst the 'number of sampled batches' (Nba) was able to elucidate part of the heterogeneity found between abattoirs. The results of the multilevel model for Nba (Table 6) showed that the estimate of the effect size of chilling is dependent on the number of times an abattoir is visited (bear in mind that this is a variable measuring *study size* and hence also related to the total sample size). This is supported by the significant coefficient β_1 and the significant QM test for the moderating variable. The estimated amount of residual heterogeneity is $\tau^2 = 0.331$, suggesting that 23.2% ((0.431–0.331)/0.431) of the total amount of heterogeneity between abattoirs can be accounted for by including the 'number of sampled batches' in the model. In agreement with the study-level multilevel meta-analyses results, it was also found at abattoir-level that the swabbed area (A) had the highest predictability for the observed outcomes (Table 6). The carcass swabbed area moderator explained a much higher proportion ((0.431–0.165)/0.431 = 61.7%) of the between-batch heterogeneity in effect sizes. The influence of the swabbed area and the number of sampled batches on the observed chilling effect size can be visually assessed in Fig. 4. The QE statistics for both multilevel models indicated that residual heterogeneities were no longer significant (Table 6). Nonetheless, this should be interpreted with some caution. As Hox and De Leeuw (2003) argue, it is highly unlikely that the available study-level variables could cover all the artefacts causing variation between abattoirs. Generally, the amount of detail in the inputs for meta-analysis, which are the research reports and articles, is not enough to cover all of the studies characteristics. Therefore, to some extent, heterogeneous results are to be expected. It is sensible to conclude instead that other study characteristics that were not coded in the present meta-analyses are likely to be also noteworthy.

In the multilevel model for *Nba*, the value of the coefficient β_1 indicates that an increase in one batch sampled in an abattoir corresponds to a reduction of 0.589 units in terms of the average log relative risk (Table 6). Fig. 4 (bottom) illustrates this relationship, showing that when an abattoir is sampled only once, the observation of an effect (either increasing or decreasing) is imprecise, and therefore, on average no effect would be observed (notice that the mean Log RR is zero when only one batch is sampled). Seemingly, the decreasing effect of chilling on *Salmonella* occurrence became apparent and significant (mean reduction ratio of ~1.8, 95% CI: 1.2–2.6) when at least two batches were sampled in an abattoir. This is in agreement to the previous finding
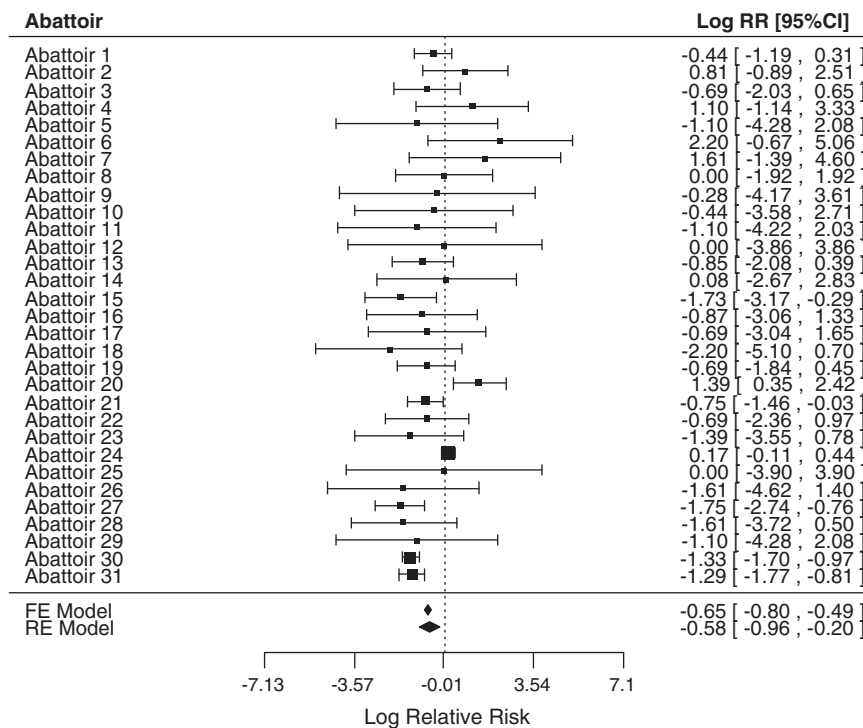
Abattoir	Log RR [95%CI]
Abattoir 1	-0.44 [-1.19 , 0.31]
Abattoir 2	0.81 [-0.89 , 2.51]
Abattoir 3	-0.69 [-2.03 , 0.65]
Abattoir 4	1.10 [-1.14 , 3.33]
Abattoir 5	-1.10 [-4.28 , 2.08]
Abattoir 6	2.20 [-0.67 , 5.06]
Abattoir 7	1.61 [-1.39 , 4.60]
Abattoir 8	0.00 [-1.92 , 1.92]
Abattoir 9	-0.28 [-4.17 , 3.61]
Abattoir 10	-0.44 [-3.58 , 2.71]
Abattoir 11	-1.10 [-4.22 , 2.03]
Abattoir 12	0.00 [-3.86 , 3.86]
Abattoir 13	-0.85 [-2.08 , 0.39]
Abattoir 14	0.08 [-2.67 , 2.83]
Abattoir 15	-1.73 [-3.17 , -0.29]
Abattoir 16	-0.87 [-3.06 , 1.33]
Abattoir 17	-0.69 [-3.04 , 1.65]
Abattoir 18	-2.20 [-5.10 , 0.70]
Abattoir 19	-0.69 [-1.84 , 0.45]
Abattoir 20	1.39 [0.35 , 2.42]
Abattoir 21	-0.75 [-1.46 , -0.03]
Abattoir 22	-0.69 [-2.36 , 0.97]
Abattoir 23	-1.39 [-3.55 , 0.78]
Abattoir 24	0.17 [-0.11 , 0.44]
Abattoir 25	0.00 [-3.90 , 3.90]
Abattoir 26	-1.61 [-4.62 , 1.40]
Abattoir 27	-1.75 [-2.74 , -0.76]
Abattoir 28	-1.61 [-3.72 , 0.50]
Abattoir 29	-1.10 [-4.28 , 2.08]
Abattoir 30	-1.33 [-1.70 , -0.97]
Abattoir 31	-1.29 [-1.77 , -0.81]
FE Model	-0.65 [-0.80 , -0.49]
RE Model	-0.58 [-0.96 , -0.20]

-7.13   -3.57   -0.01   3.54   7.1

Log Relative Risk

Fig. 3. Forest plot of the risk of *Salmonella* occurrence on pig carcasses after chilling relative to before chilling among surveyed abattoirs. Size of square markers reflects the weight assigned to the primary studies. Individual estimates and overall fixed and random effects are shown with 95% confidence intervals.
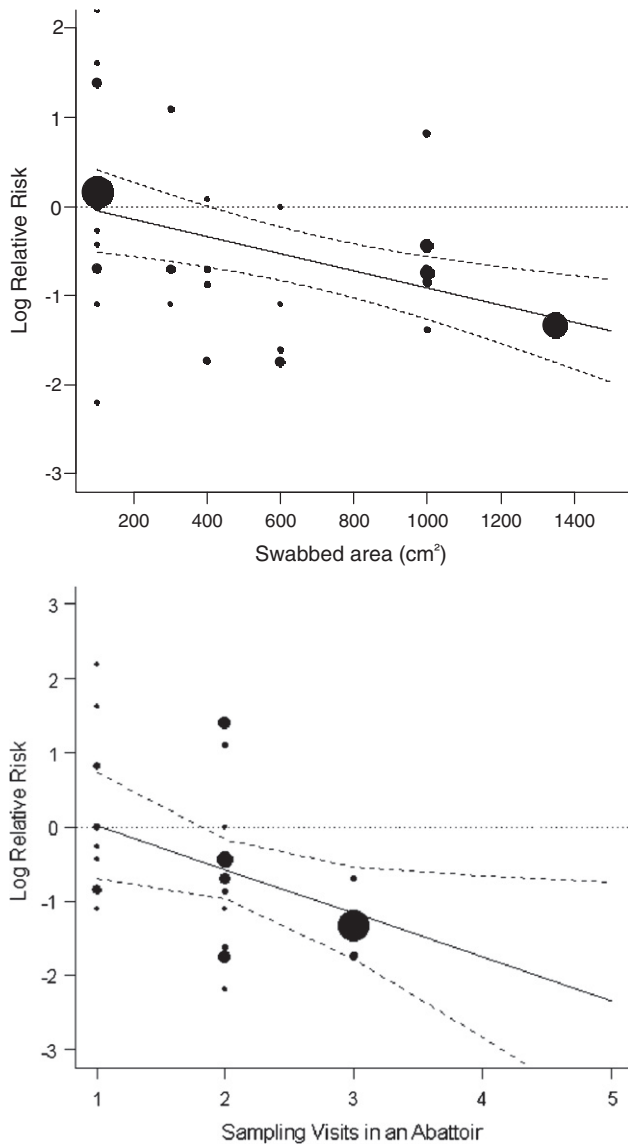
Fig. 4. Effect of the carcass swabbed area (top) and the number of batches sampled in an abattoir (bottom) on the estimation of the risk of *Salmonella* occurrence on pig carcasses after chilling relative to before chilling, as modelled by separate multilevel meta-analyses at abattoir level. Marker size reflects the weight assigned to the individual entries.

for the study-level meta-analysis that the total sample size has a significant influence on the observed beneficial chilling effect. Results of these multilevel models are related and analogous as both moderating variables ('total sample size' and 'number of sampled batches') are measures of *study size*. The comparable meta-analysis results of both the study-level and abattoir-level may be explained by the fact that, although *Salmonella* viability has been proven, at least at laboratory level, to be affected by both temperature (cold shock and refrigerated storage) and water activity (osmotic shock); still the efficacy of the chilling operation for the reduction of *Salmonella* is also affected by other equally important factors related to chilling systems, for example, abattoir logistics, cross-contamination, abattoir hygiene, etc. In addition, *Salmonella* cells are not homogeneously distributed on carcasses, which will greatly add to the *uncertainty* in the measured outcomes (this is, although a pre-chill carcass may contain *Salmonella* cells, swabbing a *Salmonella*-free area will lead to a negative result). On the other hand, the fact that pre-chill and post-chill measurements were mostly performed on different carcasses adds extra randomness to the measured outcome. Thus, it is then expected that, with so many factors

affecting the performance (and the measurement itself of the performance) of the chilling operation, the study size will have a strong influence on the measured effect size. Thus, at abattoir level, results from only one batch made up of an average of 30 pig carcasses, will not have the sufficient statistical power to precisely elucidate any effect (Fig. 4, bottom). In contrast, a greater *Salmonella* incidence reduction ratio (~3.2; 95% CI: 1.7–5.0) was noticed when three batches of an abattoir was sampled. This meta-analysis model however is not supposed to be extrapolated for four or more batches, as the number of sampling occasions is not a continuous variable but a categorical one.

A different approach to investigate publication bias is by drawing a funnel plot which is a plot of the effect sizes versus their standard errors. If the sample of available study outcomes is 'well behaved', this plot should have the shape of a funnel: the outcomes from smaller studies (normally of higher standard errors) are more variable but estimate the same underlying population parameter. If large effects are found predominantly in smaller studies, this indicates the possibility of publication bias, and the possibility of many other insignificant small studies remaining unpublished (Hox and De Leeuw, 2003). The side lines delineating a triangular area facilitate the visual assessment of blank areas within the funnel suggesting publication bias. Fig. 5 (top) shows a 'non-so-well-behaved' funnel plot based on the observed abattoir outcomes. However, as part of the variability in the plot could be due to explanatory study characteristics, it is more appropriate to use a funnel plot after removing the covariate effects; in this case, the effect of the most explanatory variable of carcass swabbed area (Fig. 5, bottom). Although there was an improvement in the funnel plot of the abattoir effect size, the funnel plot without covariate effects still suggests the likely existence of a file drawer problem for less-precise studies. As the file drawer problem is likely for publications reporting both increase and decrease of *Salmonella* occurrence during chilling (i.e., notice the left and right gaps at the bottom of the funnel plot), it may not significantly bias the overall effect size estimate. This inference on the file drawer problem is supported also by the other approach to test publication bias (multilevel model with total sample size as covariate) that was previously presented in the study-level meta-analysis.
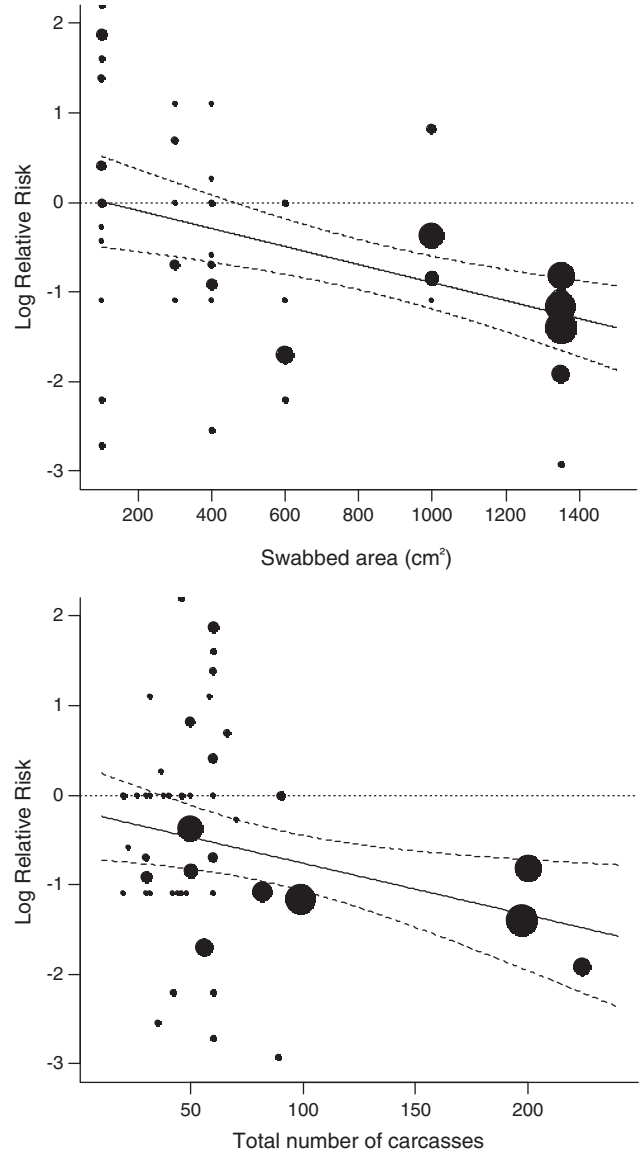
3.3. Batch-level meta-analyses

Since the outcomes of five primary studies were not included in the batch-level models as they were unavailable in such format, the batch-level meta-analyses were not conducted with the intention to produce generalisable results. Instead, their aim was to further explore the effect of the moderating variables on the measured effect size of chilling, as more observations become available at batch level. Although the Q test was significant at $p < 0.10$, the 25% rule of thumb for the intra-class correlation (Hunter and Schmidt, 1990) indicated the presence of heterogeneity in the outcomes measured at the different batches ($I^2 = 37.5\% > 25\%$; Table 7). As with the study-level multilevel meta-analyses, both the carcass swabbed area and the total sample size were study characteristics that significantly explained the effect size heterogeneity among batches. In agreement with the study-level meta-analyses, the swabbed area accounted for more variability in between-batch heterogeneity ((0.373–0.145)/0.373 = 61.1%) than the total sample size ((0.373–0.230)/0.373 = 38.3%).

The batch-level multilevel meta-analyses allowed a better interpretation of the impact of the carcass swabbed area and the total sample size taken from a batch on the measured chilling effect. Fig. 6 clearly shows that when both swabbed area and total sample size were small, the effect size values from the primary studies were more scattered and even conflicting, meaning that the measures were highly imprecise. Notice that when a carcass area of 100 cm² was swabbed, the chilling effect size in log relative risk as estimated from the different surveys, varied from −2.8 to 2.2. Likewise, when the total sample size taken in a batch was equal or lower than 60
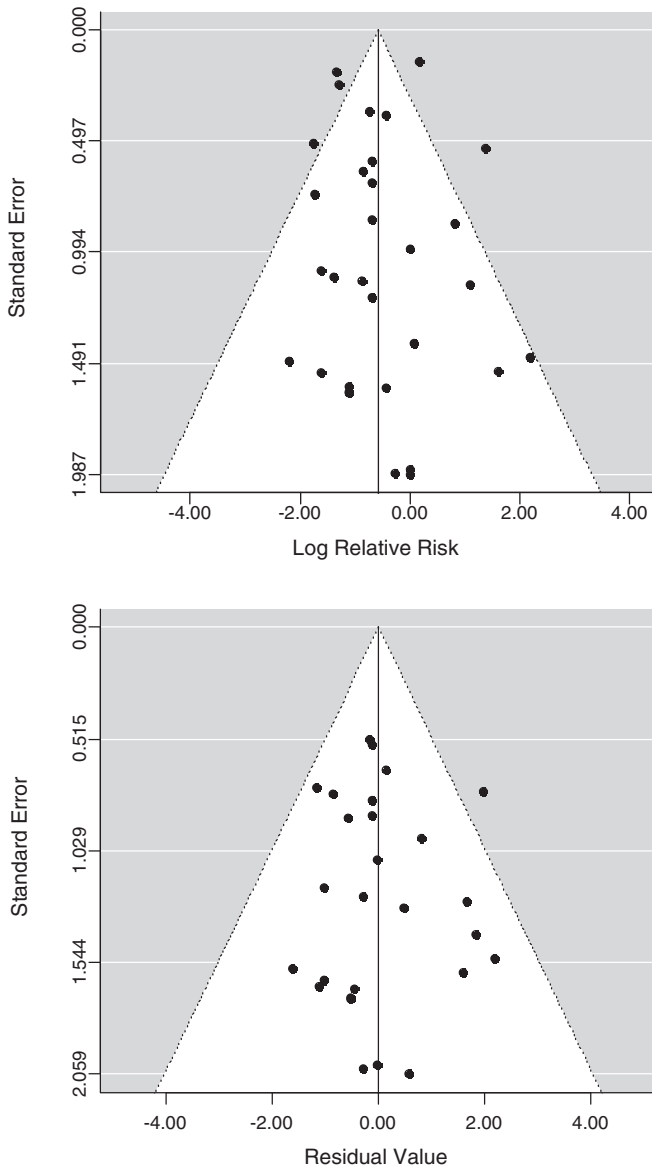
Fig. 5. Funnel plots investigating publication bias based on the observed outcomes (top) and after removal of the carcass swab effect (bottom) of the risk of *Salmonella* occurrence on pig carcasses after chilling relative to before chilling.

Fig. 6. Effect of the carcass swabbed area (top) and the total sample size (bottom) on the estimation of the risk of *Salmonella* occurrence on pig carcasses after chilling relative to before chilling, as modelled by separate multilevel meta-analyses at batch level. Marker size reflects the weight assigned to the individual entries.

pig carcasses, the chilling effect sizes measured by the studies were very disperse ranging between −2.8 and 2.2. Thus, it is not coincidence that as the swabbed area and total sample size increased, the weights assigned to the outcomes became larger (Fig. 6). Comparing the plots of effect size versus swabbed area for the study-level meta-analysis (Fig. 2, top) and for the abattoir-level meta-analysis (Fig. 4, top) with the one for the batch-level meta-analysis (Fig. 6, top), it can be noticed that they are all in agreement, despite the removal of some primary studies, and that carcass swabbed areas higher than 500 cm² should lead to more precise estimates of the decreasing effect of chilling on *Salmonella* occurrence. The higher the swab extent, the higher the measured reduction rate of chilling. The plot for total sample size at study level (Fig. 2, bottom) suggests that a study design of total number of pig carcasses higher than 400 would lead to more precise estimates of the chilling effect, whilst the respective plot for sample size in a batch (Fig. 6, bottom) suggests an approximate of a total of 50 pig carcasses sampled on a visit. At abattoir-level, a similar analysis (Fig. 4, bottom) indicates that at least two sampling visits or sampled batches are required when inferences are to be obtained from one abattoir. Therefore,

from this analysis, it can be deduced that, if we were to conduct a survey study in one abattoir only, a well-designed experiment that has the *minimum* statistical power or resolution to produce consistent and reliable results evidencing the chilling effect, would consist of visiting the abattoir at least twice, sampling a total of 50 pig carcasses (25 pre-chill carcasses and 25 post-chill) per visit, with swabs of at least 500 cm². If a large-scale survey were to be conducted with many abattoirs, the same conditions would apply per abattoir investigated, yet the study-level meta-analysis (Fig. 2, bottom) suggests that the total number of sampled carcasses should be not less than 400. This work has shown that meta-analysis is not only a tool to summarise the findings of a vast amount of primary studies investigating the same hypothesis, but it also provides valuable insight to understand the discrepancies among them, and to sustain the basis for the design of future experiments of higher statistical power.

Whilst meta-analysis is not without limitations (underpowered primary research and publication bias), its systematic approach merits consideration among food safety researchers to integrate the increasingly growing body of knowledge and data on targeted issues along

the complex continuum of agri-food production, and to furnish increased credibility to the findings. A clearer picture of the state of knowledge, evidenced by meta-analysis, can offer valuable and concise information to policy-makers on the effectiveness of treatments and interventions to control and prevent food-borne illnesses in humans. Meta-analysis can highlight areas where there is insufficient evidence of the efficacy of interventions, where there is absence of high quality studies or where there are common methodological flaws in the available research, and – as in this case – can therefore provide direction for future research.

4. Conclusion

Both fixed-effects and random-effects meta-analyses models, carried out at study, abattoir and batch levels, confirmed the significant effect of chilling to reduce the occurrence of Salmonella on pig carcasses. Given the significant heterogeneity in the measured effect size among primary studies and abattoirs, random-effects models were preferred. Whilst, based on the outcomes of 13 primary studies, the overall Salmonella incidence reduction ratio due to chilling can be said to be ~1.6 (95% CI: 1.0–2.6), the one based on the outcomes of 31 pig abattoirs was numerically close at ~1.8 (95% CI: 1.2–2.6). The heterogeneity in the true log relative risks among primary studies ($\tau^2 = 0.578$), among surveyed abattoirs ($\tau^2 = 0.431$), and among sampled batches ($\tau^2 = 0.373$) was further investigated by multilevel meta-analysis models including some study characteristics as covariates. These models revealed that study size characteristics (such as the 'total sample size' and the 'number of batches sampled in an abattoir') and the 'carcass swabbed area' significantly affect the measured effect size of chilling. A higher effect was consistently observed by the swabbed area, which explained between 56 and 62% of the total variability, in contrast to the total sample size, which explained between 23 and 38% of the total heterogeneity in the chilling true effect size. This finding represents a warning that the total sample size and swabbed area, as defined in an experimental design, pose major threats to our substantive conclusion about the effect of chilling on Salmonella incidence. Small-size studies and small swabbed areas lead to imprecise and even conflicting conclusions of increase or decrease of Salmonella incidence due to chilling. This inconsistency to elucidate the effect of chilling is however expected because of the many factors influencing both the performance of the chilling operation as a critical stage (i.e., variability in chilling systems, abattoir logistics, proximity between carcasses, cross-contamination, overall level of hygiene in abattoirs, etc.) and the measurement itself (i.e., randomness due to the heterogeneous distribution of bacterial cells on carcasses; uncertainties associated with the different carcasses sampled before and after chilling; microbial test protocol; etc.). Multilevel meta-analysis was also instrumental in the definition of what should be a well-designed study that has the *minimum* statistical power to produce precise results of the decreasing effect of chilling on Salmonella occurrence on pig carcasses. A more precise estimate of the beneficial effect of chilling that could surmount the different sources of variability and uncertainty affecting its measurement can be obtained with an experimental design of at least 50 total pre-chill and post-chill carcasses sampled from a single batch, with a minimum of two batches sampled per surveyed abattoir, with swab areas of not less than 500 cm^2. Meta-analysis applications such as the one conducted in this study are of importance for the compilation of primary studies, to produce succinct information that could be used in risk assessment modelling; for the understanding of differing outcomes found by primary studies; and for the design of future statistically-sound experiments.

Acknowledgements

References

Algino, R.J., Badtram, G.A., Ingham, B.H., Ingham, S.C., 2009. Factors associated with Salmonella prevalence on pork carcasses in very small abattoirs in Wisconsin. Journal of Food Protection 72 (4), 714–721.

Arguello, H., Carvajal, A., Collazos, J.A., Garcia-Feliz, C., Rubio, P., 2012. Prevalence and serovars of Salmonella enterica on pig carcasses, slaughtered pigs and the environment of four Spanish slaughterhouses. Food Research International 45 (2), 905–912.

Botteldoorn, N., Heyndrickx, M., Rijpens, N., Grijspeerdt, K., Herman, L., 2003. Salmonella on pig carcasses: positive pigs and cross contamination in the slaughterhouse. Journal of Applied Microbiology 95, 891–903.

Bouvet, J., Bavai, C., Rossel, R., Le Roux, A., Montet, M.P., Mazuy, C., Vernozy-Rozand, C., 2003. Evolution of pig carcass and slaughterhouse environment contamination by Salmonella. Reviews Veterinary Medicine 154 (12), 775–779.

Bryk, A.S., Raudenbush, S.W., 1992. Hierarchical Linear Models: Applications and Data Analysis Methods. Sage, Newbury Park, CA.

Cochran, W.G., 1954. The combination of estimates from different experiments. Biometrics 10, 101–129.

Cutter, C.N., 2003. Effects of commercial chilling methods for reducing bacteria on pork carcasses. Research Report. National Pork Board, Des Moines, Iowa (Available at: http://www.pork.org/ResearchDetail/146/Effectsofcommercialc.aspx).

Davies, R.H., McLaren, I.M., Bedford, S., 1999. Distribution of Salmonella contamination in two pig abattoirs. Proceedings of the 3rd International Symposium on the Epidemiology and Control of Salmonella in Pork. Iowa State University, Washington DC, pp. 267–272.

De Busser, E.V., Maes, D., Houf, K., Dewulf, J., Imberechts, H., Bertrand, S., De Zutter, L., 2011. Detection and characterization of Salmonella in lairage, on pig carcasses and intestines in five slaughterhouses. International Journal of Food Microbiology 145, 279–286.

Den Besten, H.M.W., Zwietering, M.H., 2012. Meta-analysis for quantitative microbiological risk assessments and benchmarking data. Trends in Food Science and Technology 25, 34–49.

DerSimonian, R., Laird, N., 1986. Meta-analysis in clinical trials. Controlled Clinical Trials 7, 177–188.

Duggan, S.J., Mannion, C., Prendergast, D.M., Leonard, N., Fanning, S., Gonzales-Barron, U., Egan, J., Butler, F., Duffy, G., 2010. Tracking the Salmonella status of pigs and pork from lairage through the slaughter process in the Republic of Ireland. Journal of Food Protection 73 (12), 2148–2160.

Epling, L.K., Carpenter, J.A., Blakenship, L.C., 1993. Prevalence of Campylobacter spp. and Salmonella spp. on pork carcasses and the reduction affected by spraying with lactic acid. Journal of Food Protection 56 (6), 536–537.

Glass, G.V., 1976. Primary, secondary and meta-analysis of research. Educational Researcher 5 (1,3), 3–8.

Gonzales-Barron, U., Butler, F., 2011. The use of meta-analytical tools in risk assessment for food safety. Food Microbiology 28, 823–827.

Gonzales-Barron, U., Bergin, D., Butler, F., 2008. A meta-analysis study of the effect of chilling on Salmonella prevalence on pork carcasses. Journal of Food Protection 71 (7), 1330–1337.

Gonzales-Barron, U., Soumpasis, I., Butler, F., Duggan, S., Prendergast, D., Duffy, G., 2009. Estimation of prevalence of Salmonella spp. on pig carcasses and pork joints using a quantitative risk assessment model aided by meta-analysis. Journal of Food Protection 72 (2), 274–285.

Grieg, J.D., Waddell, L., Wilhelm, B., Wilkins, W., Bucher, O., Parker, S., Rajic, A., 2012. The efficacy of interventions applied during primary processing on contamination of beef carcasses with Escherichia coli. Food Control 27 (2), 385–397.

Hedges, L.V., Vevea, J.L., 1998. Fixed- and random-effects models in meta-analysis. Psychological Methods 3, 486–504.

Higgins, J.P.T., Thompson, S.G., 2002. Quantifying heterogeneity in a meta-analysis. Statistics in Medicine 21, 1539–1558.

Hox, J.J., De Leeuw, E., 2003. Multilevel models for meta-analysis. In: Reise, S.P., Duan, N. (Eds.), Multilevel Modelling: Methodological Advances, Issues and Applications. Lawrence Erlbaum Associates, New Jersey, pp. 90–111.

Hunter, J.E., Schmidt, F.L., 1990. Methods of Meta-Analysis: Correcting Error and Bias in Research Findings. Sage, Newbury Park, CA.

McQuestin, O., Shadbolt, C.T., Ross, T., 2009. Quantification of the relative effects of temperature, pH and water activity on inactivation of Escherichia coli in fermented meat by meta-analysis. Applied and Environmental Microbiology 75, 6963–6972.

Noble, J.H., 2006. Meta-analysis: methods, strengths, weaknesses and political uses. The Journal of Laboratory and Clinical Medicine 147, 7–20.

Oosterom, J., Dekker, R., de Wilde, G.J.A., van Kempen-de Troye, F., Engels, G.B., 1985. Prevalence of Campylobacter jejuni and Salmonella during pig slaughtering. Veterinary Quarterly 7 (1), 31–34.

Patil, S.R., Morales, R., Cates, S., Anderson, D., Kendall, D., 2004. An application of meta-analysis in food safety consumer research to evaluate consumer behaviours and practices. Journal of Food Protection 67 (11), 2587–2595.

Quirke, A.M., Leonard, N., Kelly, G., Egan, J., Lynch, P.B., Rowe, T., Quinn, P.J., 2001. Prevalence of Salmonella serotypes on pig carcasses from high- and low-risk herds slaughtered in three abattoirs. Berliner und Münchener Tierärztliche Wochenschrift 114, 360–362.

Raudenbush, S.W., Bryk, A.S., 1985. Empirical Bayes meta-analysis. Journal of Educational Statistics 10, 75–98.

Saide-Albornoz, J., Knipe, C.L., Murano, E.A., Beran, G.W., 1995. Contamination of pork carcasses during slaughter, fabrication and chilled storage. Journal of Food Protection 58 (9), 993–997.

Sánchez, J., Dohoo, I.R., Christensen, J., Rajic, A., 2007. Factor influencing the prevalence of *Salmonella* spp. in swine farms: a meta-analysis approach. Preventive Veterinary Medicine 81, 148–177.

Sargeant, J.M., Amezcua, M., Rajic, A., Waddell, L., 2005. A Guide to Conducting Systematic Reviews in Agri-food Public Health. Food Safety Research and Response Network, Canada (Available at: http://www.fsrrn.net/UserFiles/File/conductingsysreviewsenglish [1].pdf).

Sargeant, J.M., Rajic, A., Read, S., Ohlsson, A., 2006. The process of systematic review and its application in agri-food public health. Preventive Veterinary Medicine 76, 141–151.

Sutton, A.J., Abrams, K.R., Jones, D.R., 2001. An illustrated guide to the methods of meta-analysis. Journal of Evaluation in Clinical Practice 7 (2), 135–148.

Van den Noortgate, W., Onghena, P., 2003. Multilevel meta-analysis: a comparison with traditional meta-analytical procedures. Educational and Psychological Measurement 63 (5), 765–790.

Vialette, M., Pinon, A., Leporq, B., Dervin, C., Membré, J.M., 2005. Meta-analysis of food safety information based on a combination of a relational database and a predictive modelling tool. Risk Analysis 25 (1), 75–83.

Viechtbauer, W., 2010. Conducting meta-analyses in R with the 'Metafor' package. Journal of Statistical Software 36 (3), 1–48.

Whitehead, A., 2002. Meta-analysis of Controlled Clinical Trials. John Wiley & Sons, West Sussex, England.

Wilhelm, B., Rajic, A., Greig, J.D., Waddell, L., Harris, J., 2011. The effect of hazard analysis critical control point programs on microbial contamination in abattoirs: a systematic review of published data. Foodborne Pathogens and Disease 8, 949–960.