International Journal of Food Science +Technology

Institute of Food Science +Technology ifst

Original article

Validation software tool (ValT) for predictive microbiology based on the acceptable prediction zones method

Thomas P. Oscar*

United States Department of Agriculture, Agricultural Research Service, Poultry Food Safety Research Worksite, Room 2111, Center for Food Science and Technology, University of Maryland Eastern Shore, Princess Anne MD 21853, USA

(Received 6 November 2019; Accepted in revised form 15 January 2020)

Abstract A study was undertaken to develop a user-friendly software tool (ValT) for validation of predictive models based on the acceptable prediction zones (APZ) method. ValT was developed in Excel and was demonstrated using a newly developed growth model for *Salmonella* and chicken skin. The model provided acceptable predictions when the proportion (p) of residuals in the APZ was ≥0.70 and there were no local prediction problems. The pAPZ were 0.97 for dependent data (n = 360), 0.93 for interpolation data (n = 160) and 0.88 for extrapolation data (n = 180) to kosher chicken skin. There were no local prediction problems, and all sets of data satisfied the test data criteria of the APZ method. Thus, the model was successfully validated for interpolation and for extrapolation to kosher chicken skin. It is hoped that the new software tool will make it easier to validate models using the APZ method.

Keywords Chicken, food safety, model, neural network, predictive microbiology, *Salmonella*, software tool, validation.

Introduction

Predictive models (de Oliveira Elias *et al*., 2018; Estilo & Gabriel, 2018; Jia *et al*., 2019; Kuroda *et al*., 2019) are valuable tools for food safety because they provide real-time predictions of changes in pathogen number in response to changes in intrinsic and extrinsic factors without the need for expensive and time-consuming microbiological tests (McKellar & Lu, 2004). Predictive models are also valuable because they can make predictions for scenarios that were not investigated but that fall within the ranges of independent variables used to develop them (Bhaduri *et al*., 1994; Baranyi *et al*., 1996). The information provided by predictive models can help support decisions aimed at protecting consumers from microbial hazards associated with food leading to a reduction in food-borne illness (Panisello & Quantick, 1998; Tamplin, 2018).

Proper validation of models is important because it provides users with confidence that predictions are reliable (Ross, 1996). The first step of model validation is to compare model predictions to the data used to develop them (Zwietering *et al*., 1994). This usually involves statistical measures of goodness of fit as well as graphical comparisons of observed and predicted values to look for systematic prediction bias (te Giffel & Zwietering, 1999). The second step of model validation is to compare model predictions to data not used to develop them to assess the ability of the model to interpolate (Oscar, 2005b). The third step of the validation process is to test the ability of the model to extrapolate to independent variables not included in model development to see how broadly the model can be applied (Martino *et al*., 2005).

Validation of models involves calculation of performance metrics for prediction bias and accuracy (Ross, 1996) and use of criteria for model performance (Ross *et al*., 2000). In addition, it is important to have criteria for test data to ensure that comparisons of observed and predicted values are valid and that the evaluation of model performance is accurate, and unbiased. The acceptable prediction zones (APZ) method (Oscar, 2005b) for validation of predictive models has criteria for test data and model performance as well as a metric (proportion of residuals in the APZ or pAPZ) for prediction bias and accuracy (Oscar, 2005b). Although the APZ method is being used in the field of predictive microbiology to evaluate models, its criteria for test data are not being used and its criteria for model performance are not being correctly used because they are complex (Min & Yoon, 2010; Li *et al*., 2011; Ostergaard *et al*., 2014; Mohr *et al*., 2015). Therefore, the current study was undertaken to develop a user-friendly software tool (ValT) for proper validation of models using the APZ method. The main advantages are calculation of pAPZ

*Correspondent: Fax: + 1 410-651-8498;
e-mail: thomas.oscar@usda.gov

doi:10.1111/ijfs.14534

Published 2020. This article is a U.S. Government work and is in the public domain in the USA
The peer review history for this article is available at https://publons.com/publon/10.1111/ijfs.14534.

and provision of decision trees to organise and integrate criteria for test data and model performance into an objective decision about model validation. The decision is objective because it is based on well-defined criteria that are science-based and the same result is obtained regardless of who performs the evaluation. The main drawback of ValT is that it is not completely automated, and thus, false answers can be entered by accident or on purpose. The new software tool (ValT) was demonstrated using a new model (multiple-layer feedforward neural network) for growth of *Salmonella* Typhimurium DT104 on chicken skin. Although it was not demonstrated in this study, the APZ method and by inference ValT can and has been used to evaluate models for other pathogens like *Listeria monocytogenes* (Abou-Zeid *et al*., 2009) and *Campylobacter* spp. (Oyarzabal *et al*., 2010).

Materials and methods

Software tool

The validation software tool (ValT) was developed in Excel (version 2016, MicroSoft Corporation, Redmond, WA) and was based on the test data and model performance criteria of the acceptable prediction zones (APZ) method (Oscar, 2005a; Oscar, 2005b). It consisted of eight spreadsheets: (i) table of contents; (ii) dependent data input; (iii) decision tree for dependent data (Fig. 1); (iv) interpolation data input; (v) decision tree for interpolation data (Fig. 2); (vi) extrapolation data input; (vii) decision tree for extrapolation data (Fig. 3); and (viii) array of APZ values for partially acceptable residuals.

Users enter independent variable, observed and predicted values into spreadsheets 2, 4 and 6. Formula in these spreadsheets then calculate residuals, assign individual APZ values (0–1) and generate a code for counting prediction cases as a function of independent variable levels and combinations. Spreadsheets 3, 5 and 7 contain decision trees with questions for test data and model performance criteria and validation, and pivot tables for test data and model performance. Decision trees integrate criteria for test data and model performance into an objective decision about model validation where answers of 'yes' lead to validation and a single answer of 'no' leads to model failure. Questions about criteria are answered by users, whereas as questions about validation are answered by ValT. The software works on the honor system because it is not fully automated.

Criteria for test data

The criteria for test data depend on the type of data being evaluated (Oscar, 2005b). The criteria for dependent data are (i) used for model development; (ii) even spacing of values for independent variables; (iii) minimum of four prediction cases per combination of independent variables; and (iv) same number of prediction cases per combination of independent variables. The criteria for interpolation data are (i) not used for model development; (ii) obtained with the same methods (same strain(s), same previous growth conditions, same inoculum size, same food matrix and same enumeration method) as dependent data; (iii) values of independent variables that are intermediate to those used in model development; (iv) minimum of two prediction cases per combination of independent variables; and (v) same number of prediction cases per combination of independent variables. The criteria for extrapolation data are (i) not used in model development; (ii) obtained with the same methods as dependent data except for the new independent variable being evaluated; (iii) obtained using the same experimental design as dependent data except for the new independent variable being evaluated; (iv) minimum of two prediction cases per combination of independent variables; and (v) same number of prediction cases per combination of independent variables.

Criteria for model performance

Criteria for model performance are the same for all types of data (Oscar, 2005a). An APZ value (Y) is assigned to each residual (X) for calculation of the proportion (p) of residuals (observed – predicted) in the APZ (pAPZ) as follows:

$$Y = 0 \quad \text{if } X \leq -2$$

$$Y = 1X + 2 \quad \text{if } X > -2 \text{ and} < -1$$

$$Y = 1 \quad \text{if } X \geq -1 \text{ and } \leq 0.5$$

$$Y = -2X + 2 \quad \text{if } X > 0.5 \text{ and } < 1$$

$$Y = 0 \quad \text{if } X \geq 1$$

where pAPZ = $\Sigma Y_i/n$ where i was the ith APZ value (Y), and n was the number of residuals (X).

Special prediction cases occur when the observed or predicted number is zero, for which the logarithm is undefined. To include these cases in the calculation of pAPZ, a value of −0.01 log is entered in ValT for the observed or predicted value that was zero and an APZ value was assigned as follows:

$$Y = 0 \quad \text{if } O = -0.01 \text{ log and } P \geq 0 \text{ log or } O \geq 0 \text{ log and } P = -0.01 \text{ log}$$

$$Y = 1 \quad \text{if } O = -0.01 \text{ log and } P = -0.01 \text{ log}$$

where O was the observed log count, and P was the predicted log count. Any negative value could be used here without affecting the result. The value of −0.01

 Published 2020. This article is a U.S. Government work and is in the public domain in the USA

Q	A	Decision Tree for Dependent Data
1	Yes	Were the data used to develop the model?
2	Yes	Were the values of the independent variables evenly spaced?
3	Yes	Was there a minimum of four prediction cases per combination of independent variables?
4	Yes	Did all combinations of independent variables have the same number of prediction cases?
5	Yes	Was the overall pAPZ ≥ 0.70?
6	Yes	Was pAPZ for all individual levels of independent variables ≥ 0.70?
7	Yes	Was a single pAPZ ≥ 0.70 for every three consecutive combinations of the independent variables?
8	Yes	Was the model validated for dependent data?

Pivot Table for Test Data (Prediction Case Count)

	t					
T	0	2	4	6	8	Total
10	8	8	8	8	8	40
15	8	8	8	8	8	40
20	8	8	8	8	8	40
25	8	8	8	8	8	40
30	8	8	8	8	8	40
35	8	8	8	8	8	40
40	8	8	8	8	8	40
45	8	8	8	8	8	40
50	8	8	8	8	8	40
Total	**72**	**72**	**72**	**72**	**72**	**360**

Pivot Table for Model Performance (pAPZ)

	t					
T	0	2	4	6	8	Total
10	1.00	1.00	0.93	1.00	1.00	0.98
15	1.00	0.90	1.00	1.00	1.00	0.98
20	1.00	1.00	1.00	1.00	1.00	1.00
25	1.00	1.00	1.00	0.85	1.00	0.97
30	1.00	1.00	0.97	0.87	0.89	0.95
35	1.00	1.00	0.93	0.96	0.97	0.97
40	1.00	1.00	0.96	1.00	0.92	0.97
45	1.00	0.97	1.00	0.91	0.80	0.93
50	1.00	0.95	0.89	0.98	0.97	0.96
Total	**1.00**	**0.98**	**0.96**	**0.95**	**0.95**	**0.97**

Figure 1 Screenshot of the validation software tool (ValT) for dependent data. T, temperature; t, time, pAPZ, proportion of residuals in the acceptable prediction zones.

log was used because it is the convention used in ComBase for samples with zero pathogens, for which the logarithm is undefined (Baranyi & Tamplin, 2004). ComBase is an internationally used microbial modelling database.

Finally, a local prediction problem occurred when pAPZ < 0.70 for a single level of an independent variable or when pAPZ < 0.70 for three consecutive combinations of independent variables (Oscar, 2018a).

Model validation

A model is classified as validated in the APZ method when it satisfies all criteria for test data and model performance for dependent data and independent data for interpolation (Oscar, 2005b). Predictions of models that fail validation for dependent data are not reliable. Thus, validation for interpolation requires validation for dependent data. Likewise, predictions of models that fail validation for interpolation are not reliable. Thus, validation for extrapolation requires validation for interpolation. A model that fails validation can be repaired by further data collection, interpolation of missing data, data pruning and(or) use of other models (Oscar, 2005a). Thus, all models can in theory pass this validation process.

Case study

To demonstrate ValT, published log count data (Oscar, 2009b) for growth of *Salmonella* Typhimurium DT104 on chicken skin as a function of time (0–8 h) and temperature (10 °C–50 °C) were used to develop a new model (multiple-layer feedforward neural network with two hidden layers of two nodes each). The new model was developed in Excel using NeuralTools (version 7.6, Palisade Corporation, Ithaca, NY) as previously described (Oscar, 2018a). The ability of the model to predict dependent data, independent data for interpolation and independent data for extrapolation to kosher chicken skin was evaluated using ValT.

Results

Predictive model

The model for growth of *Salmonella* Typhimurium DT104 on chicken skin is shown in Fig. 4. A temperature from 10 °C to 50 °C is entered, and then, the model predicts the growth curve from 0 to 8 h. For example, in Fig. 4, a temperature of 39 °C was entered and then the model predicted that *Salmonella* Typhimurium DT104 would grow on chicken skin from an

Published 2020. This article is a U.S. Government work and is in the public domain in the USA

Q	A	Decision Tree for Interpolation
1	Yes	Was the model validated for dependent data?
2	Yes	Were the data independent?
3	Yes	Were the data collected using the same methods as dependent data?
4	Yes	Were the independent variables at values intermediate to those used in model development?
5	Yes	Was there a minimum of two prediction cases per combination of independent variables?
6	Yes	Did all combinations of independent variables have the same number of prediction cases?
7	Yes	Was the overall pAPZ ≥ 0.70?
8	Yes	Was pAPZ for all individual levels of independent variables ≥ 0.70?
9	Yes	Was a single pAPZ ≥ 0.70 for every three consecutive combinations of the independent variables?
10	Yes	Was the model validated for interpolation?

Pivot Table for Test Data (Prediction Case Count)

	t						
T	0	2	4	6	8	Total	
12.5	4	4	4	4	4	20	
17.5	4	4	4	4	4	20	
22.5	4	4	4	4	4	20	
27.5	4	4	4	4	4	20	
32.5	4	4	4	4	4	20	
37.5	4	4	4	4	4	20	
42.5	4	4	4	4	4	20	
47.5	4	4	4	4	4	20	
Total	**32**	**32**	**32**	**32**	**32**	**160**	

Pivot Table for Model Performance (pAPZ)

	t					
T	0	2	4	6	8	Total
12.5	1.00	1.00	1.00	1.00	1.00	1.00
17.5	1.00	1.00	1.00	1.00	0.99	1.00
22.5	1.00	1.00	1.00	1.00	1.00	1.00
27.5	1.00	1.00	1.00	1.00	0.46	0.89
32.5	1.00	0.93	0.93	0.91	0.76	0.91
37.5	1.00	0.96	0.94	0.95	0.97	0.96
42.5	1.00	0.93	0.96	0.95	0.87	0.94
47.5	1.00	1.00	0.74	0.51	0.53	0.76
Total	**1.00**	**0.98**	**0.95**	**0.91**	**0.82**	**0.93**

Figure 2 Screenshot of the validation software tool (ValT) for interpolation data. T, temperature; t, time, pAPZ, proportion of residuals in the acceptable prediction zones.

initial log count of 0.79 at 0 h to a final log count of 6.40 at 8 h.

Dependent data

The data used in model development and their predicted growth curves are shown in Fig. 5. These data satisfied all criteria for test data as indicated by answers of 'yes' to questions 1–4 in the decision tree for dependent data (Fig. 1). The overall pAPZ for dependent data ($n = 360$) was 0.97 (Fig. 6A), whereas the range of pAPZ was 0.95–1.00 for individual times, 0.93–1.00 for individual temperatures and 0.80–1.00 for individual combinations of time and temperature (Fig. 1). The maximum number of consecutive combinations of time and temperature with pAPZ < 0.70 was zero. Thus, the model satisfied all criteria for model performance as indicated by answers of 'yes' to question 5–7 in the decision tree for dependent data (Fig. 1). Because the data and model satisfied all the test data and model performance criteria, the answer to question 8 in the decision tree was 'yes' indicating that the model was validated for dependent data.

Interpolation data

Because the model was validated for dependent data, the answer to question 1 in the decision tree for interpolation (Fig. 2) was 'yes' indicating that the model could be evaluated for interpolation. The data used to evaluate the model for interpolation and their predicted growth curves are shown in Fig. 7. These data satisfied all criteria for test data as indicated by answers of 'yes' to questions 2–6 in the decision tree for interpolation (Fig. 2). The overall pAPZ for interpolation data ($n = 160$) was 0.93 (Fig. 6B), whereas the range of pAPZ was 0.82–1.00 for individual times, 0.76–1.00 for individual temperatures and 0.51–1.00 for individual combinations of time and temperature (Fig. 2). The maximum number of consecutive time and temperature combinations with pAPZ < 0.70 was two. Thus, the model satisfied all criteria for model performance as indicated by answers of 'yes' to questions 7–9 in the decision tree for interpolation (Fig. 2). Because the data and model satisfied all test data and model performance criteria of the APZ method, the answer to question 10 in the decision tree for interpolation was 'yes' indicating that the model was validated for interpolation.

Published 2020. This article is a U.S. Government work and is in the public domain in the USA

Q	A	Decision Tree for Extrapolation
1	Yes	Was the model validated for interpolation?
2	Yes	Were data independent?
3	Yes	Were data collected with the same methods as dependent data except for the new independent variable being evaluated?
4	Yes	Were the independent variables at the same values as those used in model development?
5	Yes	Was there a minimum of two prediction cases per combination of independent variables?
6	Yes	Did all combinations of the independent variables have the same number of prediction cases?
7	Yes	Was the overall pAPZ ≥ 0.70?
8	Yes	Was pAPZ for all individual levels of independent variables ≥ 0.70?
9	Yes	Was a single pAPZ ≥ 0.70 for every three consecutive combinations of the independent variables?
10	Yes	Was the model validated for extrapolation?

Pivot Table for Test Data (Prediction Case Count)

	t					
T	0	2	4	6	8	Total
10	4	4	4	4	4	20
15	4	4	4	4	4	20
20	4	4	4	4	4	20
25	4	4	4	4	4	20
30	4	4	4	4	4	20
35	4	4	4	4	4	20
40	4	4	4	4	4	20
45	4	4	4	4	4	20
50	4	4	4	4	4	20
Total	**36**	**36**	**36**	**36**	**36**	**180**

Pivot Table for Model Performance (pAPZ)

	t					
T	0	2	4	6	8	Total
10	1.00	0.99	0.90	0.83	1.00	0.94
15	1.00	1.00	1.00	1.00	1.00	1.00
20	1.00	1.00	1.00	1.00	1.00	1.00
25	1.00	1.00	1.00	0.91	0.88	0.96
30	1.00	1.00	0.88	0.64	0.23	0.75
35	1.00	0.75	0.79	0.86	0.89	0.86
40	1.00	0.83	0.92	0.98	1.00	0.94
45	1.00	0.75	0.83	0.44	0.47	0.70
50	1.00	1.00	0.50	0.50	0.74	0.75
Total	**1.00**	**0.92**	**0.87**	**0.79**	**0.80**	**0.88**

Figure 3 Screenshot of the validation software tool (ValT) for extrapolation data. T, temperature; t, time, pAPZ, proportion of residuals in the acceptable prediction zones.

Extrapolation data

Because the model satisfied all test data and model performance criteria for interpolation, the answer to question 1 in the decision tree for extrapolation (Fig. 3) was 'yes' indicating that the model could be evaluated for extrapolation. The data used to evaluate the model for extrapolation to kosher chicken skin and predicted growth curves are shown in Fig. 8. These data satisfied all criteria for test data as indicated by answers of 'yes' to questions 2–6 in the decision tree for extrapolation (Fig. 3). The overall pAPZ for the extrapolation data (n = 180) was 0.88 (Fig. 6c), whereas the range of pAPZ was 0.79–1.00 for individual times, 0.70–1.00 for individual temperatures and 0.23–1.00 for individual combinations of time and temperature. The maximum number of consecutive time and temperature combinations with pAPZ < 0.70 was two. Thus, the model satisfied all criteria for model performance as indicated by answers of 'yes' to questions 7–9 in the decision tree for extrapolation (Fig. 3). Because the data and model satisfied all criteria for test data and model performance, the answer to question 10 in the decision tree for extrapolation was 'yes' indicating that the model was validated for extrapolation to kosher chicken skin.

Discussion

Although there is general agreement that validation of models in predictive microbiology is important, there is no consensus as to what this means. A goal of the APZ method and ValT is to bring clarity to the term 'validated model' through development and integration of criteria for test data and model performance that lead to an accurate, unbiased and objective decision about model validation. Without stated criteria that are well-defined and science-based, the decision about model validation would be subjective (authors' opinion). The criteria in the APZ method are well-defined and based on statistical principles and analysis (Oscar, 2005a; Oscar, 2005b). They have been vetted by the scientific peer-review process many times (Oscar, 2005a; Oscar, 2005b; Oscar, 2006; Oscar, 2007; Abou-Zeid *et al*., 2009; Oscar, 2009a; Oscar, 2009b; Oyarzabal *et al*., 2010; Oscar, 2011a; Oscar, 2011b; Oscar,

Published 2020. This article is a U.S. Government work and is in the public domain in the USA

Temp	Time	Log
39	0	0.79
39	1	0.92
39	2	1.15
39	3	1.55
39	4	2.23
39	5	3.28
39	6	4.56
39	7	5.68
39	8	6.40

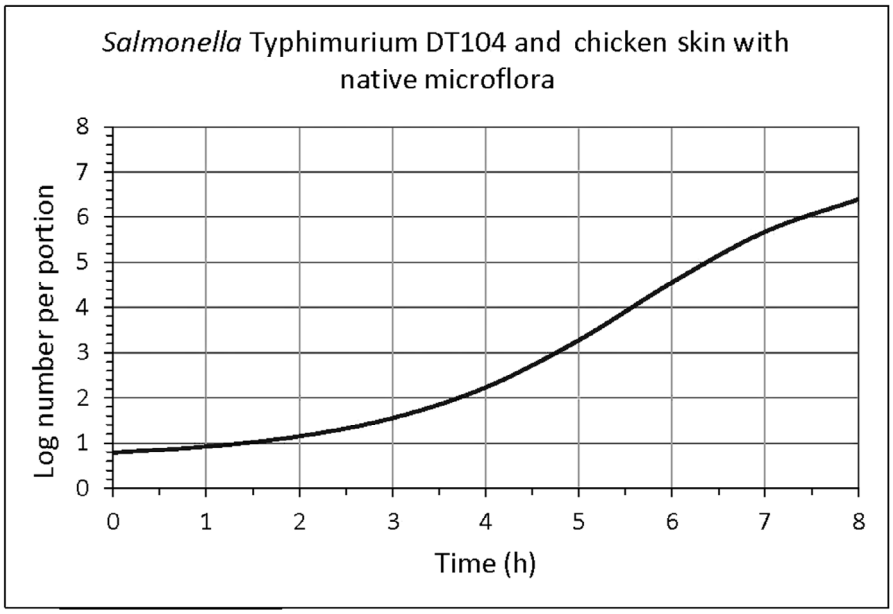

Figure 4 Screenshot of the predictive model for growth of *Salmonella* Typhimurium DT104 on chicken skin as a function of time (0–8 h) and temperature (10 °C–50 °C)

2013; Oscar, 2014; Oscar, 2015; Oscar, 2017a; Oscar, 2017b; Oscar, 2018a; Oscar, 2018b). Thus, they provide a proper evaluation and validation of predictive models for food-borne pathogens that is objective because the result is the same regardless of who performs the evaluation using these criteria. Proper validation of models is important because it provides users with confidence that predictions are reliable. This is especially important when models are used to make important food safety decisions that impact public health.

The APZ method was first developed and applied to secondary models that predict lag time and growth rate as a function of independent variables (Oscar, 2005b). The prediction error used in the APZ method for lag time and growth rate is the relative error. Subsequently, the APZ method was expanded to include models that predict log counts as a function of time, temperature and other independent variables using the residual as the prediction error (Oscar, 2005a). The test data criteria in the APZ method are similar for all types of models, whereas the APZ are specific for each type of model. When the APZ method is applied to models that predict log counts as a function of time, it is important to realise that the models are being evaluated for how well they predict log counts in all phases of the pathogen life cycle.

A model is classified as validated in the APZ method when it satisfies all criteria for test data and model performance for dependent data and independent data for interpolation (Oscar, 2005b). Validation for extrapolation is optional but important because it saves time and money by identifying conditions for which new models are not needed (Oscar, 2011b; Oscar, 2013; Oscar, 2015; Oscar, 2018a). In the present study, a new model (multiple-layer feedforward neural network) for growth of a low initial number (0.85 log) of *Salmonella* Typhimurium DT104 on chicken skin as a function of time (0–8 h) and temperature (10 °C 50 °C) was successfully validated for dependent data, for interpolation and for extrapolation to kosher chicken skin. Thus, the new model can be used with confidence to make predictions within the ranges of time and temperature used to develop it and a new model for kosher chicken skin is not needed. How well the model predicts growth of other inoculum sizes and serotypes remains to be determined. The current model passed validation indicating that the criteria were not too rigorous. In theory, all models should be able to pass this validation process.

The criteria for test data in the APZ method are important because they ensure that comparisons of observed and predicted values are valid (not confounded) and that the evaluation of model performance is accurate and unbiased. To accomplish this, it is important to collect validation data using the same methods as used to collect data for model development so that valid (not confounded) comparisons can be made between observed and predicted values (Oscar, 2005b). It is also important that observed and

 Published 2020. This article is a U.S. Government work and is in the public domain in the USA

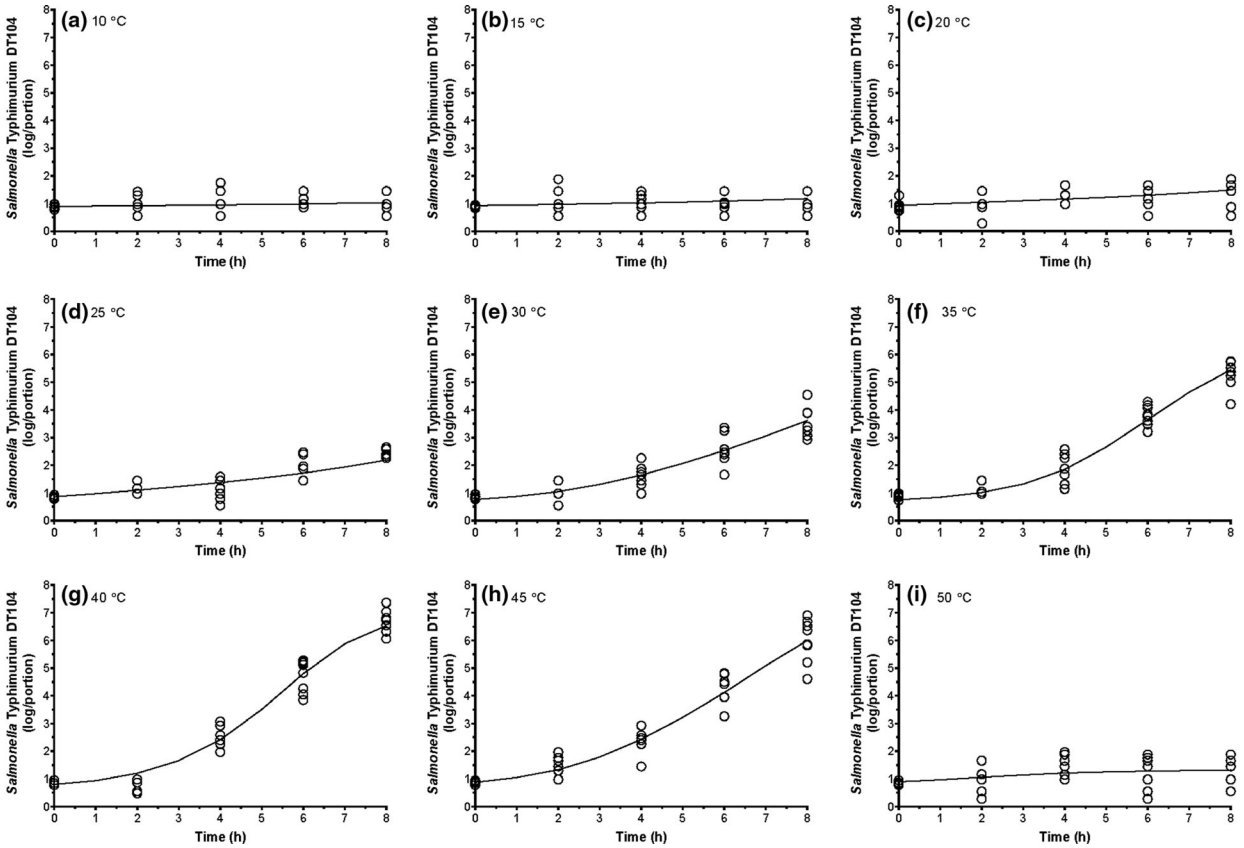

Figure 5 Growth curves for *Salmonella* Typhimurium DT104 and chicken skin incubated at (a) 10 °C; (b) 15 °C; (c) 20 °C; (d) 25 °C; (e) 30 °C; (f) 35 °C; (g) 40 °C; (h) 45 °C; or (i) 50 °C. Open symbols are observed log counts, and solid lines are predicted log counts for data used in model development.

predicted values be evenly distributed throughout the prediction region of the model so that the evaluation of model performance is not biased (Oscar, 2018a). Likewise, it is important that each combination of independent variables be sufficiently replicated so that the evaluation of model performance is accurate. Enough replication in the APZ method is four observed values for dependent data and two observed values for independent data. Too many or too few data (uneven distribution) in one or more parts of the prediction region of the model can bias the evaluation of model performance, whereas too little data can result in an inaccurate evaluation of model performance and a false conclusion about model validation.

The first validation method in the field of predictive microbiology to have criteria for model performance was the bias factor (B_f) and accuracy factor (A_f) method (Ross, 1996). The B_f is the antilog of the mean log ratio of predicted and observed values, whereas A_f is the antilog of the mean absolute log ratio of predicted and observed values. A B_f of 1.0 indicates no average prediction bias, whereas an A_f of 1.0 indicates perfect agreement between predicted and observed values. In this method, a model provides predictions with acceptable bias when B_f is between 0.7 (fail – safe) and 1.15 (fail – dangerous) and a model provides predictions with acceptable accuracy when A_f is $<1.0 + 0.15x$, where x is the number of independent variables in the model (Ross *et al.*, 2000). These metrics and criteria are the most widely used in the field of predictive microbiology for model validation; yet, they were not used in ValT for the following reasons.

First, B_f and A_f are average values. Thus, they can be affected by outliers (Delignette-Muller *et al.*, 1995) resulting in a biased and inaccurate evaluation of model performance and a false conclusion about model validation. This is especially true for small sets of data. In contrast, pAPZ is a proportion that is not affected by outliers because each prediction case contributes equally to its final value (Oscar, 2005b).

Second, when least squares regression methods are used to fit models to data, B_f is always 1.0 for dependent data (Baranyi *et al.*, 1999; Oscar, 2005b). Thus, B_f is not an informative metric for assessing goodness

Published 2020. This article is a U.S. Government work and is in the public domain in the USA

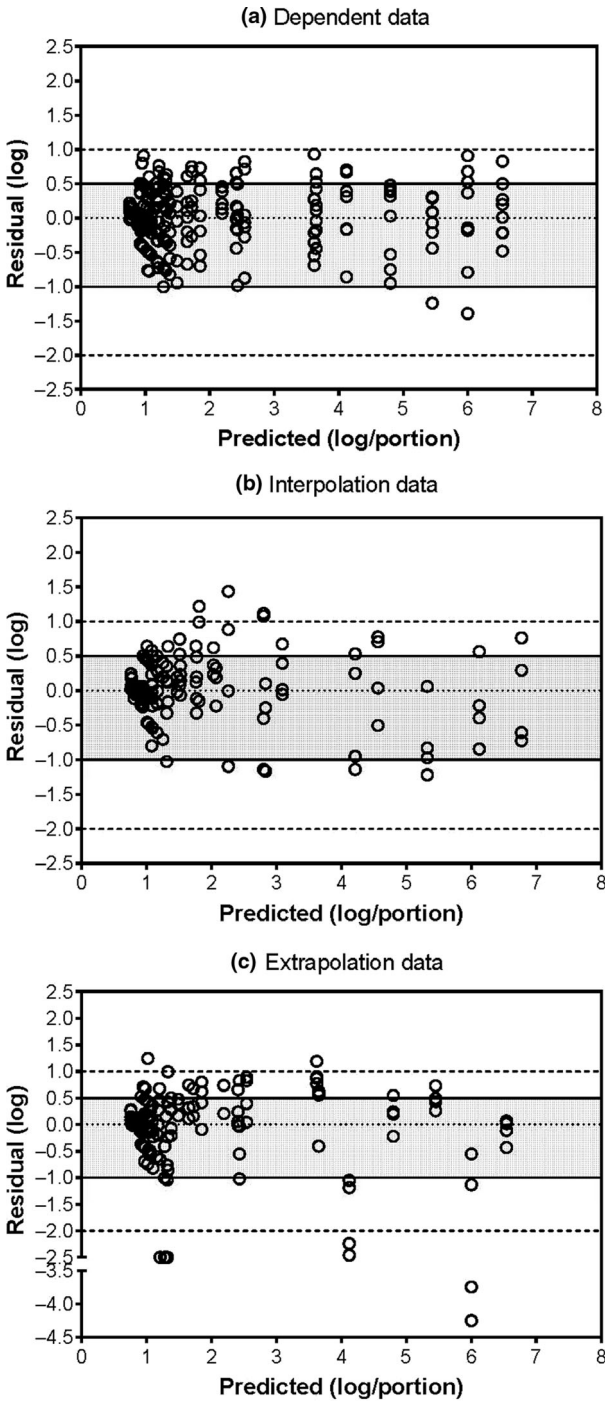

Figure 6 Acceptable prediction zones analyses for (a) dependent data; (b) independent data for interpolation; and (c) independent data for extrapolation to kosher chicken skin.

of fit. In fact, the B_f and A_f method was developed for external evaluation of models and thus is not intended for internal evaluation or validation (Ross, 1996). In contrast, pAPZ can be any value from zero to one, even for dependent data. Thus, it provides information about whether prediction bias for dependent data is acceptable or not (Oscar, 2005b).

Third, B_f does not detect some forms of systematic prediction bias because it is an average value (Ross, 1996). In contrast, pAPZ can detect all forms of prediction bias because it is a proportion and not an average value (Oscar, 2005b). Therefore, it determines whether systematic prediction bias is acceptable or not.

Fourth, the B_f and A_f method overestimates performance of models that predict log counts because it takes the log of a log ratio of observed and predicted values (Oscar, 2009a). In contrast, the APZ method does not overestimate performance of models that predict log counts because it uses residuals to compare observed and predicted values (Oscar, 2005a).

Fifth, as the denominator of B_f or A_f approaches a log count of zero, small differences between observed and predicted values are inflated (Oscar, 2009a), whereas as the denominator of B_f or A_f approaches larger log counts like ten, large differences between observed and predicted values are deflated resulting in a systematic bias in the calculation of these metrics (Oscar, 2014). In contrast, in the APZ method, as observed or predicted values approach a log count of zero or a larger log count like ten, the calculated residual is not inflated or deflated (Oscar, 2009a). Thus, the APZ method provides an unbiased evaluation of prediction error.

Sixth, criteria for B_f and A_f are based on opinion (Hu *et al.*, 2018) rather than a statistical analysis of data. Thus, they are arbitrary. In contrast, criteria for model performance in the APZ method are based on a statistical analysis of experimental error associated with log counts (Oscar, 2005a). Thus, they are not arbitrary.

Seventh, special prediction cases involving observed or predicted values of zero are not and cannot be included in calculation of B_f and A_f (Oscar, 2005b). In contrast, special prediction cases are included in the calculation of pAPZ as described above.

Eighth, there are no criteria for test data in the B_f and A_f method (Ross, 1996). In contrast, the APZ method has criteria for test data that ensure that the comparisons of observed and predicted values are valid and that the evaluation of model performance is accurate and unbiased (Oscar, 2005b).

Ninth, the criterion for an acceptable value of the A_f is dependent on the number of independent variables (Ross *et al.*, 2000). Thus, as a model gets more complex by addition of independent variables (x), the value for an acceptable A_f increases from 1.15 ($x = 1$) to 1.3 ($x = 2$) to 1.45 ($x = 3$) to 1.6 ($x = 4$)... until an inaccurate model is considered accurate. In contrast, the APZ method has criteria for model performance that are fixed and do not change as a function of the number of independent variables (Oscar, 2005a; Oscar,

Published 2020. This article is a U.S. Government work and is in the public domain in the USA

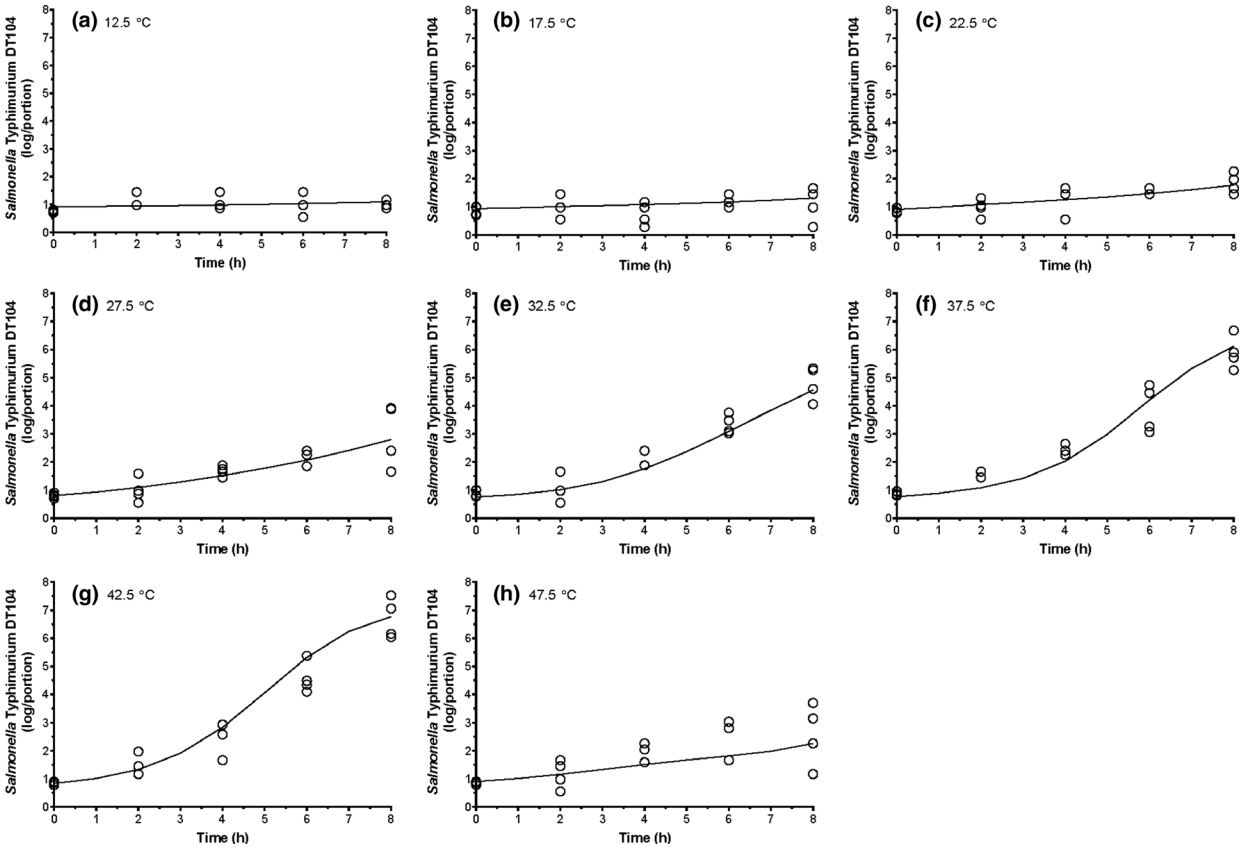

Figure 7 Growth curves for *Salmonella* Typhimurium DT104 and chicken skin incubated at (a) 12.5 °C; (b) 17.5 °C; (c) 22.5 °C; (d) 27.5 °C; (e) 32.5 °C; (f) 37.5 °C; (g) 42.5 °C; or (h) 47.5 °C. Open symbols are observed log counts, and solid lines are predicted log counts for interpolation data not used in model development.

2005b). Thus, the APZ method informs (pAPZ < 0.70) when a model has become too complex to provide reliable predictions.

Tenth, the B_f and A_f method does not evaluate models for local prediction problems. In contrast, the APZ method evaluates models for local prediction problems (Oscar, 2014; Oscar, 2018a). It is important to evaluate a model for its performance in different phases of the life cycle because higher model performance is expected when bacterial numbers are small and not changing in lag phase and lower model performance is expected when bacterial numbers are rapidly changing in exponential phase or are of greater magnitude in stationary phase (Oscar, 2006; Oscar, 2018b).

Eleventh, in a head-to-head comparison, the APZ method outperformed the B_f and A_f method at detecting prediction problems in secondary models for lag time and growth rate (Oscar, 2005b).

Twelfth, if a method cannot validate a model for dependent data, then it cannot reliably validate a model for independent data (Oscar, 2005b). In other words, if a method cannot perform internal validation, then it cannot reliably perform external validation. Thus, the B_f and A_f method (Ross, 1996) is not a reliable method for validating models in the field of predictive microbiology because it cannot perform internal validation, has no criteria for test data and has arbitrary criteria for model performance. Yet, it is the most widely used and accepted method for validation of models in the field of predictive microbiology (McKellar & Lu, 2004). In contrast, the APZ method can perform both internal and external validations of models using science-based and objective criteria for test data and model performance (Oscar, 2005a; Oscar, 2005b). Yet, it is not as widely used and accepted as the B_f and A_f method. The scientific basis for this situation is not clear.

These are twelve reasons why the APZ method (Oscar, 2005a; Oscar, 2005b) and not the B_f and A_f method (Ross, 1996) is used in ValT to validate models for predictive microbiology. Once ValT is published, it will be made available at: www.ars.usda.gov/nea/errc/PoultryFARM. It is hoped that ValT will make it easier for predictive microbiologists to properly develop and validate models using the APZ

Published 2020. This article is a U.S. Government work and is in the public domain in the USA

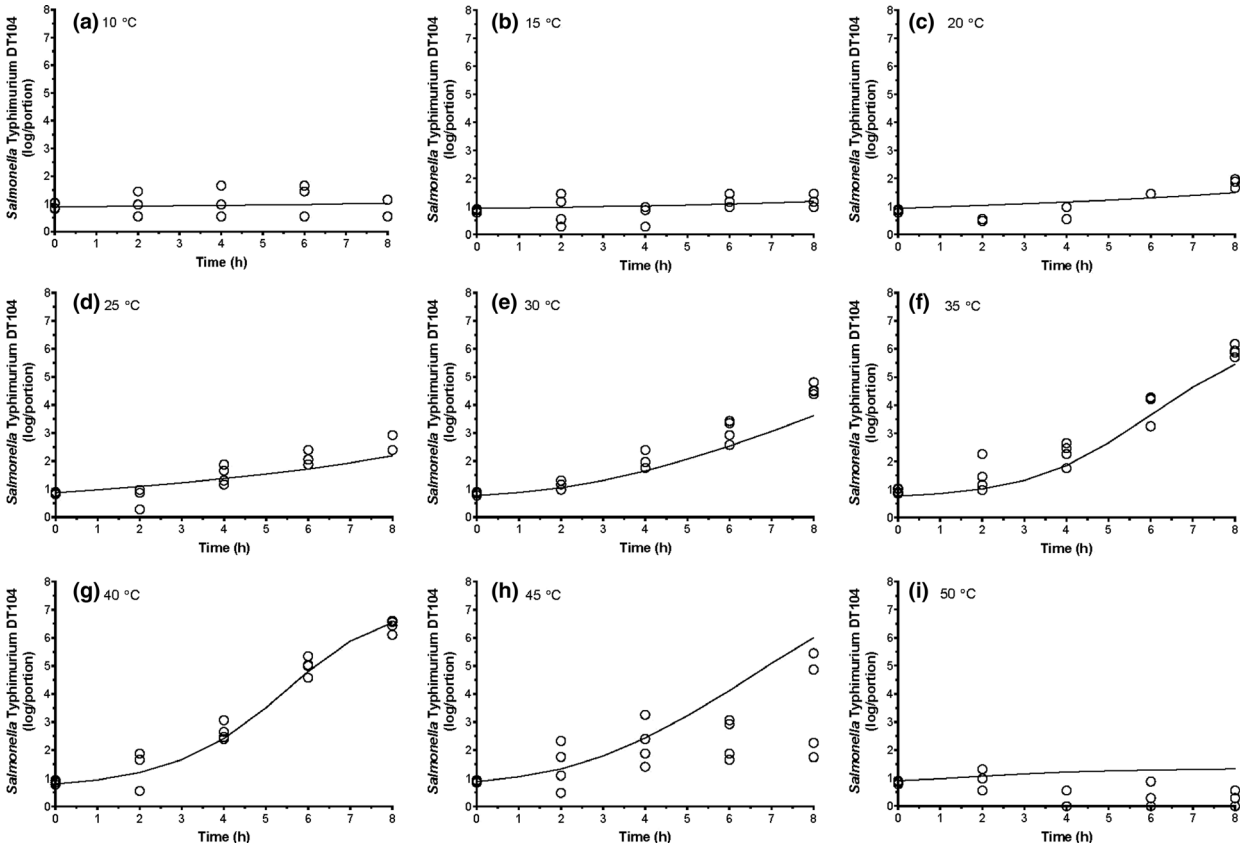

Figure 8 Growth curves for *Salmonella* Typhimurium DT104 and kosher chicken skin incubated at (a) 10 °C; (b) 15 °C; (c) 20 °C; (d) 25 °C; (e) 30 °C; (f) 35 °C; (g) 40 °C; (h) 45 °C; or (i) 50 °C. Open symbols are observed log counts, and solid lines are predicted log counts for extrapolation data not used in model development.

method. Ultimately, a shift from the B_f and A_f method (Ross, 1996) to the APZ method (Oscar, 2005a; Oscar, 2005b) in the field of predictive microbiology will benefit the food industry and regulatory agencies by providing models that are properly developed and validated using objective and science-based criteria for test data and model performance. This will allow users of predictive microbiology models to use them with greater confidence to make important food safety decisions that impact public health.

Acknowledgments

The author appreciates the outstanding technical assistance provided by Jacquelyn Ludwig (retired) and Stacey Engster of the U.S. Department of Agriculture, Agricultural Research Service, and Sharif Walker, Celia Whyte, Catherine Katambo and Elsie Dennis of the University of Maryland Eastern Shore. This project was funded by in-house project CRIS-8072-42000-079-00D of the Agricultural Research Service of the U. S. Department of Agriculture. Mention of trade names or commercial products is solely for providing specific information and does not imply recommendation or endorsement by the U. S. Department of Agriculture, which is an equal opportunity provider and employer.

Conflict of Interest

There is no conflict of interest.

Data availability statement

Research data are not shared.

References

Abou-Zeid, K.A., Oscar, T.P., Schwarz, J.G., Hashem, F.M., Whiting, R.C. & Yoon, K. (2009). Development and validation of a predictive model for *Listeria monocytogenes* Scott A as a function of temperature, pH, and commercial mixture of potassium lactate and sodium diacetate. *Journal of Microbiology and Biotechnology*, **19**, 718–726.

Baranyi, J. & Tamplin, M.L. (2004). ComBase: a common database on microbial responses to food environments. *Journal of Food Protection*, **67**, 1967–1971.

 Published 2020. This article is a U.S. Government work and is in the public domain in the USA

Baranyi, J., Ross, T., McMeekin, T.A. & Roberts, T.A. (1996). Effects of parameterization on the performance of empirical models used in 'predictive microbiology'. *Food Microbiology*, **13**, 83–91.

Baranyi, J., Pin, C. & Ross, T. (1999). Validating and comparing predictive models. *International Journal of Food Microbiology*, **48**, 159–166.

Bhaduri, S., Turner-Jones, C., Buchanan, R.L. & Phillips, J.G. (1994). Response surface model of the effect of pH, sodium chloride and sodium nitrite on growth of *Yersinia enterocolitica* at low temperatures. *International Journal of Food Microbiology*, **23**, 333–343.

Delignette-Muller, M.L., Rosso, L. & Flandrois, J.P. (1995). Accuracy of microbial growth predictions with square root and polynomial models. *International Journal of Food Microbiology*, **27**, 139–146.

Estilo, E.E.C. & Gabriel, A.A. (2018). A model for the influences of soluble and insoluble solids, and treated volume on the ultraviolet-C resistance of heat-stressed *Salmonella enterica* in simulated fruit juices. *Food Microbiology*, **69**, 72–81.

te Giffel, M.C. & Zwietering, M.H. (1999). Validation of predictive models describing the growth of *Listeria monocytogenes*. *International Journal of Food Microbiology*, **46**, 135–149.

Hu, J., Lin, L., Chen, M. & Yan, W. (2018). Modeling for predicting the time to detection of *Staphylococcal* enterotoxin A in cooked chicken product. *Frontiers in Microbiology*, **9**, 1536.

Jia, Z., Li, C., Fang, T. & Chen, J. (2019). Predictive modeling of the effect of epsilon-polylysine hydrochloride on growth and thermal inactivation of *Listeria monocytogenes* in fish balls. *Journal of Food Science*, **84**, 127–132.

Kuroda, S., Okuda, H., Ishida, W. & Koseki, S. (2019). Modeling growth limits of *Bacillus* spp. spores by using deep-learning algorithm. *Food Microbiology*, **78**, 38–45.

Li, M., Pradhan, A., Cooney, L. *et al*. (2011). A predictive model for the inactivation of *Listeria innocua* in cooked poultry products during postpackage pasteurization. *Journal of Food Protection*, **74**, 1261–1267.

Martino, K.G., Marks, B.P., Campos, D.T. & Tamplin, M.L. (2005). Quantifying the robustness of a broth-based model for predicting *Listeria monocytogenes* growth in meat and poultry products. *Journal of Food Protection*, **68**, 2310–2316.

McKellar, R.C. & Lu, X. (2004). *Modeling Microbial Responses in Food*. Boca Raton: CRC Press LLC.

Min, K.J. & Yoon, K.S. (2010). Development and validation of a predictive model for foodborne pathogens in ready-to-eat pork as a function of temperature and a mixture of potassium lactate and sodium diacetate. *Journal of Food Protection*, **73**, 1626–1632.

Mohr, T.B., Juneja, V.K., Thippareddi, H.H. *et al*. (2015). Assessing the performance of *Clostridium perfringens* cooling models for cooked, uncured meat and poultry products. *Journal of Food Protection*, **78**, 1512–1526.

de Oliveira Elias, S., Noronha, T.B. & Tondo, E.C. (2018). Assessment of *Salmonella* spp. and *Escherichia coli* O157:H7 growth on lettuce exposed to isothermal and non-isothermal conditions. *Food Microbiology*, **72**, 206–213.

Oscar, T.P. (2005a). Development and validation of primary, secondary and tertiary models for predicting growth of *Salmonella* Typhimurium on sterile chicken. *Journal of Food Protection*, **68**, 2606–2613.

Oscar, T.P. (2005b). Validation of lag time and growth rate models for *Salmonella* Typhimurium: acceptable prediction zone method. *Journal of Food Science*, **70**, M129–M137.

Oscar, T.P. (2006). Validation of a tertiary model for predicting variation of *Salmonella* Typhimurium DT104 (ATCC 700408) growth from a low initial density on ground chicken breast meat with a competitive microflora. *Journal of Food Protection*, **69**, 2048–2057.

Oscar, T.P. (2007). Predictive models for growth of *Salmonella* Typhimurium DT104 from low and high initial density on ground chicken with a natural microflora. *Food Microbiology*, **24**, 640–651.

Oscar, T.P. (2009a). General regression neural network and Monte Carlo simulation model for survival and growth of *Salmonella* on raw chicken skin as a function of serotype, temperature, and time for use in risk assessment. *Journal of Food Protection*, **72**, 2078–2087.

Oscar, T.P. (2009b). Predictive model for survival and growth of *Salmonella* Typhimurium DT104 on chicken skin during temperature abuse. *Journal of Food Protection*, **72**, 304–314.

Oscar, T.P. (2011a). Development and validation of a predictive microbiology model for survival and growth of *Salmonella* on chicken stored at 4 to 12°C. *Journal of Food Protection*, **74**, 279–284.

Oscar, T.P. (2011b). Extrapolation of a predictive model for growth of a low inoculum size of *Salmonella* Typhimurium DT104 on chicken skin to higher inoculum sizes. *Journal of Food Protection*, **74**, 1630–1638.

Oscar, T.P. (2013). Validation of a predictive model for survival and growth of *Salmonella* Typhimurium DT104 on chicken skin for extrapolation to a previous history of frozen storage. *Journal of Food Protection*, **76**, 1035–1040.

Oscar, T.P. (2014). General regression neural network model for behavior of *Salmonella* on chicken meat during cold storage. *Journal of Food Science*, **79**, M978–M987.

Oscar, T.P. (2015). Neural network model for survival and growth of *Salmonella enterica* serotype 8,20:-:z_6 in ground chicken thigh meat during cold storage: extrapolation to other serotypes. *Journal of Food Protection*, **78**, 1819–1827.

Oscar, T.P. (2017a). Modeling the effect of inoculum size on the thermal inactivation of *Salmonella* Typhimurium to elimination in ground chicken thigh meat. *American Journal of Food Science and Technology*, **5**, 135–142.

Oscar, T.P. (2017b). Neural network model for thermal inactivation of *Salmonella* Typhimurium to elimination in ground chicken: acquisition of data by whole sample enrichment, miniature most-probable-number method. *Journal of Food Protection*, **80**, 104–112.

Oscar, T.P. (2018a). Development and validation of a neural network model for predicting growth of *Salmonella* Newport on diced Roma tomatoes during simulated salad preparation and serving: extrapolation to other serotypes. *International Journal of Food Science and Technology*, **53**, 1789–1801.

Oscar, T.P. (2018b). Neural network model for growth of *Salmonella* Typhimurium in brain heart infusion broth. *International Journal of Food Science and Technology*, **53**, 2610–2616.

Ostergaard, N.B., Eklow, A. & Dalgaard, P. (2014). Modelling the effect of lactic acid bacteria from starter- and aroma culture on growth of *Listeria monocytogenes* in cottage cheese. *International Journal of Food Microbiology*, **188**, 15–25.

Oyarzabal, O.A., Oscar, T.P., Speegle, L. & Nyati, H. (2010). Survival of *Campylobacter jejuni* and *Campylobacter coli* on retail broiler meat stored at -20, 4, or 12°C and development of Weibull models for survival. *Journal of Food Protection*, **73**, 1438–1446.

Panisello, P.J. & Quantick, P.C. (1998). Application of FoodMicro-Model predictive software in the development of hazard analysis critical control point (HACCP) systems. *Food Microbiology*, **15**, 425–439.

Ross, T. (1996). Indices for performance evaluation of predictive models in food microbiology. *Journal of Applied Bacteriology*, **81**, 501–508.

Ross, T., Dalgaard, P. & Tienungoon, S. (2000). Predictive modeling of the growth and survival of *Listeria* in fishery products. *International Journal of Food Microbiology*, **62**, 231–245.

Tamplin, M.L. (2018). Integrating predictive models and sensors to manage food stability in supply chains. *Food Microbiology*, **75**, 90–94.

Zwietering, M.H., Cuppers, H.G.A.M., deWit, J.C. & van 'T Riet, K. (1994). Evaluation of data transformations and validation of a model for the effect of temperature on bacterial growth. *Applied and Environmental Microbiology*, **60**, 195–203.

Published 2020. This article is a U.S. Government work and is in the public domain in the USA