Validation of Lag Time and Growth Rate Models for *Salmonella* Typhimurium: Acceptable Prediction Zone Method

THOMAS P. OSCAR

ABSTRACT: The prediction bias (B_f) and accuracy (A_f) factors are the most widely used measures of performance of predictive models for food pathogens. However, B_f and A_f have limitations that can produce inaccurate assessments of model performance. Consequently, an objective of the current study was to develop a method for quantifying model performance that overcomes limitations of B_f and A_f. Performance of published lag time and growth rate models for *Salmonella* Typhimurium were evaluated for data used in model development and for data not used in model development but that were inside (interpolation) or outside (extrapolation) the response surface of the models. In addition, performance of published models for growth of *Escherichia coli* O157:H7 was evaluated for data used in model development. Observed and predicted values were compared using B_f, A_f, and pRE, a new performance factor that quantified the proportion of relative errors (RE) in an acceptable prediction zone from an RE of –0.3 (fail-safe) to 0.15 (fail-dangerous). A decision diagram based on criteria for test data and model performance was used to validate the models. When B_f and A_f were used to quantify model performance, all models were validated. In contrast, when pRE was used to evaluate model performance, 2 models for *S.* Typhimurium and both models for *E. coli* O157:H7 failed validation. Overall, pRE was a more sensitive and reliable indicator of model performance than B_f and A_f because unacceptable pRE, which indicated a performance problem, were obtained for 8 of 20 evaluations, all of which had acceptable B_f and A_f. A limitation of pRE was the inability to distinguish between global and regional prediction problems. However, when used in combination with an RE plot, pRE provided a complete evaluation of model performance that overcame limitations of B_f and A_f.

Keywords: performance evaluation, growth model, verification, validation, *Salmonella* Typhimurium

Introduction

Mathematical models that predict the growth of foodborne pathogens or spoilage organisms are used in the food industry to help assess the microbiological safety and shelf-life of food, respectively. Most models for pathogen growth were derived from kinetic data collected in broth media of different pH and water activities and incubated at temperatures commonly encountered in food processing and storage (McClure and others 1994). The underlying principle of such models is that by accounting for the major factors that control microbial growth, models developed in broth can provide acceptable predictions when extrapolated to food (Buchanan 1991).

An important step in development of broth models is evaluation of predictions against data collected with food (Ross 1996). The most common method of evaluation is to compare broth model predictions to published data for pathogen growth in food (Gibson and others 1988; Buchanan and Phillips 1990). However, these types of comparisons are not valid when food data are obtained with different strains, previous growth conditions, or modeling methods. The latter statement is controversial because it is widely accepted in predictive microbiology that use of a broth model to predict pathogen growth in food is not extrapolation but application of the model. Interestingly, predictive microbiologists often state that it is important not to use a model outside the range of the independent variables used to develop the model because the model may not provide reliable predictions (Buchanan and others 1993). In other words, it is widely accepted that if a broth model for pathogen growth were developed for a sodium chloride range of 0.5 to 2.5%, that use of the model to predict pathogen growth at a sodium chloride level of 3% would be extrapolation. Thus, it is unclear why it is not considered extrapolation when a broth model for pathogen growth is developed with a fat level of 0% and then used to predict the growth of the pathogen in food with a fat level of 20%. Moreover, if a model were developed to predict the growth of strain A in broth and the performance of the model were evaluated for predicting the growth of strain B in food, how could this be a valid evaluation of the ability of the model to extrapolate to another strain or to food when the comparisons of observed and predicted values are confounded by differences in strain and growth medium? The approach adopted in the current study, but 1st published by others (Gibson and others 1988; Ross 1996), was to evaluate broth model predictions with food data obtained in controlled challenge studies with the same strain, previous growth conditions, and modeling methods.

Calculation of prediction bias and accuracy factors (Delignette-Muller and others 1995; Ross 1996) and use of graphical methods (Bratchell and others 1990; McClure and others 1993) to assess systematic prediction bias are important components of the evaluation process. However, criteria for acceptable values of performance factors that allow a determination of whether a model provides valid predictions of pathogen growth have not been

MS 20040596 Submitted 9/3/04, Accepted 11/29/04. The author is with Microbial Food Safety Research Unit, Agricultural Research Service, U.S. Dept. of Agriculture, Univ. of Maryland Eastern Shore, Room 2111, Center for Food Science and Technology, Princess Anne, MD 21853. Direct inquiries to author Oscar (E-mail: toscar@umes.edu).

Published on Web 2/14/2005

established (te Giffel and Zwietering 1999). Establishment of standard methods and criteria for validating predictive models will provide users of predictive models with greater confidence in using the models to assess food safety. Currently, pathogen-detection methods are used to a greater extent in the food industry to assess food safety than predictive models. In part, this is due to the official validation process provided by the Association of Official Analytical Chemists (AOAC). If predictive models were subjected to an AOAC-like validation process, they may find much wider use in the food industry.

The most widely used method in predictive microbiology for quantifying model performance is the ratio method of Ross (1996), which involves calculating prediction bias (B_f) and accuracy (A_f) factors. However, limitations of this method that can result in an inaccurate assessment of model performance and improper validation are as follows: (1) B_f and A_f are mean values that do not detect some forms of prediction bias (Ross 1996), (2) B_f and A_f are mean values that are subject to bias by outliers (Delignette-Muller and others 1995), and (3) prediction cases involving no growth are excluded from calculation of B_f and A_f, resulting in an overestimation of model performance (Augustin and Carlier 2000). A method of performance evaluation is needed that overcomes these limitations.

In the current study, published response surface models for lag time and growth rate of *Salmonella* Typhimurium (Oscar 1999a, 1999b, 1999c) were evaluated for the ability to predict the data used to develop them and for the ability to predict data not used in model development but that were inside (interpolation) or outside (extrapolation) the response surface. In addition, published response surface models for growth of *Escherichia coli* O157:H7 were evaluated for the ability to predict the data used to develop them.

Test data for performance evaluation of the *S.* Typhimurium growth models were collected in controlled challenge studies with the same strain, previous growth conditions, and modeling methods so as not to confound the comparisons of observed and predicted values. Although prediction bias and accuracy factors were reported and model predictions were assessed graphically for systematic prediction bias in the previous studies (Oscar 1999a, 1999b, 1999c), criteria for determining whether the models provided predictions with acceptable bias and accuracy were not established. In addition, the models were not tested for the ability to extrapolate to other variables (for example, growth media).

Model performance in the present study was quantified using a new acceptable prediction zone method that overcomes the aforementioned limitations of B_f and A_f. In addition, a decision diagram based on criteria for test data and acceptable performance factors was developed and used to determine whether models provided valid predictions of pathogen growth.

Materials and Methods

Salmonella Typhimurium growth models

Response surface models for lag time (λ) and maximum specific growth rate (μ_{max}) of the same strain of *Salmonella* Typhimurium (ATCC 14028, American Type Culture Collection, Manassas, Va., U.S.A.) were developed in a series of 3 studies (Oscar 1999a, 1999b, 1999c) using the same primary and secondary modeling methods. Growth kinetic data (that is, log colony-forming units [CFU]/mL or cm^2) were fit to a 2-phase linear primary model (Buchanan and others 1997) to determine λ and μ_{max}. Natural logarithm (ln) transformations of λ and μ_{max} were used in regression analysis to obtain secondary response surface models (Oscar 1999a, 1999b, 1999c).

Details of the experimental methods used to collect and model the kinetic data have been published (Oscar 1999a, 1999b, 1999c). In brief, growth kinetics of stationary phase cells of *S.* Typhimurium were measured in the absence of microbial competition in brain heart infusion (BHI) broth and on sterilized cooked chicken breast meat (sBM). Changes in *S.* Typhimurium density as a function of time were assessed using viable cell counts on BHI agar and then modeled as described previously (Oscar 1999a, 1999b, 1999c).

The experimental design in Study I (Oscar 1999c) for model development was a full 4 × 6 × 3 factorial arrangement of previous growth pH (5.7, 6.7, 7.8, 8.6), temperature (15 °C, 20 °C, 25 °C, 30 °C, 35 °C, 40 °C), and pH (5.2, 6.3, 7.4) in BHI broth. Response surface models were of the following form:

$$\ln \lambda \text{ or } \ln \mu_{max} = b_0 + b_1A + b_2T + b_3P + b_4AT + b_5AP + \\ + b_6TP + b_7A^2 + b_8T^2 + b_9P^2$$

where b_0 to b_9 were regression coefficients that were published previously (Oscar 1999c), A was previous growth pH, T was temperature, and P was pH. The model for ln λ was designated Model 1, whereas the model for ln μ_{max} was designated Model 2 in the present study.

The experimental design for model development in Study II (Oscar 1999b) was a full 4 × 4 factorial arrangement of previous growth temperature (16, 22, 28, 34 °C) in BHI broth and temperature (16, 22, 28, 34°C) on sBM. Response surface models were of the form:

$$\ln \lambda \text{ or } \ln \mu_{max} = b_0 + b_1B + b_2T + b_3BT + b_4B^2 + b_5T^2$$

where b_0 to b_5 were regression coefficients that were published previously (Oscar 1999c), B was previous growth temperature and T was temperature. The model for ln λ was designated Model 3, whereas the model for ln μ_{max} was designated Model 4 in this study.

The experimental design for model development in Study III (Oscar 1999a) was a full 3 × 9 factorial arrangement of previous growth sodium chloride (0.5%, 2.5%, 4.5%) in BHI broth and temperature (10 °C, 12 °C, 14 °C, 16 °C, 20 °C, 24 °C, 28 °C, 34 °C, 40 °C) on sBM. Response surface models were of the form:

$$\ln \lambda \text{ or } \ln \mu_{max} = b_0 + b_1C + b_2T + b_3CT + b_4C^2 + b_5T^2$$

where b_0 to b_5 were regression coefficients that were published previously (Oscar 1999a), C was previous growth sodium chloride, and T was temperature. The model for ln λ was designated Model 5, whereas the model for ln μ_{max} was designated Model 6 in the current study.

Escherichia coli O157:H7 growth models

Response surface models for aerobic growth of a 3 strain cocktail of *E. coli* O157:H7 in BHI broth were developed using a fractional factorial design of temperature (5 °C to 42 °C), pH (4.5 to 8.5), and sodium chloride level (5 to 50 g/L) as described previously (Buchanan and others 1993). Growth kinetic data were fit to the Gompertz model and ln transformations of the Gompertz parameters B and M were used in regression analysis to obtain the secondary response surface models:

$$\ln B \text{ or } \ln M = b_0 + b_1S + b_2T + b_3P + b_4ST + b_5SP + \\ + b_6TP + b_7S^2 + b_8T^2 + b_9P^2$$

where b_0 to b_9 were regression coefficients that were published previously (Buchanan and others 1993), S was sodium chloride, T was temperature, and P was pH.

Table 1—Observed and predicted lag times (λ) and maximum specific growth rates (μ_{max}) of *Salmonella* Typhimurium ATCC 14028 on sterilized cooked chicken breast and thigh meat: data for performance evaluation of extrapolation of Models 1 and 2 in Study I

	Response surface model parameters			λ (h)		μ_{max} (/h)	
Meat type	Previous growth pH	Temperature	pH	Observed	Predicted	Observed	Predicted
Breast	6.3	15	6.04	12.97	11.79	0.092	0.114
Breast	8.3	15	6.04	17.34	12.29	0.099	0.112
Breast	6.7	20	6.07	8.05	6.47	0.140	0.219
Breast	7.8	20	6.07	8.59	6.67	0.171	0.220
Breast	6.7	25	6.04	4.10	3.90	0.386	0.370
Breast	6.3	25	6.26	4.06	3.71	0.483	0.386
Breast	5.7	30	6.07	2.92	2.36	0.521	0.564
Breast	6.3	30	6.09	1.82	2.53	0.493	0.550
Breast	6.3	35	6.26	1.43	1.87	0.596	0.710
Breast	8.6	35	6.26	1.53	1.99	0.689	0.716
Breast	6.3	40	6.09	1.26	1.61	0.628	0.747
Breast	7.4	40	6.09	1.86	1.74	0.643	0.732
Thigh	6.3	15	6.84	13.49	11.70	0.082	0.124
Thigh	8.3	15	6.84	14.99	11.88	0.090	0.123
Thigh	5.7	20	6.93	6.73	5.94	0.200	0.252
Thigh	6.7	20	6.93	8.13	6.42	0.208	0.243
Thigh	6.3	25	6.84	2.69	3.71	0.387	0.415
Thigh	5.7	25	6.99	3.14	3.55	0.348	0.433
Thigh	7.4	30	6.93	2.32	2.63	0.568	0.600
Thigh	8.3	30	7.00	2.25	2.61	0.556	0.613
Thigh	6.3	35	6.99	1.22	1.89	0.688	0.778
Thigh	5.7	35	6.99	1.24	1.77	0.753	0.803
Thigh	6.3	40	7.00	0.85	1.59	0.736	0.846
Thigh	7.4	40	7.00	1.63	1.69	0.895	0.825

Test data for *Salmonella* Typhimurium growth models

Independent data for performance evaluation of interpolation were collected with the same strain, growth media, and modeling methods but different combinations of the independent variables that were within the response surface of the model. Independent data for performance evaluation of extrapolation were collected in the same manner except that the growth media used to measure growth kinetics was different and thus, the response surface models were evaluated for the ability to extrapolate to a different growth medium.

The experimental design for interpolation in Study I (Oscar 1999c) was a full 3 × 5 × 2 factorial arrangement of previous growth pH (6.3, 7.4, 8.3), temperature (17.5 °C, 22.5 °C, 27.5 °C, 32.5 °C, 37.5 °C) and pH (5.7, 6.7) in BHI broth. Growth conditions presented in Table 1 were used for evaluating extrapolation of broth Models 1 and 2 to sBM and to sterilized cooked chicken thigh meat (sTM). The pH of sBM and sTM homogenates (6 g meat:94 mL distilled water) was determined with a combination pH electrode attached to a model Φ34 pH meter (Beckman Instruments, Fullerton, Calif., U.S.A.).

The experimental design for interpolation in Study II (Oscar 1999b) was a full 3 × 3 factorial arrangement of previous growth temperature (19 °C, 25 °C, 31 °C) in BHI broth and temperature (19 °C, 25 °C, 31 °C) on sBM. Data from an experiment with a full 4 × 4 factorial arrangement of previous growth temperature (16 °C, 22 °C, 28 °C, 34 °C) in BHI broth and temperature (16 °C, 22 °C, 28 °C, 34 °C) on sTM and data from an experiment with a full 3 × 3 factorial arrangement of previous growth temperature (19 °C, 25 °C, 31 °C) in BHI broth and temperature (19 °C, 25 °C, 31 °C) on sTM (Oscar, unpublished) were combined for evaluating extrapolation of Models 3 and 4.

The experimental design for interpolation in Study III (Oscar 1999a) was a full 2 × 8 factorial arrangement of previous growth sodium chloride (1.5%, 3.5%) in BHI broth and temperature (11 °C, 13 °C, 15 °C, 18 °C, 22 °C, 26 °C, 31 °C, 37 °C) on sBM. Ability of Models 5 and 6 for extrapolation to another growth medium was not tested.

Test data for *Escherichia coli* O157:H7 growth models

Observed and predicted values (ln transformation) of lag time and generation time (τ) (Buchanan and others 1993) for aerobic growth of *E. coli* O157:H7 were used to evaluate the ability of the *E. coli* O157:H7 models to predict the data used in model development. Predicted values of lag time and generation time were derived using predictions from the response surface models for the B and M parameters of the Gompertz model and a fixed value of 6.34 for the C parameter of the Gompertz model (Buchanan and others 1993). This dataset contained 25 no growth prediction cases where the models predicted growth but no growth occurred.

Performance evaluation

Prediction bias (B_f) and accuracy (A_f) factors were calculated as described by Ross (1996) except that different ratios of observed and predicted values were used for λ and μ_{max} so that B_f less than 1 represented fail-safe predictions and B_f above 1 represented fail-dangerous predictions. Likewise, relative errors (RE) of individual prediction cases were calculated (Delignette-Muller and others 1995):

$$\text{RE for } \lambda = (\text{predicted} - \text{observed})/\text{predicted}$$

$$\text{RE for } \mu_{max} = (\text{observed} - \text{predicted})/\text{predicted}$$

such that RE less than zero represented fail-safe predictions and RE above zero represented fail-dangerous predictions. For prediction cases in which observed or predicted lag time was infinity or predicted lag time, generation time or growth rate was zero, the RE signed a value of –1 for graphical presentation. This is an important

Table 2—Performance of growth models for *Salmonella* Typhimurium ATCC 14028 in brain heart infusion broth or on sterilized cooked chicken breast and thigh meat: prediction bias and accuracy factors[a]

Study	Model	Growth parameter	Growth medium	Model parameters	Dataset	*n*	B_f	A_f	pRE
I	1	λ	Broth	A, T, P	Development	75	1.000	1.105	0.893
I	2	μ_{max}	Broth	A, T, P	Development	75	1.000	1.059	0.933
II	3	λ	Breast	B, T	Development	32	1.000	1.157	0.688
II	4	μ_{max}	Breast	B, T	Development	32	1.000	1.074	0.875
III	5	λ	Breast	C, T	Development	55	1.000	1.206	0.673
III	6	μ_{max}	Breast	C, T	Development	55	1.000	1.149	0.875
I	1	λ	Broth	A, T, P	Interpolation	30	0.979	1.082	0.933
I	2	μ_{max}	Broth	A, T, P	Interpolation	30	1.031	1.069	0.900
II	3	λ	Breast	B, T	Interpolation	18	0.951	1.151	0.889
II	4	μ_{max}	Breast	B, T	Interpolation	18	0.913	1.109	1.000
III	5	λ	Breast	C, T	Interpolation	16	1.089	1.262	0.438
III	6	μ_{max}	Breast	C, T	Interpolation	16	0.937	1.156	0.813
I	1	λ	Breast	A, T, P	Extrapolation	12	0.981	1.225	0.583
I	1	λ	Thigh	A, T, P	Extrapolation	12	1.128	1.275	0.667
I	2	μ_{max}	Breast	A, T, P	Extrapolation	12	0.884	1.182	0.833
I	2	μ_{max}	Thigh	A, T, P	Extrapolation	12	0.861	1.178	0.917
II	3	λ	Thigh	B, T	Extrapolation	50	1.011	1.190	0.680
II	4	μ_{max}	Thigh	B, T	Extrapolation	50	1.050	1.102	0.820

[a] A = previous growth pH; A_f = accuracy factor; B = previous growth temperature; B_f = bias factor; *n* = number of prediction cases; C = previous growth sodium chloride; *P* = pH; pRE = proportion of relative errors (RE) in the acceptable prediction zone; λ = lag time; μ_{max} = maximum specific growth rate; T = temperature.

Table 3—Validation of growth models for *Salmonella* Typhimurium ATCC 14028: results of the decision diagram (Figure 1) for the acceptable prediction zone method

	Development			Interpolation				Extrapolation			
Model	Q1	Q2	Outcome	Q3	Q4	Q5	Outcome	Q6	Q7	Q8	Outcome
1	Yes	Yes	Verified	Yes	Yes	Yes	Validated	Yes	Yes	No	Not Validated
2	Yes	Yes	Verified	Yes	Yes	Yes	Validated	Yes	Yes	Yes	Validated
3	Yes	No	Not Verified	No	—	—	Not Validated	No	—	—	Not Validated
4	Yes	Yes	Verified	Yes	Yes	Yes	Validated	Yes	Yes	Yes	Validated
5	Yes	No	Not Verified	No	—	—	Not Validated				
6	Yes	Yes	Verified	Yes	Yes	Yes	Validated				

feature of the acceptable prediction zone method because it allows the inclusion of no growth prediction cases in the calculation of its performance factor pRE. In contrast, no growth prediction cases are excluded from the calculation of the performance factors B_f and A_f, which results in an overestimation of model performance.

In the original publications of the growth models for *S.* Typhimurium (Oscar 1999a, 1999b, 1999c), RE was calculated using the observed value in the denominator, whereas in the current study, RE was calculated using the predicted value in the denominator. Statistically similar but not identical results are obtained with both of the methods for calculating RE. For example, the mean absolute RE (MARE) ± standard error of the mean (SEM) for Model 1 for interpolation, which had 30 prediction cases (Table 2), was 7.8% ± 1.1% when the observed value was used in the denominator to calculate RE (Oscar 1999c). In comparison, MARE was 8.1% ± 1.2% when the predicted value was used in the denominator to calculate RE. When compared using a “*t*-test, the MARE was not different (*P* = 0.87) between the 2 methods for calculating RE.

The proportion of RE (pRE) that fell in an acceptable prediction zone (that is, the number of RE in the acceptable prediction zone/total number of prediction cases) from an RE of –0.3 (fail-safe) to 0.15 (fail-dangerous) was calculated and used as a new measure of model performance. The boundaries of the acceptable prediction zone were equivalent to those proposed by Ross and others (2000) for acceptable values of B_f, which are 0.7 (fail-safe) to 1.15 (fail-dangerous). Models with pRE ≥ 0.700 were considered to provide predictions with acceptable bias and accuracy. The acceptable value for pRE was based on the gold standard in the U.S. education system for an acceptable test grade, which is 70% correct answers. In the acceptable prediction zone method, 70% correct answers correspond to 70% of the prediction errors falling inside the acceptable prediction zone.

The acceptable prediction zone was wider in the fail-safe direction because greater prediction error can be tolerated in the fail-safe direction when using models to predict food safety (Ross and others 2000). In fact, use of models that provide overly fail-safe predictions results in destruction of safe food that otherwise would benefit public health by maintaining consumer health and resistance to infectious disease. On the other hand, use of models that provide overly fail-dangerous predictions results in consumption of unsafe food and an increase in foodborne illness.

pRE is a relative measure of model performance because the width of the acceptable prediction zone affects its value. Pathogen incidence is an example of a relative performance factor that is widely accepted and used in the food industry. The size of sample used to assess pathogen incidence affects its value in much the same way that the width of the acceptable prediction zone affects the value of pRE. As the size of sample used to determine pathogen incidence increases, pathogen incidence increases in a nonlinear manner toward 100% (Oscar 2004). As an example, Surkiewicz and others (1969) found that the incidence of *Salmonella* contamination of whole chickens was 4.9% when a 10-mL sample of carcass rinse was used but was 20.5% when a 270-mL sample of carcass rinse was used. In an analogous manner, pRE increases toward 1.000 as the width of the acceptable prediction zone is increased. Both pRE and pathogen incidence are reliable and reproducible performance factors as long as standard methods are used to determine them.

Model validation

The decision diagram in Figure 1 was used to determine whether models provided valid predictions of pathogen growth. For a model to be validated for interpolation it had to be verified, it had to meet criteria for test data for interpolation, which are described subsequently, and it had to meet performance criteria for pRE (that is, pRE ≥ 0.700). For a model to be validated for extrapolation, it had to be validated for interpolation, meet criteria for test data for extrapolation (described subsequently), and meet performance criteria for pRE.

Results and Discussion

Evaluation of the ability of a model to predict the data used to develop it (verification) is important because a model that does not properly fit the data used to develop it would not provide valid predictions within or beyond its response surface. In the current study, models that failed performance evaluation for verification automatically failed performance evaluations for interpolation and extrapolation (Figure 1).

As expected, B_f for all model verifications were 1.000, a value that indicates no average bias. The expected value for B_f was 1 because models were developed by least squares regression in which the average deviation of observed values from predicted values is zero and because B_f is the ratio of observed and predicted values its average value for data used in model development is one. In contrast to B_f, which did not vary among models, A_f for model verification ranged from 1.059 to 1.206, where an A_f of 1.000 indicates perfect agreement between observed and predicted values (Table 2). Model 5 from Study III was the least accurate with an A_f of 1.206. The proportion of RE for verification that resided in the acceptable prediction zone (Figure 2) ranged from 0.673 to 0.933 (Table 2). Overall, pRE was higher, indicating better performance, for broth models of Study I than sBM models of Studies II and III (Table 2). In addition, pRE was higher for μ_{max} than λ models (Table 2). This was also observed by Wei and others (2001) who reported higher A_f for λ models than μ_{max} models for *Yersinia enterocolitica* and chicken meat. The poorer performance of λ models might be related to the higher biological variation of this parameter that is often observed in challenge studies (Lebert and others 1998; Oscar 2000).

Evaluation of the ability of a model to interpolate within its response surface is important because a model that does not provide predictions with acceptable accuracy and bias within its response surface is a model that does not properly define its response surface. Proper evaluation for interpolation requires an independent set of data collected with the same strain, previous growth conditions, and modeling methods. In addition, test data should uniformly cover the response surface to provide a complete and unbiased test of model performance. Failure to collect test data in the proper manner can invalidate the entire performance evaluation or result in only a partial validation of the model for interpolation (that is, only for that portion of the response surface for which test data were collected). In the current study, models were evaluated for interpolation using independent data collected with the same strain, previous growth conditions, and modeling methods and experimental designs that provided uniform and complete coverage of the response surface. Thus, all performance evaluations for interpolation in this study met the aforementioned criteria for test data for interpolation.

Model 4 had the lowest B_f of 0.913 for interpolation, whereas Model 5 had the highest B_f of 1.089 for interpolation (Table 2). Most RE for interpolation of Model 2 were greater than 1 (Figure 3b), which indicated fail-dangerous predictions, whereas most RE for Models 1, 3, and 4 were less than 1 (Figure 3a, 3c, 3d, respectively), which indicated fail-safe predictions. However, all or most RE for these models were inside the acceptable prediction zone (Figure 3a to 3d) for pRE of 0.889 to 1.000 (Table 2). Thus, prediction bias of Model 2 for interpolation was not overly fail-dangerous and prediction bias of Models 1, 3, and 4 for interpolation was not overly fail-safe.

Model 5 had the highest B_f of 1.089, highest A_f of 1.262, and lowest pRE of 0.438 for interpolation (Table 2). Predictions of Model 5 were overly fail-dangerous (that is, RE > 0.15) for 7 of 16 prediction cases (Figure 3e). Model 6 had a B_f of 0.937, an A_f of 1.156, and a pRE of 0.813 for interpolation (Table 2). Two of 3 prediction cases for Model 6 that were outside the acceptable prediction zone were very close to the boundaries (Figure 3f).

Successful extrapolation of model predictions to variables (for example, other strains or growth media) not included in the model can save time and money by eliminating the need to develop additional models. Proper evaluation of model performance for extrapolation requires an independent set of data that differs from the data used in model development by only 1 variable. Use of datasets that differ by more than 1 variable from the dataset used in model development will confound the comparison of observed and predicted values and thus invalidate the performance evaluation for extrapolation. In the present study, model performance for extrapolation was evaluated using independent sets of data

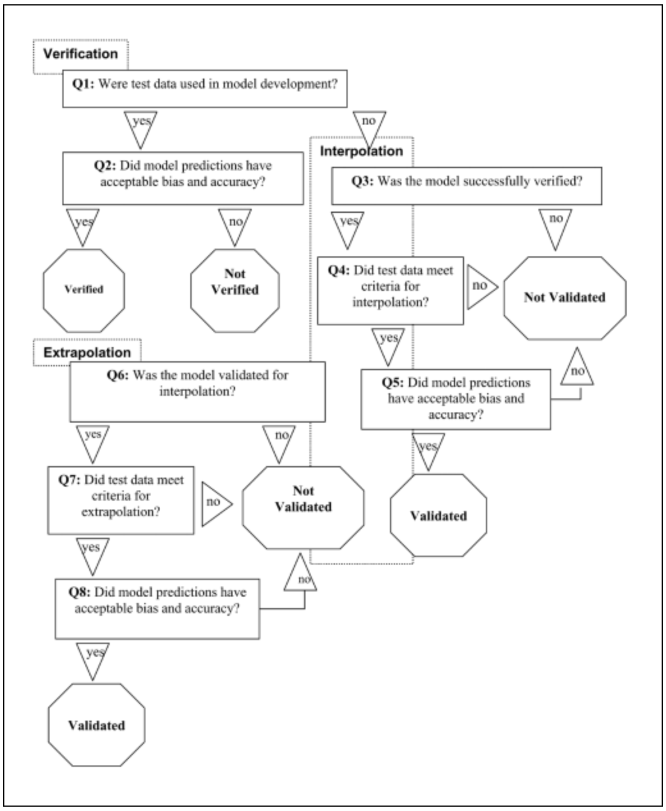
Verification
Q1: Were test data used in model development?
yes
no
Interpolation
Q2: Did model predictions have acceptable bias and accuracy?
Q3: Was the model successfully verified?
yes
no
yes
no
Verified
Not Verified
Q4: Did test data meet criteria for interpolation?
no
Not Validated
yes
no
Extrapolation
Q6: Was the model validated for interpolation?
Q5: Did model predictions have acceptable bias and accuracy?
yes
no
yes
Q7: Did test data meet criteria for extrapolation?
no
Not Validated
Validated
yes
no
Q8: Did model predictions have acceptable bias and accuracy?
yes
Validated

Figure 1—Decision diagram for validation of predictive models

obtained with the same strain, previous growth conditions, and modeling methods but different growth media. Thus, all performance evaluations for extrapolation in the current study met the aforementioned criteria for test data for extrapolation.

B_f for extrapolation of Model 1, which was developed in broth, to sBM was 0.981 and to sTM was 1.128 (Table 2). Although B_f for extrapolation to sBM and sTM were acceptable, the RE plot (Figure 4a) indicated that broth Model 1 provided biased predictions of λ when extrapolated to sBM and sTM. Specifically, Model 1 provided overly fail-dangerous predictions at short λ (<4 h) and fail-safe but not overly fail-safe predictions at longer λ (Figure 4a). Model 1 had an A_f of 1.225 for sBM and an A_f of 1.275 for sTM. pRE for extrapolation of broth Model 1 were 0.583 for sBM and 0.667 for sTM (Table 2).

Broth Model 2 when extrapolated to sBM and sTM made fail-safe predictions at most μ_{max} but only 3 RE were outside the acceptable prediction zone (Figure 4b) and thus, the biased predictions of Model 2 were not overly fail-safe. B_f for extrapolation of broth Model 2 to sBM and sTM were 0.884 and 0.861, respectively (Table 2). Broth Model 2 had A_f of 1.182 for sBM and 1.178 for sTM and pRE of 0.833 and 0.917 for sBM and sTM, respectively (Table 2).

Models 3 and 4 were developed on sBM and tested for extrapolation to sTM. Model 3 for λ had a slightly lower B_f (1.011 versus 1.050), a higher A_f (1.190 versus 1.102), and a lower pRE (0.680 versus 0.820) than Model 4 for μ_{max} (Table 2). Most RE that were outside the acceptable prediction zone for extrapolation of Model 3 to sTM were at short λ and were fail-dangerous (Figure 4c). Most RE for μ_{max} that were outside the acceptable prediction zone for extrapolation of Model 4 to sTM were fail-dangerous but close to the upper bound of the acceptable prediction zone (Figure 4d).

There is currently no consensus as to what values of B_f and A_f constitute a model that provides acceptable predictions of pathogen growth in broth or on food. However, for growth rate, B_f from 0.700 to 1.150 are considered acceptable (Ross and others 2000). In the current study, all B_f for *S.* Typhimurium growth models were in this range (Table 2). In general, A_f increases by 0.1 to 0.15 per independent variable in the model (Ross and others 2000). Thus, models with 2 independent variables, such as Models 3 to 6 in the present study, would be expected to have A_f < 1.300 and models with 3 independent variables, such as Models 1 and 2 in this study, would be expected to have A_f < 1.450. All *S.* Typhimurium models had A_f below 1.300 (Table 2). In contrast to B_f and A_f, pRE was unacceptable (that is, <0.700) for 6 of 18 performance evaluations (Table 2).

A limitation of B_f and A_f for evaluation of model performance is that criteria for acceptable B_f and A_f are not consistent with each other. More specifically, acceptable values for B_f are fixed (that is, 0.700 to 1.150) and independent of the number of model variables and criteria for acceptable B_f consider whether predictions err more in the fail-safe direction. In contrast, acceptable A_f are variable because they are dependent on the number of model variables and A_f does not consider that predictions can err more in the fail-safe direction. This discrepancy is important because B_f and A_f are supposed to work together to provide a complete evaluation of model performance. A proposed correction is to use the same criteria for acceptable B_f and A_f. In other words, if a model has a fail-dangerous B_f of >1.000, then A_f should be <1.150, whereas if a model has a fail-safe B_f of $\leq$ 1.000, then A_f should be <1.300. When this correction was applied, 3 of 18 evaluations were unacceptable for B_f and A_f compared with 6 of 18 unacceptable evaluations for pRE (Table 2). Thus, even after correction for this limitation of B_f and A_f, pRE was a more sensitive and reliable indicator of model performance

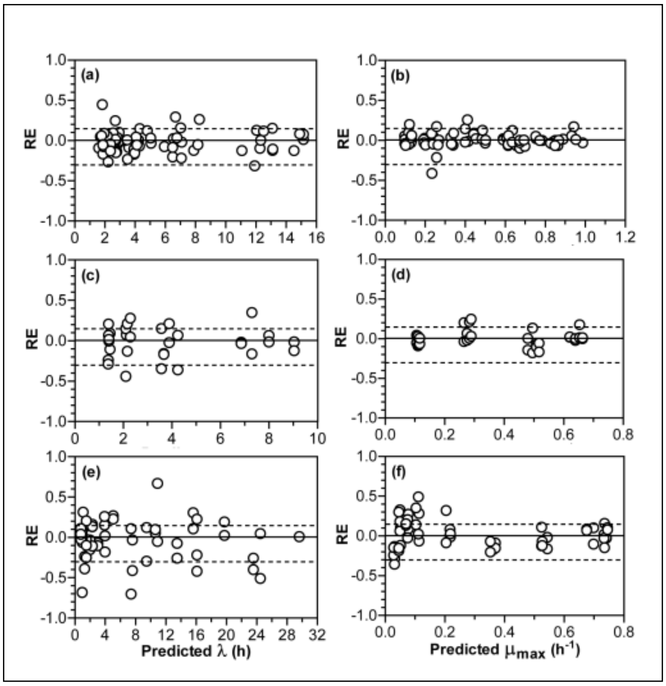

Figure 2—Relative error (RE) plots with an acceptable prediction zone from an RE of –0.3 (fail-safe) to 0.15 (fail-dangerous) for comparison of observed and predicted values of lag time (λ) and maximum specific growth rate (μ_{max}) of *Salmonella* Typhimurium ATCC 14028 for data used in verification of (a) Model 1, (b) Model 2, (c) Model 3, (d) Model 4, (e) Model 5, and (f) Model 6.

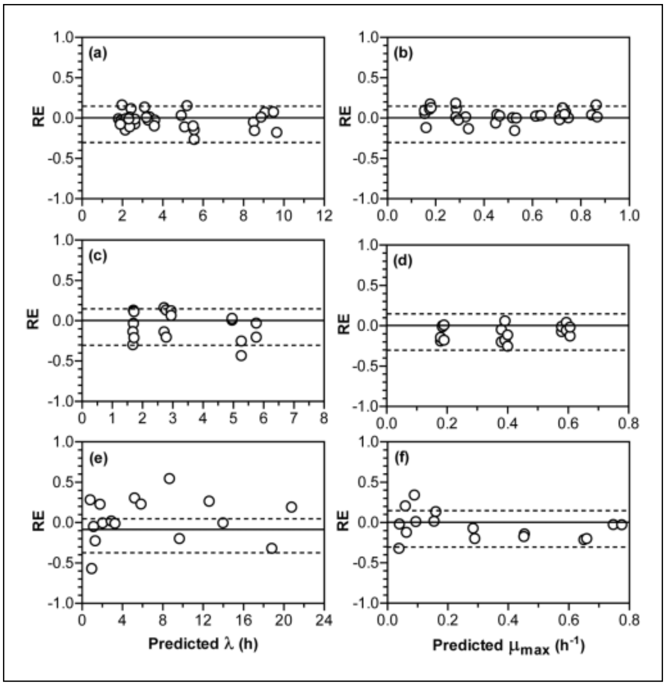

Figure 3—Relative error (RE) plots with an acceptable prediction zone from an RE of –0.3 (fail-safe) to 0.15 (fail-dangerous) for comparison of observed and predicted values of lag time (λ) and maximum specific growth rate (μ_{max}) of *Salmonella* Typhimurium ATCC 14028 for data used in evaluation of interpolation for (a) Model 1, (b) Model 2, (c) Model 3, (d) Model 4, (e) Model 5, and (f) Model 6.

than B_f and A_f. This occurred because of the other limitations of B_f and A_f that are discussed next.

Another limitation of B_f and A_f as performance factors is the inability to detect some forms of prediction bias such as under prediction in 1 region of the response surface and over prediction in another region of the response surface (Ross and others 2000). For example, broth Model 1 had an acceptable B_f of 0.981 and an acceptable A_f of 1.225 but an unacceptable pRE of 0.583 (Table 2). When the RE plot was examined, it was found that broth Model 1 provided overly fail-dangerous predictions at short λ (<4 h) and slightly fail-safe but not overly fail-safe predictions at longer λ (Figure 4a). Thus, pRE was able to detect a performance problem for extrapolation of broth Model 1 that was the result of systematic prediction bias and was not detected by B_f and A_f. It should be noted that in contrast to other methods for evaluating systematic prediction bias (for example, normal distribution of residuals around zero and the runs test), a defined amount of systematic prediction bias is acceptable in the methods used here. In other words, as long as the systematic prediction bias resides mostly within the acceptable prediction zone or within the acceptable range for B_f it is acceptable, as was the case for extrapolation of broth Model 2 to sBM and sTM (Figure 4b).

Another limitation of B_f and A_f, which are ratios of observed and predicted values, is that they cannot be calculated for prediction cases where no growth is predicted by the model and growth is observed or no growth is observed and the model predicts growth (Dalgaard and Jorgensen 1998). To illustrate this point, the acceptable prediction zone method was applied to data used to develop models for aerobic growth of *Escherichia coli* O157:H7 in broth (Buchanan and others 1993). This dataset contained 25 prediction cases where no growth was observed but the models predicted growth. An acceptable B_f of 1.087 and an acceptable A_f of 1.428 were obtained for the λ model, which had 3 variables and an expected A_f of <1.450 (Figure 5a). Likewise, an acceptable B_f of 1.000 and an acceptable A_f of 1.298 were obtained for the τ model (Figure 5b), which also had 3 independent variables and an expected A_f of <1.450. However, in this situation, B_f and A_f overestimate the performance of these models because the 25 no growth prediction cases were excluded from the calculation of B_f and A_f. In contrast, pRE provides an accurate assessment of model performance because pRE considers no growth prediction cases in its calculation of model performance. In fact, an unacceptable pRE of 0.265 was obtained for the λ model (Figure 5a), and an unacceptable pRE of 0.422 was obtained for the τ model (Figure 5b). Thus, in contrast to B_f and A_f, pRE indicated that the models for aerobic growth of *E. coli* O157:H7 in broth do not provide acceptable predictions of the data used to develop them. In other words, the models failed the evaluation for verification, which is a prerequisite for validation of the models (Figure 1). Failure of the *E. coli* O157:H7 growth models resulted mainly from the models not predicting well near the growth/no growth interface as there were 25 no growth prediction cases where the models predicted growth but growth did not occur. Thus, it may be possible to repair these models by adding data in the regions of the growth/no growth interface.

A limitation of pRE is that it is unable to distinguish between models with global (for example, Model 5 for interpolation in Figure 3e) and regional (for example, Model 1 for extrapolation in Figure 4a) performance problems. However, use of pRE and an RE plot with an acceptable prediction zone was found to provide a reliable and complete evaluation of model performance. In particular, this combination was effective at identifying specific regions in the response surface where predictions were overly fail-safe or overly fail-dangerous. Together, pRE and the RE plot form the acceptable prediction zone method, a new method for evaluating the performance of predictive models that overcomes the limitations of B_f and A_f.

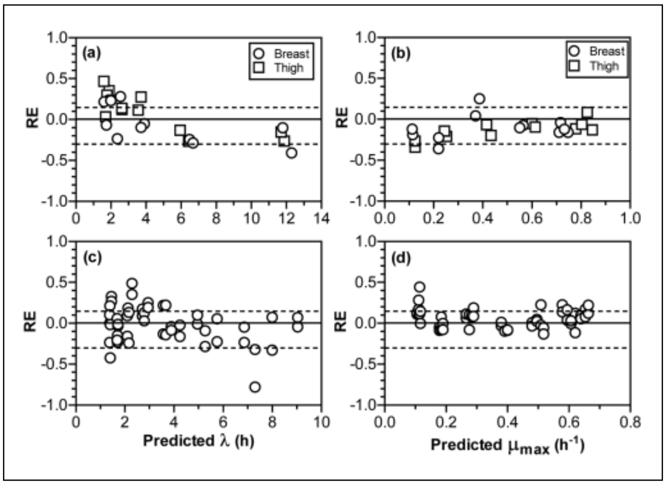

Figure 4—Relative error (RE) plots with an acceptable prediction zone from an RE of –0.3 (fail-safe) to 0.15 (fail-dangerous) for comparison of observed and predicted values of lag time (λ) and maximum specific growth rate (μ_{max}) of *Salmonella* Typhimurium ATCC 14028 for data used in evaluation of extrapolation for (a) Model 1, (b) Model 2, (c) Model 3, and (d) Model 4.

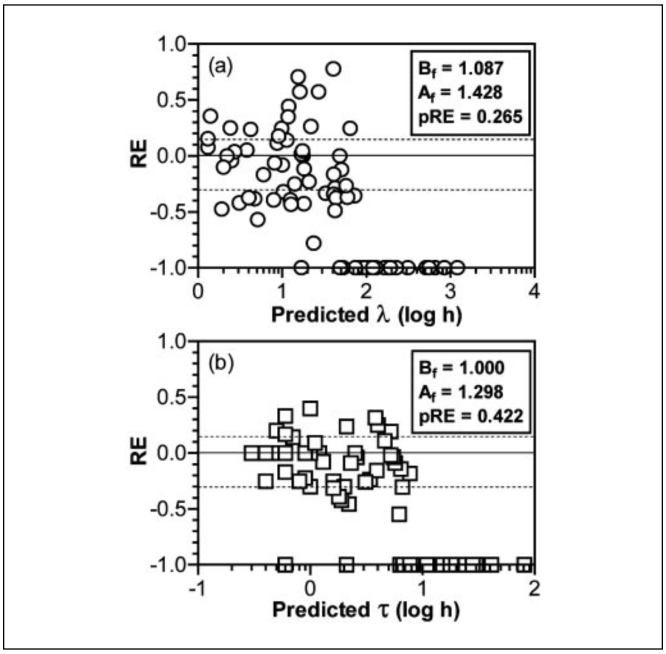

Figure 5—Relative error (RE) plots with an acceptable prediction zone from an RE of –0.3 (fail-safe) to 0.15 (fail-dangerous) for comparison of observed and predicted values of (a) lag time (λ) and (b) generation time (t) of published data for *Escherichia coli* O157:H7 (Buchanan and others 1993). A_f = accuracy factor; B_f = bias factor; pRE = proportion of RE in the acceptable prediction zone. By default, no growth prediction cases fell outside the acceptable prediction zone and were assigned a value of –1 for graphical presentation. Here, the no growth prediction cases (*n* = 25) resulted from the models predicting growth when no growth occurred.

Variation of growth kinetics among strains of a pathogen is an important factor to consider when evaluating the ability of a model to extrapolate to an independent set of data obtained with a different strain than the 1 used to develop the model (Lebert and others 1998). The growth kinetics of *S.* Typhimurium ATCC 14028, which was used to develop the models evaluated in this study, were previously compared with growth kinetics of other strains of *Salmonella* (Oscar 1998). At 40 °C in BHI broth, λ of *S.* Typhimurium ATCC 14028 was shorter than λ of 5 other strains and not different from λ of 10 other strains, whereas μ_{max} was less than μ_{max} of 6 other strains and not different from μ_{max} of 9 other strains tested (Oscar 1998). In general, variation of growth kinetics among strains of a pathogen is greater under nonoptimal growth conditions (Begot and others 1997). For example, the coefficient of variation for τ among 45 strains of *Salmonella* Enteritidis increases from 4% at the optimal growth temperature of 37 °C to 22% at the suboptimal growth temperature of 9 °C (Fehlhaber and Kruger 1998).

In the present study, all sets of data used in evaluation of model performance were obtained using the same strain as that used in model development so as not to confound the comparison of observed and predicted values. This was important because had a faster growing strain such as *Salmonella* Simsbury, with a μ_{max} of 0.88/h at 40 °C versus 0.78/h for *S.* Typhimurium ATCC 14028 (Oscar 1998), been used to collect test data set for extrapolation broth Model 2 to sBM, B_f for this comparison would have been very close to 1.000 rather than 0.884. Thus, it would have been falsely concluded that the broth model provided unbiased predictions of *Salmonella* growth on sBM when in fact broth Model 2 on average overpredicted μ_{max} by 12%. Of note, Baranyi and others (1999) also observed that broth models tend to overestimate the growth rate of pathogens on food.

Previous growth conditions also affect microbial growth kinetics and if not controlled when collecting test data, could result in a confounded evaluation of model performance. The most well-documented effect of previous growth conditions on microbial growth kinetics is the effect of shifts in temperature on λ. Mellefont and Ross (2003) reported that relative λ increased in a nonlinear manner when *E. coli* SB1 grown at a previous temperature of 44.4 °C were downshifted to growth temperatures from 44.4 °C to 10 °C. Buchanan and Klawitter (1991) found that λ of *Listeria monocytogenes* at 5 °C increases from 37 to 50 h when previous growth temperature increases from low (5 °C to 28 °C) to high (37 °C to 42 °C) temperatures. Likewise, increasing previous growth temperature from 5 °C to 35 °C increases λ at 5 °C from 12 to 354 h or from 4.5 to 18 h, depending on the strain of *Aeromonas hydrophila* tested (Hudson 1993). In contrast, upshifts and downshifts from previous growth temperatures of 16 °C to 34 °C to subsequent growth temperatures of 16 °C to 34 °C in challenge studies with sBM and *S.* Typhimurium ATCC 14028 did not alter λ (Oscar 1999b). In all the aforementioned studies, μ_{max} was not affected by previous growth conditions. Consequently, in the current study, data used in evaluating the performance of models were collected using the same previous growth conditions so as not to confound the comparison of predicted and observed values for λ.

The primary model used to fit the growth kinetic data also affects the values for λ and μ_{max} and thus, the comparison of observed and predicted values. Buchanan and others (1997) reported that the 3-phase linear model provided shorter λ and higher μ_{max} than the Gompertz and Baranyi models when fit to the same growth data for *E. coli* O157:H7 in broth. The differences in λ and μ_{max} among the 3-phase linear, Gompertz and Baranyi models occur because the models make different assumptions and thus, provide different estimates of the growth parameters (Buchanan and others 1997). For example, μ_{max} in the 3-phase linear model is the slope of the line through the assumed linear exponential growth phase, whereas μ_{max} is the slope of the line that is tangent to the inflection point in the Gompertz model, which does not assume constant growth rate throughout the exponential phase. Thus, although all 3 models fit the same growth data well, they provide slightly different values for λ and μ_{max} (Buchanan and others 1997). Consequently, it is important to compare only λ and μ_{max} obtained with the same primary growth model when evaluating the performance of a model or the performance evaluation will be confounded. In the present study, the 2-phase linear primary model was used to fit all growth curves used in model development and performance evaluation so as not to confound the comparison of observed and predicted values of λ and μ_{max}.

The decision diagram in Figure 1 and the acceptable prediction zone method were used to determine whether the models evaluated provided valid predictions of *S.* Typhimurium growth. All model evaluations met the criteria for test data as indicated by a "yes" response to questions 4 and 7 in Figure 1 (Table 3). Thus, no models failed validation for this reason. However, some models had unacceptable performance (that is, pRE <0.700) as indicated by a "no" response to questions 2 or 8 in Figure 1 (Table 3). Specifically, Models 3 and 5 failed verification and by default failed validation, whereas Model 1 failed validation for extrapolation to sBM and sTM. Models 1, 2, 4, and 6 were validated for interpolation, and Models 2 and 4 were validated for extrapolation. Interestingly, Model 3 failed verification but had an acceptable pRE for interpolation. Thus, it may be possible to "repair" this model by adding the test data for interpolation to the data used for model development and refitting the secondary model. Overall, the evaluation in Table 3 indicated that models can be developed and validated using the criteria established in the current study. Thus, the criteria are not overly restrictive.

The use of the terms verification and validation in Figure 1 is controversial because in predictive microbiology, these terms are used as synonyms, whereas in Figure 1 and other fields of science they are not. More specifically, verification in Figure 1 is the successful outcome of the performance evaluation process where the model predictions were compared with the data used in model development (that is, dependent data). In contrast, validation in Figure 1 is the successful outcome of the performance evaluation process where model predictions were compared with data that was not used in model development (that is, independent data). Although use of the terms verification and validation in the current study may be at odds with their current usage in predictive microbiology, separate usage of these terms has the advantage of providing an easy and needed distinction between the 2 types of evaluation processes, that is, 1 with dependent data and 1 with independent data. Furthermore, the use of the terms here is consistent with their usage in other scientific disciplines.

Conclusions

Even when proper methods are used for collecting test data to evaluate model performance and acceptable values of B_f and A_f are obtained, this is not sufficient for validation of predictive models because B_f and A_f have important limitations that can result in inaccurate assessments of model performance and improper validation of models. In contrast, the acceptable prediction zone method and decision diagram for validation developed here provide a complete evaluation of model performance that overcomes limitations of B_f and A_f and provides an accurate assessment of model performance and validation of models even in situations of systematic prediction bias and no growth prediction cases.

Acknowledgments

The author thanks Jaci Ludwig and Pat Shannon of the Agricultural Research Service and Keith Peacher, Robert Sudler, Tiana Richardson, Paul Jackson, Nisha Oatman, and Jamil Johnson of the Univ. of Maryland Eastern Shore for their assistance on this project. Mention of brand or firm name does not constitute an endorsement by the U.S. Dept. of Agriculture over others of a similar nature not mentioned.

References

Augustin JC, Carlier V. 2000. Mathematical modeling of the growth rate and lag time for *Listeria monocytogenes*. Int J Food Microbiol 56:29–51.

Baranyi J, Pin C, Ross T. 1999. Validating and comparing predictive models. Int J Food Microbiol 48:159–66.

Begot C, Lebert I, Lebert A. 1997. Variability of the response of 66 *Listeria monocytogenes* and *Listeria innocua* strains to different growth conditions. Food Microbiol 14:403–12.

Bratchell N, McClure NJ, Kelly TM, Roberts TA. 1990. Predicting microbial growth: graphical methods for comparing models. Int J Food Microbiol 11:279–88.

Buchanan RL. 1991. Using spreadsheet software for predictive microbiology applications. J Food Saf 11:123–34.

Buchanan RL, Bagi LK, Goins RV, Phillips JG. 1993. Response surface models for the growth kinetics of *Escherichia coli* O157:H7. Food Microbiol 10:303–15.

Buchanan RL, Klawitter LA. 1991. Effect of temperature history on the growth of *Listeria monocytogenes* Scott A at refrigeration temperature. Int J Food Microbiol 12:235–46.

Buchanan RL, Phillips JG. 1990. Response surface model for predicting the effects of temperature, pH, sodium chloride content, sodium nitrite concentration and atmosphere on the growth of *Listeria monocytogenes*. J Food Prot 53:370–6.

Buchanan RL, Whiting RC, Damert WC. 1997. When is simple good enough: a comparison of the Gompertz, Baranyi and three-phase linear models for fitting bacterial growth curves. Food Microbiol 14:313–26.

Dalgaard P, Jorgensen LV. 1998. Predicted and observed growth of *Listeria monocytogenes* in seafood challenge tests and in naturally contaminated cold-smoked salmon. Int J Food Microbiol 40:105–15.

Delignette-Muller ML, Rosso L, Flandrois JP. 1995. Accuracy of microbial growth predictions with square root and polynomial models. Int J Food Microbiol 27:139–46.

Fehlhaber F, Kruger G. 1998. The study of *Salmonella enteritidis* growth kinetics using rapid automated bacterial impedance technique. J Appl Microbiol 84:945–9.

Gibson AM, Bratchell N, Roberts TA. 1988. Predicting microbial growth: growth responses of salmonellae in a laboratory medium as affected by pH, sodium chloride and storage temperature. Int J Food Microbiol 6:155–78.

Hudson JA. 1993. Effect of pre-incubation temperature on the lag time of *Aeromonas hydrophila*. Lett Appl Microbiol 16:274–6.

Lebert I, Begot C, Lebert A. 1998. Development of two *Listeria monocytogenes* growth models in a meat broth and their application to beef meat. Food Microbiol 15:499–509.

McClure PJ, Baranyi J, Boogard E, Kelly TM, Roberts TA. 1993. A predictive model for the combined effect of pH, sodium chloride and storage temperature on the growth of *Brochothrix thermosphacta*. Int J Food Microbiol 19:161–78.

McClure PJ, Blackburn CW, Cole MB, Curtis PS, Jones JE, Legan JD, Ogden ID, Peck MW, Roberts TA, Sutherland JP, Walker SJ. 1994. Modelling the growth, survival and death of microorganisms in foods: the UK Food Micromodel approach. Int J Food Microbiol 23:265–75.

Mellefont LA, Ross T. 2003. The effect of abrupt shifts in temperature on the lag phase duration of *Escherichia coli* and *Klebseilla oxytoca*. Int J Food Microbiol 83:295–305.

Oscar TP. 1998. Growth kinetics of *Salmonella* isolates in a laboratory medium as affected by isolate and holding temperature. J Food Prot 61:964–8.

Oscar TP. 1999a. Response surface models for effects of temperature and previous growth sodium chloride on growth kinetics of *Salmonella* Typhimurium on cooked chicken breast. J Food Prot 62:1470–4.

Oscar TP. 1999b. Response surface models for effects of temperature and previous temperature on lag time and specific growth rate of *Salmonella* Typhimurium on cooked ground chicken breast. J Food Prot 62:1111–4.

Oscar TP. 1999c. Response surface models for effects of temperature, pH, and previous growth pH on growth kinetics of *Salmonella* Typhimurium in brain heart infusion broth. J Food Prot 62:106–11.

Oscar TP. 2000. Variation of lag time and specific growth rate among 11 strains of *Salmonella* inoculated onto sterile ground chicken breast burgers and incubated at 25C. J Food Saf 20:225–36.

Oscar TP. 2004. Simulation model for enumeration of *Salmonella* on chicken as a function of PCR detection time score and sample size: implications for risk assessment. J Food Prot 67:1201–8.

Ross T. 1996. Indices for performance evaluation of predictive models in food microbiology. J Appl Bacteriol 81:501–8.

Ross T, Dalgaard P, Tienungoon S. 2000. Predictive modeling of the growth and survival of *Listeria* in fishery products. Int J Food Microbiol 62:231–45.

Surkiewicz BF, Johnston RW, Moran AB, Krumm GW. 1969. A bacteriological survey of chicken eviscerating plants. Food Technol 23:80–5.

te Giffel MC, Zwietering MH. 1999. Validation of predictive models describing the growth of *Listeria monocytogenes*. Int J Food Microbiol 46:135–49.

Wei QK, Fang TJ, Chen WC. 2001. Development and validation of growth model for *Yersinia enterocolitica* in cooked chicken meats packaged under various atmosphere packaging and stored at different temperatures. J Food Prot 64:987–93.