Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Food Science

Review

From data to models and predictions in food microbiology

József Baranyi[1], Maha Rockaya[1] and Mariem Ellouze[2]

This paper emphasizes the importance of structured databases, visualization techniques, statistics, and mathematical models as milestones when developing predictive models of bacterial responses to food environments. Predictions generated by such models are vital in decision-making on food safety and quality issues. The paper suggests that while refinements, such as reparameterization, rescaling, and fine-tuning smoothness-characterizing parameters, are useful for numerical/statistical point of view, the result should not be considered as new models. It is proposed that novel predictive models should be linked to those widely accepted in related disciplines, such as biotechnology, systems biology, or biochemistry.

Addresses
[1] University of Debrecen, 138 Böszörményi út, 4032 Debrecen, Hungary
[2] Nestlé Research, 57 Route Du Jorat, 1000 Lausanne, Switzerland

Corresponding author: Baranyi, József (baranyi.jozsef@med.unideb.hu)

Current Opinion in Food Science 2024, **57**:101177

This review comes from a themed issue on **Food Microbiology**

Edited by **Maristela da Silva do Nascimento**

For complete overview of the section, please refer to the article collection, "Food Microbiology 2024"

Available online 17 May 2024

https://doi.org/10.1016/j.cofs.2024.101177

2214–7993/© 2024 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Introduction

Food safety and quality are a top priority for all the stakeholders of the food chain. All have a shared responsibility to assure that food is safe and suitable for consumption [9]. Predictive microbiology has emerged as a powerful tool to ensure food safety and quality in a cost and time effective manner. In this paper, we explore the steps needed to generate predictions via database building, statistics and mathematical modelling (Figure 1).

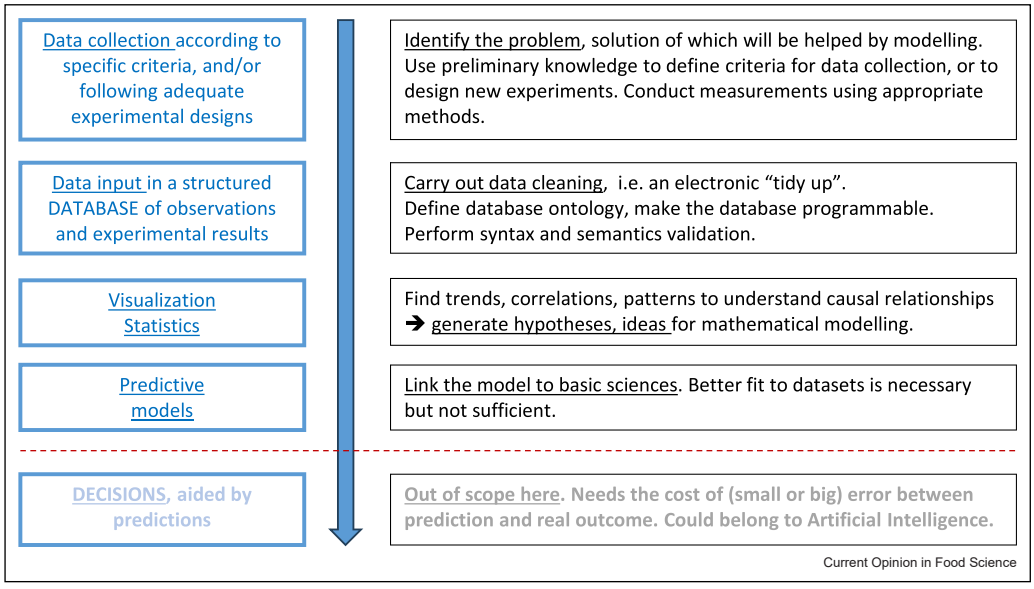
Following the steps above, in what follows, we will clarify some misleading concepts and suggest where progress could be made when developing new models for predictive food microbiology.

Database

Collecting and digitizing data and structuring and archiving them in databases are of paramount importance in various life sciences fields, including microbiology. It requires efficient organization, retrieval, and analysis of data, leading to valuable insights and informed decision-making. Scientific papers have consistently emphasized the significance of database ontology in data management, knowledge accumulation and pattern identification/discovery. Tamplin and Ratkowsky [21] highlighted the power of the ComBase food microbiology database to identify causal relationships between bacterial responses and their environments, then using them to provide recommendations and guidelines for Food Business Operators. The structure of that database has been found useful in related areas, too, where the aim was to record observations on causal effects. For instance, Pacza et al. [14] adapted the ontology of ComBase to provide a template for digitizing experimental results on the molecular composition of human milk as a function of the mother–infant–environment triplet. Farkas et al. [7] analyzed large number of milk samples and used food consumption databases as well as existing literature to fill knowledge gaps and to gain realistic parameters for their Monte-Carlo simulations estimating the exposure of different age groups to aflatoxin; this way revealing elevated risk for toddlers and children. Similarly, Messens et al. [13] analyzed national and European databases to draw conclusions and provide recommendations.

These studies demonstrate the value of structured databases with well-defined syntax and semantics in enabling researchers to extract meaningful information. A well-constructed ontology should be able to formally represent knowledge, relationships, and other properties within a specific domain, thus allowing algorithms to check the syntax and semantics of the records. It should also facilitate integration and interoperability of data across different systems [8]. Ontologies are vital for artificial intelligence (AI), information science and knowledge management to facilitate data sharing, reasoning and semantic understanding.

Figure 1



Milestones when making sense of data.

Visualization and statistics

Visualization and statistics play a crucial role in understanding and interpreting complex information. By means of visualization techniques, patterns, trends and relationships can be revealed from the studied data sets. Statistics, as a complementary tool, provide a robust methodology to analyze the data, identify outliers and potential correlations and highlight significance. For instance, Zongur et al. [23] used descriptive statistical and visualization methods to analyze their data before selecting a model to describe its main features. The authors aimed at modelling the inhibition of various *Fusarium* spp strains in diseased potato tubers by a fermented fruit (gilaburu) extract. They inoculated agar plates with the tested mold and measured the inhibition diameters around spots of the studied compound. They used three variables (extraction type, compound concentration and the type of Fusarium spp strain) and one response (the diameter of the inhibition zone). They applied rescaling and normalization techniques to improve the performance of the tested models and decrease the bias in the comparisons. Synergy and feedback mechanism between the steps detailed by Figure 1 is in fact typical throughout the process of developing predictive models from raw data.

Mathematical models

While visualization and statistics are key to explore data sets, mathematical models provide a deeper understanding of the mechanism governing the system studied. Unlike statistical models that focus on empirical relationships and correlations, mechanistic models aim to describe the causal relationships and dynamics of the system by means of (possibly differential) equations. This enables researchers to predict the behaviour of the micro-organisms under different conditions throughout the value chain or as a function of specific processing steps (e.g. thermal treatment, fermentation).

Mathematical models are about the common features of different phenomena, and these common features, described in the language of mathematics, form an ideal world in our 'cognitive space'. Predictions are inferences from current observations, and the means to generate them is mathematical modelling. To this end, it is rather unfortunate that 'predictive modelling' was named like this in the 80s, as it is rather trivial that we do modelling to be able to predict events.

Predictive microbiology models are commonly divided into primary and secondary models, the first focusing on the bacterial responses as a function of time, in a constant environment, and the second describing the parameters of the primary model as a function of some environmental factors.

In the early years of the subject, researchers used multivariate functions to describe the combined effects of time and environmental factors, and only later they recognized that temporal variation should not be treated in the same way as the variation with other factors. Indeed, time is a special explanatory variable; it is irreversible

and cannot be affected by other factors. (Note that the 'time to a certain event' can, of course, be affected by other factors, but this is an interval, not the same as the concept of time flowing independently of anything else.) For the above reasons, primary models have been separated and normally never merged with secondary models.

Primary models

Primary models can describe both growth and survival of bacteria (see Ref. [11]). The 'ideal' response variable is the bacterial concentration, estimable by plating; that is counting Colony Forming Units of a sample, on agar plates. It can also be measured by some physical projection of the bacterial concentration, like optical density, in certain conditions (e.g. transparent liquid media). Lately, qPCR (Quantitative Polymerase Chain Reaction) techniques have been used for such estimates. A fundamental problem of the latter two techniques is that they do not detect VIABLE cells only; therefore, they are unsuitable to measure survival curves.

Here, we must mention the concept of *dynamic model*. It does not simply refer to the fact that one or more response variables (such as the bacterial concentration) change with time. Dynamic models are about how the *rates* of temporal (time-dependent) variables change as a function of other (possibly also time-dependent) temporal variables. Such models are described by differential equations expressing how the instantaneous *change* of a system at the time t depends on the *state* of the system in that moment. Since Newton, virtually all great temporal models of natural science have been described by differential equations. To some extent, we can say that the use of differential equations indicates the intention to provide a mechanistic, causal understanding of the system (which does not mean that a model becomes mechanistic just because it is described by differential equations).

Primary models should be dynamic models, and it is unfortunate that most publications only quote their algebraic solutions (valid only at constant environment), just because those forms are easier for curve fitting. This can be confusing when combining primary and secondary models to generate predictions for situations when the environment changes with time. In fact, for this case, the only valid approach is based on differential equations. An iterative algorithm may be able to simulate the process; but, finally, it becomes equivalent to the numerical (approximate) solution of a differential equation, as happened in Haque et al. [10]. As that paper also demonstrates, it is especially difficult to model the lag time and the 'time to reach a certain level' without the concept of differential equations because it would need another rate, the 'consumption rate' of that time interval.

Secondary models

As mentioned, secondary models describe how the parameters of the primary model vary as a function of environmental factors. A recent summary of primary and secondary models, and their properties, can be found in Dantigny [6], including a new model developed by the author.

The most quoted examples for secondary models describe the variation of the maximum specific growth rate of an organism as a function of temperature.

While primary models for growth are typically sigmoid curves, the shape of such secondary model is like that of an asymmetric triangle or trapezoid. The two most frequently used secondary models, describing how the maximum specific growth rate depends on temperature, are that of Ratkowsky et al. [18] and of Rosso and Robinson [20].

The model of Ratkowsky et al. [18] states that in the suboptimal region of temperature, the square root of the maximum specific growth rate of an organism linearly depends on the temperature. Linearity is a holy grail for modellers; our numerical and statistical arsenal is much more exact for linear than for nonlinear cases, which made this model popular. An extension for it maintained the (at least close-to) linear shape for suboptimal temperatures but transformed it to the typical triangle shape for the entire growth range of the temperature. The controversial issue here is the claim that the variance of the square-root-transformed specific rate remained constant in the entire temperature range (Figure 2). Intuitively, close to the boundary of growth, any uncertainty measure should be greater than in the 'happy growth' region, far from the possibility of no growth.

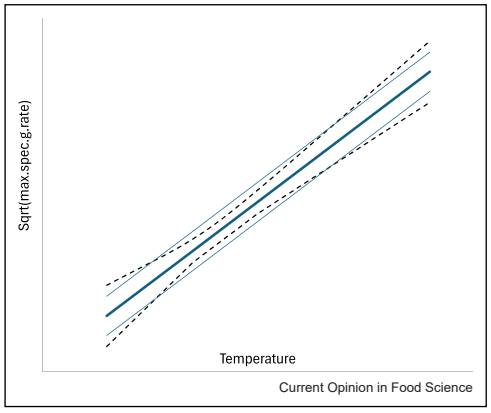
The model of Rosso and Robinson [20] does not have linear origins, and the authors did not investigate what link function would be best for regression. However, they included a curvature parameter n in their model, which controls the smoothness of the transition from no growth to grow domain around the minimum temperature, T_{min} (Figure 3).

In the last section, we make some recommendations on such model extensions, especially on the use of link functions and curvature parameters.

Tertiary models

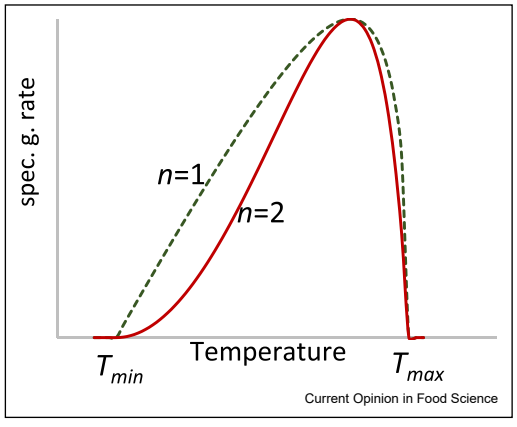
Logically, tertiary models should describe how the parameters of a secondary model depend on explanatory variables other than those characterizing the environment; for example, categories such as species/strain or food [3]. It is unfortunate that the name 'tertiary models' is still used for the simple *combination* of primary and secondary models [5]. One of the early examples for tertiary modelling was Rosso et al. [19] who found strong correlations between the respective cardinal values of many species. This is not a

Figure 2



The secondary model of Ratkowsky et al. [18] (thick continuous line) is linear for the square root of the maximum specific growth rate at sub-optimal temperatures. It is questionable whether the measurement errors generate a constant confidence band around it (thin continuous lines). It is more likely that, as the temperature decreases, the width of this band decreases together with the rate, then closer to the minimum temperature, it becomes wider again,due to the increasing biological variability (broken lines).

Figure 3



Effect of the tuning (curvature) parameter n on the Cardinal Values Model of Rosso and Robinson [20]. With $n = 1$, the transition between the 'growth/no growth' domain is sudden; with $n = 2$, it is smooth (differentiable).

correlation a typical regression procedure would show between the estimates of two parameters. The latter is due to the noise of the observations, not due to the variation of the biological species in question. Such clarifications are still necessary to improve the mathematical rigour of predictive microbiology.

Recommendations for model developments

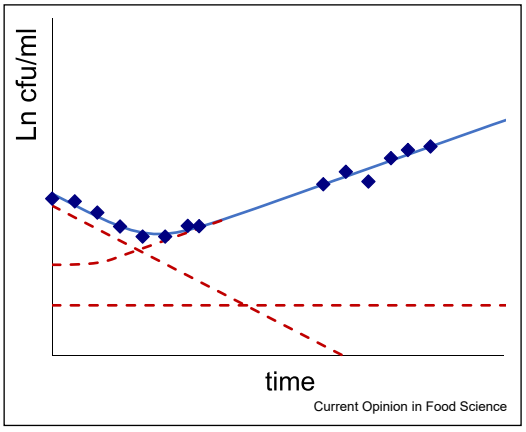
History effect for primary models

The commonly used lag model of Baranyi and Roberts [2] is an example for dynamic modelling. In fact, it is a refinement/extension of the fundamental exponential growth, with $\mu = q/(1+q)\cdot\mu_{max}$ where the μ_{max} parameter is the maximum exponential rate that the organism can reach in the studied environment. The $q(t)$ temporal variable quantifies how far the process is in the adjustment period, which is a transition from the history to the current environment. A real mechanistic result would be to find what biochemical process $q(t)$ could represent, which would probably require different measurement methods compared with the classical and easy log count methodology used today. With the new developments in imaging techniques, this might become possible in the years to come.

Another potential development (that would be a milestone in food microbiology modelling) is this: predict, from the history of the cells, what proportion of the inoculum will be able to produce exponentially growing progeny (a subpopulation originating from a single initial cell) in the new environment or die in response to an intense stress. Namely, some initial cells, especially in stress environment, may not divide but die or become *persisters* [1]. The result for the log cell concentration could be a nonsigmoid curve as shown by Figure 4, with the following question: what proportion of the initial cells will be the ancestors of those that reach the stationary phase? In the same way, what proportion just survives; not dividing and not dying either. These cells play a crucial role in stress resistance modelling, as the fast-growing cell population may be more vulnerable, while the persistent cells may become more resistant to other stresses, too. The population could have been still homogeneous at the inoculation, while the subpopulations generated by the initial individual cells follow one of the three trajectories of Figure 2 randomly, at α_0, α_1, $(1-\alpha_0-\alpha_1)$ probabilities, respectively. This distribution of the three fractions is a good candidate for primary model parameters depending both on the current environment and the history of the cells. Such questions have also been discussed by Paganini et al. [15].

This leads us to an important response variable: the probability of growth for a single cell. By this, we mean the probability that an inoculated single cell can produce an exponentially growing progeny. Such stochastic models belong to primary modelling. In our view, it would be more important to find connections between primary models at population level and 'probability of growth' models at single cell level instead of refining the structures of existing primary models or inventing newer ones. For example, it is obvious that the probability of division for a single cell correlates with the specific growth rate of the population; just think about the rearrangement of the equation $dx/dt = \mu\cdot x$ for $dx/x = \mu\cdot dt$, where $x(t)$ is the concentration of a cell population growing at the μ specific growth rate. The term dx/x can be conceived as the probability of division of a single cell in a small dt interval. Especially around the growth/no

Figure 4



Nonsigmoid primary model. Diamonds: Ln-cell concentration of *Salmonella spp.* under osmotic stress [22]. Thick line: fitted model assuming a dying, a nongrowing and a growing subpopulation; the curves of these latter three subpopulations are represented by broken lines. Their proportion in the inoculum depends on the history of the cells as well as on the current environment.

growth boundary of the environmental factor(s) on which μ depends, this could be utilized to develop generic or linked models for the two responses.

Reparameterization, rescaling, and tuning parameters

Reparameterization of a model is a transformation of the parameters, normally for the sake of better statistical features, such as accurately estimating the standard error of the parameters when fitting the model to data, or for lower correlation between the error of the estimates, or simply for easier interpretation of the new parameter set, that has significance also when providing initial estimates for them during fitting. Reparameterization does not change the model, but it can significantly change its regression, easily transforming it from linear to nonlinear and vice versa. A recent example of this technique is that of Boonruang and Lerkkasemsan [4]. A reparameterized version of a model certainly should not be called a new model; even the 'modified model' is an exaggeration (like the 'modified Gompertz model', which expression is still in use; see Petruzzi et al. [16]).

Rescaling is applying a monotone function for one of the variables, generally to make its effect closer to linear. An example for this is the Arrhenius rescaling of temperature [12]. If the rescaling is on the response variable, it is called the 'link function', as we referred to it when discussing the model of Ratkowsky et al. [18].

Conclusions

In this paper, we highlighted the importance of structured databases, visualization methods, statistics, and mathematical models to improve the safety and quality of food products. We pointed out that the combination of visual representation, statistical analysis and mathematical modelling enhances data-driven decision-making and supports evidence-based research. Refinements such as reparameterization, rescaling, and fine-tuning of parameters should be done for the easier application/implementation of the models, justified by statistical reasoning, and should not be considered as new models.

Novel predictive microbiology models could be inspired from accepted models of other disciplines such as biotechnology, systems biology, and biochemistry. Examples for this is the mentioned model of Baranyi and Roberts [2], whose starting point was the exponential growth, modulated by logistic limitation due to the carrying capacity of the environment and by inhibition during the lag phase inspired by Michaelis-Menten kinetics. The model of Ratkowsky et al. [18] was inspired by an old biochemical model (and the author himself referred to it as Belehrádek-type model).

Another possibility for development could be to leverage the already available databases and develop AI models considering different responses. For example, machine learning techniques have been applied for anomaly detection, defect identification, and quality assessment in foods using sensor-based devices. Those data could be coupled to microbiological responses to identify through adequate correlations, early warning indicators that microbiological growth to unacceptable levels might be observed. Agent-based models have also been utilized for environmental monitoring programmes, for simulating sampling strategies and for supporting proposals for corrective actions. This could be used to answer one of the most asked questions in predictive microbiology and get to know the initial microbial load when cross-contamination occurs from the production environment. These examples could increase AI adoption in the field of food safety. Indeed, while AI has been widely adopted in areas such as marketing and agricultural production, its use in food safety applications remains relatively low mainly because of the limited availability and sharing of microbial data, concerns about data privacy and business risks, and the lack of a clear legal and regulatory framework [17].

Data Availability

Data will be made available on request.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest.

1. • Arvaniti M, Skandamis PN: **Defining bacterial heterogeneity and dormancy with the parallel use of single-cell and population level approaches**. *Curr Opin Food Sci* 2022, **44**:100808.

A possible direction for modelling: develop 'statistical physics' of predictive microbiology.

2. Baranyi J, Roberts TA: **A dynamic approach to predicting bacterial growth in food**. *Int J Food Microbiol* 1994, **23**:277-294.

3. Baranyi J, Buss da Silva N, Ellouze M: **Rethinking tertiary models: relationships between growth parameters of *Bacillus cereus* strains**. *Front Microbiol* 2017, **8**:293731.

4. Boonruang P, Lerkkasemsan N: **Re-parameterization of the asymmetric model for fungal spore germination**. *Int J Food Microbiol* 2023, **384**:109974.

5. Chen Q, Zhao Z, Wang X, Xiong K, Shi C: **Microbiological predictive modeling and risk analysis based on the one-step kinetic integrated Wiener process**. *Innov Food Sci Emerg Technol* 2022, **75**:102912.

6. Dantigny P: **Applications of predictive modeling techniques to fungal growth in foods**. *Curr Opin Food Sci* 2021, **38**:86-90.

7. •• Farkas Z, Kerekes K, Ambrus Á, Süth M, Peles F, Pusztahelyi T, Jóźwiak ÁB: **Probabilistic modeling and risk characterization of the chronic aflatoxin M1 exposure of Hungarian consumers**. *Front Microbiol* 2022, **13**:1000688.

A comprehensive tour-de-force from databases through statistics and modelling to recommendations.

8. Filter M, Nauta M, Pires SM, Guillier L, Buschhardt T: **Towards efficient use of data, models and tools in food microbiology**. *Curr Opin Food Sci* 2022, **46**:100834.

9. Gerardi A: **Global Food Safety Initiative (GFSI): underpinning the safety of the global food chain, facilitating regulatory compliance, trade, and consumer trust**. Present Knowledge in Food Safety. Academic Press; 2023:1089-1098.

10. Haque M, Wang B, Mvuyekure AL, Chaves BD: **Validation of competition and dynamic models for Shiga toxin-producing *Escherichia coli* (STEC) growth in raw ground pork during temperature abuse**. *Food Microbiol* 2024, **117**:104400.

11. • Koutsoumanis KP, Lianou A, Gougouli M: **Latest developments in foodborne pathogens modeling**. *Curr Opin Food Sci* 2016, **8**:89-98.

A nice summary of predictive models.

12. Koutsoumanis KP, Taoukis PS, Drosinos EH, Nychas GJE: **Applicability of an Arrhenius model for the combined effect of temperature and CO_2 packaging on the spoilage microflora of fish**. *Appl Environ Microbiol* 2000, **66**:3528-3534.

13. •• Messens W, Bover-Cid S, Hempen M, Lindqvist R, Nauta M, Skandamis PN, Koutsoumanis K: **Use of risk assessment and predictive microbiology in regulatory science related to the scientific opinions of the EFSA BIOHAZ Panel**. *Int J Food Microbiol* 2023, **403**:110302.

An insightful report on how regulatory agencies use predictive models.

14. Pacza T, Martins ML, Rockaya M, Müller K, Chatterjee A, Barabási AL, Baranyi J: **MilkyBase, a database of human milk composition as a function of maternal-, infant- and measurement conditions**. *Sci Data* 2022, **9**:557.

15. Paganini CC, Longhi DA, de Aragão GMF, Carciofi BAM: **Modelling the inactivation, survival and growth of *Salmonella enterica* under osmotic stress considering inoculum phase and serotype**. *J Appl Microbiol* 2022, **132**:3973-3986.

16. Petruzzi L, Campaniello D, Corbo MR, Speranza B, Altieri C, Sinigaglia M, Bevilacqua A: **Wine microbiology and predictive microbiology: a short overview on application, and perspectives**. *Microorganisms* 2022, **10**:421.

17. •• Qian C, Murphy SI, Orsi RH, Wiedmann M: **How can AI help improve food safety?** *Annu Rev Food Sci Technol* 2023, **14**:517-538.

A glimpse into the future.

18. Ratkowsky DA, Lowry RK, McMeekin TA, Stokes AN, Chandler R: **Model for bacterial culture growth rate throughout the entire biokinetic temperature range**. *J Bacteriol* 1983, **154**:1222-1226.

19. Rosso L, Lobry JR, Flandrois JP: **An unexpected correlation between cardinal temperatures of microbial growth highlighted by a new model**. *J Theor Biol* 1993, **162**:447-463.

20. Rosso L, Robinson TP: **A cardinal model to describe the effect of water activity on the growth of moulds**. *Int J Food Microbiol* 2001, **63**:265-273.

21. Tamplin ML, Ratkowsky DA: **Pathogen growth when implementing 'Time as a Public Health Control'**. *Food Microbiol* 2021, **96**:103718.

22. Zhou K, George SM, Métris A, Li PL, Baranyi J: **Lag phase of Salmonella enterica under osmotic stress conditions**. *Appl Environ Microbiol* (5) 2011, **77**:1758-1762.

23. Zongur A, Kavuncuoglu H, Kavuncuoglu E, Capar TD, Yalcin H, Buzpinar MA: **Machine learning approach for predicting the antifungal effect of gilaburu (*Viburnum opulus*) fruit extracts on Fusarium spp. isolated from diseased potato tubers**. *J Microbiol Methods* 2022, **192**:106379.