mathematics

MDPI

Article

Bias Due to Averaging the Logistic and SI Models

Dorothy I. Wallace

Department of Mathematics, Dartmouth College, Hanover, NH 03755, USA; dwallace@math.dartmouth.edu

Abstract: Modelers have choices in how they approach a problem, with different approaches potentially leading to different outcomes. Sometimes one approach gives a consistently lower (or higher) result than another. The theorem and corollaries in this study show that if the logistic equation or, equivalently, the SI model, are perturbed at time zero by a range of values with mean zero, the resulting trajectories must average to a value below (for logistic and I) or above (for S) the solution with average initial condition. The proof of the theorem shows that this phenomenon is the result of algebraic properties of the nonlinear quadratic term, although we note it can be extended to a larger class of systems. More importantly it shows that the only necessary criterion is that the perturbations average to zero. The source of them and the properties of their distribution does not matter to the result of the theorem but does affect the magnitude of the proven difference.

Keywords: logistic equation; SI model; average; model comparison; stochastic

MSC: 62J12; 34D10; 34F99

1. Introduction

Modelers of physical and biological phenomena are faced with a choice of approaches; process based systems of ordinary differential equations (ODEs), agent based models, and stochastic models are all possible ways of addressing the same question. Often these different approaches lead to different outcomes, which is not surprising. But sometimes researchers notice that one approach leads to consistently higher or lower output values than another, a phenomenon which is not expected and which could be described as a bias introduced by the modeling process. Simple models for population growth and disease transmission, the logistic and SIR models, exhibit this behavior, as seen in several studies [1–4].

The logistic equation is foundational in mathematical biology, describing growth phenomena in population biology for many organisms as well as tumor growth in mathematical oncology [5–7]. In a recent study of tumor growth, Boucharas et al compared the ability several common growth models to fit a data set of 289 female mice, and concluded that the logistic model did better than the others [8]. Boucharas et al introduced a known quantity of carcinogen into the tissue of the mouse, and searched for the best single set of growth parameters that would fit each mouse tumor separately.

Another common type of cancer study introduces human tumor tissue into an immunodeficient mouse, with the resulting implant called a xenograft [9]. In all cases tumor growth is measured at various times and the methods for doing so are not as precise as one would like [10]. One might hope to improve the data fit by averaging the data over all mice in the study. If, as Boucharas et al suggests, the logistic growth curve is a good representation of a single mouse tumor, then averaging the logistic curves fit to each individual mouse should give a good fit to the averaged data, although that average curve is not itself a solution to the nonlinear logistic equation. Whether or not to average solutions or data over a pool of subjects is a decision every researcher in this field has to make.

The logistic equation is closely related to the commonly used susceptible/infected models in the epidemiology of infectious disease, introduced in 1927 and in common use

check for updates

Citation: Wallace, D.I. Bias Due to Averaging the Logistic and SI Models. *Mathematics* **2023**, *11*, 2321. https://doi.org/10.3390/math11102321

Academic Editors: Juan Francisco Sánchez-Pérez, Gonzalo García Ros and Manuel Conesa

Received: 1 March 2023
Revised: 24 April 2023
Accepted: 10 May 2023
Published: 16 May 2023

Copyright: © 2023 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

today, including, most recently, analysis of the COVID-19 epidemic [11–16]. Parameter estimation, including initial conditions, for this model is of utmost importance in predicting the size and timing of an epidemic, and the model fit to data is determined in part by the modeler's decision to pool data or consider separate locations, resulting in considerations similar to those for xenograft tumor growth [17,18].

This study addresses the difference between these two approaches at a single point of variation, namely initial conditions. One approach is the deterministic logistic equation with given initial condition, f_0, and the other approach is the average of many runs with different initial conditions, f_i, that average to f_0.

A visual example is shown in Figure 1. Figure 1a, shows eleven runs of the logistic equation with different initial conditions, one of which has the average initial condition. Figure 1b compares the run with average initial condition with the average of all eleven runs, which is lower, and the average of the two extreme runs is lower still. Figure 1c shows the difference of the two averages with the solution to the ODE with average initial condition. Figure 1d looks at the susceptible population in a simple susceptible/infectious model, in which the pattern is reversed, with the average of solutions lying above the single solution to the ODE with average initial conditions.

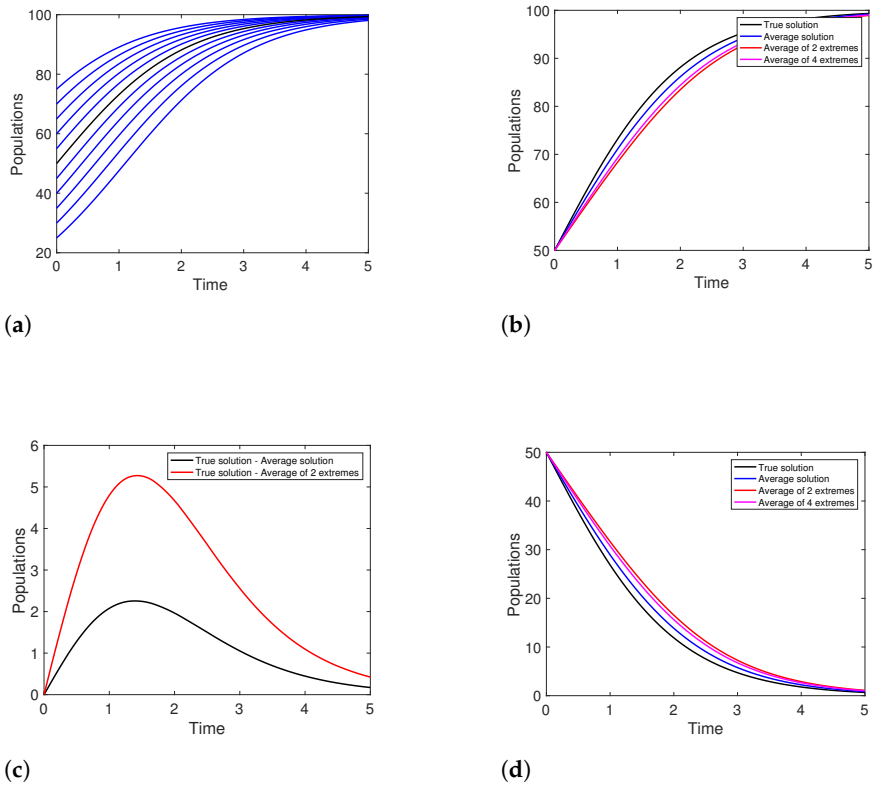


Figure 1. Consequences of averaging. (**a**) Eleven runs shown for the logistic equation ; black has average initial condition. (**b**) Run with average initial condition (black), average of all runs (blue), average of the two extreme runs (red), and average of the four extreme runs (magenta).The run with the average initial condition, labeled "true solution", dominates the average of all solutions, with the average of two or four extreme solutions being even lower. (**c**) The deficit between the runs in part b. The differences between run with average initial condition and average of all runs (black), or average of two extreme runs (red) are shown. (**d**) Here the sign of the growth parameter is negative, as in the susceptible population of an SI model. The runs with average initial (black), average of all runs (blue), average of two extreme runs (red) and average of four extreme runs (magenta) are shown.Here the average of all solutions dominates the trajectory associated to the average initial condition.

The main theorem in this study compares the trajectory of a particular solution to the logistic equation with the average of multiple trajectories whose average initial condition

matches the particular solution. The point of the theorem is that the patterns seen in Figure 1 hold for any such set of initial conditions. In other words, the process of averaging multiple solutions in this manner produces values that are not only different from a single solution but biased in a consistent direction. Two corollaries apply the theorem to a situation with multiple perturbations and to a simple Susceptible/Infectious (SI) model. It is noted that the theorem can be extended to systems of the form $f' = G(f)$ where G is convex.

2. Main Theorem for the Logistic Equation

Theorem 1. *Let f_i, $i = 1, \ldots, n$ be n solutions to the logistic equation given by $f_i' = af_i - bf_i^2$, $a, b > 0$, where each f_i has a different positive initial condition, with the average of initial conditions denoted $f_{av}(0) = \frac{1}{n}\Sigma_i f_i(0)$. Let Y be the solution to the logistic equation with initial condition $f_{av}(0)$. Let $\bar{F} = \frac{1}{n}\Sigma_i f_i$. Then $\bar{F} \leq Y$ with equality only when $n = 1$.*

Proof of Theorem 1. The average of the family of n solutions, $\bar{F} = \frac{1}{n}\Sigma_i f_i$ solves the differential equation

$$\bar{F}' = \frac{1}{n}\Sigma_i f_i' = \frac{1}{n}\Sigma_i(af_i - bf_i^2). \tag{1}$$

Rearranging terms we have

$$\bar{F}' = \frac{1}{n}\Sigma_i(af_i) - \frac{1}{n}\Sigma_i(bf_i^2) = A - B \tag{2}$$

where $A = \frac{1}{n}\Sigma_i(af_i)$ and $B = \frac{1}{n}\Sigma_i(bf_i^2)$.

Further we have that

$$A = \frac{1}{n}\Sigma_i(af_i) = \frac{a}{n}\Sigma_i(f_i) = a\bar{F}. \tag{3}$$

The expression given by B may be compared with $b\bar{F}^2$.

The differential equation given by Equation (1) may now be expressed as

$$\bar{F}' = \frac{1}{n}\Sigma_i f_i' = a\bar{F} - \frac{1}{n}\Sigma_i(bf_i^2) \tag{4}$$

where, for convenience we denote

$$J = (\bar{F}^2 - \frac{1}{n}\Sigma_i(f_i^2)). \tag{5}$$

A preliminary calculation gives:

$$\bar{F}^2 = (\frac{1}{n}\Sigma_i f_i)^2 = \frac{1}{n^2}(\Sigma_i f_i^2 + \Sigma_{i\neq j} f_i f_j)$$

$$= \frac{1}{n^2}(\Sigma_i f_i^2 + \Sigma_{i>j} 2f_i f_j)$$

and thus

$$-n^2 J = -(\Sigma_i f_i^2 + \Sigma_{i>j} 2f_i f_j) + n\Sigma_i(f_i^2)) = (n-1)\Sigma_i(f_i^2) - \Sigma_{i>j} 2f_i f_j. \tag{6}$$

Accounting for all terms in this expression gives

$$-n^2 J = \Sigma_{j=2}^n (f_1^2 - 2f_1 f_j + f_j^2)$$
$$+\Sigma_{j=3}^n (f_2^2 - 2f_2 f_j + f_j^2)$$
$$+ \ldots$$
$$+\Sigma_{j=k}^n (f_{k-1}^2 - 2f_{k-1} f_j + f_j^2)$$

$$+\ldots$$

A count shows that each f_j^2 occurs in this sum exactly $n-1$ times, and each $2f_if_j$ occurs exactly once, giving

$$-n^2J = \Sigma_{k=2}^{n}\Sigma_{j=k}^{n}(f_{k-1}^2 - 2f_{k-1}f_j + f_j^2) = \Sigma_{k=2}^{n}\Sigma_{j=k}^{n}(f_{k-1} - f_j)^2 > 0. \quad (7)$$

Thus the expression J is negative. Substituting $(\bar{F}^2 - J)$ for $\frac{1}{n}\Sigma_i(f_i^2)$ from Equation (5) into Equation (4) gives

$$\bar{F}' = (a\bar{F} - b\bar{F}^2) + bJ \leq (a\bar{F} - b\bar{F}^2) \quad (8)$$

where $J < 0$.

Therefore the derivative of $\bar{F}'$ is dominated by the logistic equation. Because $\bar{F}$ and Y have the same initial conditions, it follows that $\bar{F} \leq Y$ with equality only when $n = 1$. □

A reviewer of this paper points out that the theorem may be generalized to any equation of the form $f' = G(f)$ where G is a concave function and suitable constraints are placed on parameters and initial conditions. The proof replaces the algebraic argument given in Equations (6) and (7) with a convexity result due to Jensen [19]. We keep the algebraic argument here because Equation (7) gives an expression for J, the quantity controlling the difference in growth between $\bar{F}$ and Y, the solution to the logistic equation at the average initial value.

Corollary 1. *Let f_i, $i = 1, \ldots, n$ be n solutions to the logistic equation given by $f_i' = af_i - bf_i^2$, $a, b > 0$ where each f_i is updated stochastically at times t_j by the addition of $\epsilon_{i,j}$ to the function f_i such that $\Sigma_i\epsilon_{i,j} = 0$ and all f_i remain positive after perturbation. Let $\bar{F} = \frac{1}{n}\Sigma_i f_i$. Let Y be the solution to the logistic equation with initial condition $Y(0) = \frac{1}{n}\Sigma_i f_i(0)$. Then $\bar{F} \leq Y$ with equality only when $n = 1$.*

Proof. Note that $\bar{F}$ remains the same at t_j after each f_i is perturbed by $\epsilon_{i,j}$ because $\Sigma_i\epsilon_{i,j} = 0$ by assumption. Without loss of generality, Theorem 1 may be applied to a system starting at time j. Let W_j be the solution of the logistic equation with initial condition $W_j(t_j) = \bar{F}(t_j)$. By the theorem, $\bar{F}(t_{j+1}) \leq W_j(t_{j+1})$. Thus at each perturbation, $\bar{F}(t_{j+1})$ is less than the associated solution to the logistic equation passing through the $\bar{F}(t_j)$ at the prior time of perturbation. By induction, the average at every time remains less than Y. □

Corollary 2. *Under the same hypotheses as the corollary, adding one more stochastic update (of the same sort) to the sequence will further lower $\bar{F}$ and increase its deficit from Y.*

3. The SI Model

Consider the simplest model of disease transmission where a positive susceptible population (S) contracts a disease and becomes infectious (I) and never recovers. In this model the total population (N) is fixed, thus $S + I = N$. This system is described by two equations. Let $a > 0$ be a parameter describing the rate of disease transmission.

$$S' = -aSI \quad (9)$$

and

$$I' = aSI. \quad (10)$$

Using the relationship $S + I = N$ results in two decoupled equations:

$$S' = -aS(N - S) \quad (11)$$

and

$$I' = a(N - I)I. \tag{12}$$

Both of these are versions of the logistic equation and amenable to the same analysis, leading to another corollary to Theorem 1.

Corollary 3. *Let S_j, I_j, $j = 1, \ldots, n$ be n solutions to the SI model described above, where each S_j, I_j may have a different positive initial condition, with the average of initial conditions denoted $S_{av}(0) = \frac{1}{n}\Sigma_j S_j(0)$ and $I_{av}(0) = \frac{1}{n}\Sigma_j I_j(0)$. Let X, Y be the solution to the logistic equation with initial condition $S_{av}(0), I_{av}(0)$. Let $\bar{S} = \frac{1}{n}\Sigma_j S_j$ and let $\bar{I} = \frac{1}{n}\Sigma_j I_j$. Then $\bar{S} \geq X$ and $\bar{I} \leq Y$ with equality only when $n = 1$.*

4. Discussion

The theorem and its corollaries show that if the logistic equation or, equivalently, the SI model, are perturbed at a single time point by a range of values with mean zero, the resulting trajectories must average to a value below (for logistic and I) or above (for S) the solution with average value at that time. The proof of the theorem shows that this phenomenon is the result of algebraic properties of the nonlinear quadratic term, although it can be extended to more general equations of the form $f' = G(f)$ where G is a convex function. More importantly it shows that the only necessary criterion is that the perturbations average to zero. The source of them does not matter and the properties of their distribution matters only to the magnitude of the difference.

4.1. Implications for Sensitivity Analysis and Parameter Estimation

The logistic equation has been used extensively to measure the regrowth trajectories of tumors after radiation therapy, both of a single subject and also aggregating the results of many subjects [7,20]. It is assumed that, after radiation, the tumor of a single individual regrows according to a logistic equation. At the time of radiation the tumor size is reduced depending on the radiation dosage, and is measured. This measurement becomes the initial condition for the regrowth trajectory, and there may be error in this measurement. To quantify the effects of this error one would use a sensitivity analysis of the expected tumor size at a later date with respect to perturbed values of the initial condition centered around the actual measurement taken.

In any sensitivity analysis, the researcher must decide in what domain to take the perturbed values. Suppose one conducts a sensitivity analysis of the logistic equation with respect to initial conditions, with these varying round a mean ($f_{av}(0)$), and the output values are solutions at some particular time, t_{out}. By the theorem proven here, this analysis would produce a histogram of output values with a mean value that is less than $f_{av}(t_{out})$). Equation (7) further shows that the difference between the mean of trials and the value of $f_{av}(t_{out})$) depends on the differences between individual trajectories at the time of measurement, which in turn depends on the spread of the initial conditions. The distribution of these initial conditions will affect how large the difference is between $f_{av}(t_{out})$) and the distribution mean.

If, in turn, the distribution mean is used to estimate the overall growth parameter for these tumors, it will necessarily be lower than the growth parameter for a single tumor, because the average of trajectories at (t_{out}) is dominated by $f_{av}(t_{out})$ for any output time chosen. This observation is also relevant to the process of model choice [6].

4.2. Observations on Perturbed, Agent Based, or Stochastic Systems

Variation can be added to a system in a variety of ways. If shocks are given at discrete time points, as in the case of fractionated radiation therapy where multiple doses are given, the consequences of averaging are described in Corollaries 1 and 2 [20]. However, most modelers use different approaches, varying derivatives rather than values or varying parameters or adding noise into an agent based model. These are not covered by the

theorems in this study, but similar phenomenon can still be seen in the literature. Four examples are given.

Petroni et al analyze a stochastic version of the logistic equation, giving an integral expression for the expected value of the solution [1]. The integrals could not be resolved analytically but were illustrated computationally. It is seen in their figures that the expected value of solutions to the stochastic problem was less than the solution to the original deterministic logistic equation. As the variance of the noise increased, this difference became larger. The authors state that, as the variance of noise decreased, the expected value approached the deterministic trajectory from below [1]. Both observations are consistent with the results given here, one by virtue of Theorem 1 and the other because of the nature of Equation (7).

Rahmandad and Sterman compare agent based with differential equation models for disease transmission using the susceptible/exposed/infectious/recovered (SEIR) dynamic [2]. In the agent based model, contact frequencies for each individual varied according to a network structure. For five types of networks tested, the agent based simulation exhibits on average a lower and later peak of infection than the corresponding deterministic ODE model [2]. A comparison for the susceptible/infectious/recovered (SIR) model done by Macal exhibits a similar phenomenon, with the infectious compartment in the deterministic model peaking higher and earlier than the comparable agent based simulation [3]. These observations are consistent with Corollary 3. The author attributes the difference to the ability of the agent based model to capture the stochasticity in the system, but does not comment on why that would produce an average that is lower than the deterministic model.

Srivastava et al study intraviral kinetics using both stochastic and deterministic models, based on three coupled ODEs [4]. Tracking one of the molecules involved, they observe that the average of 3000 stochastic simulations was always lower than the corresponding deterministic model. They point out that fitting these two types of model to data would result in different parameters being assigned to the model, even though these parameters are supposed to describe the same phenomenon [4]. Figueredo et al compare ODE, agent-based and stochastic models in three different models, two of which include quadratic terms and show the ODE output as consistently higher than the other models [21].

These studies share some important commonalities. All of the models considered are nonlinear, and in all cases the nonlinearity is quadratic. The variation in the agent-based and stochastic models can be done in a variety of ways, either varying different parameters or inserting a variety of forms of stochastic noise in various terms of the deterministic equation. Yet, in all cases the average of runs for a quantity in the agent-based or stochastic version of the problem is consistently above or below the deterministic outcome. Authors offer various explanations for why the results differ, but offer no explanation for the sign of the difference. The theorem and corollaries proven here do not cover these cases, but do offer an explanation for some of the observed phenomena based on the nonlinear nature of the model.

4.3. Final Comments

Although the situation analyzed here is not exactly the same as the examples described in the introduction, parallel phenomena occur. As in Petroni et al, larger variation leads to a greater difference between the solution and the average of solutions, as seen in Figure 1b,c [1]. As with the SIR and SEIR examples, the average of solutions for S is higher than the solution, while for I it is lower [2,3]. These observations lead to two lines of questioning.

As variation can be added to the system in a variety of ways, it would be worth exploring the consequences of varying parameters, whether at a single time point or in an ongoing fashion. Theorems addressing these types of variation would be useful to understand discrepancies between models and in selecting what sources of variation to use in a model. The theorem presented in this study could also perhaps be generalized to higher

dimensional systems with quadratic nonlinearity, such as SEIR or the Ross-Macdonald model for malaria.

When two approaches to modeling the same phenomenon not only disagree, but can be proven to disagree in a consistent manner, one is forced to wonder which is the better representation of reality. This question is neither mathematical nor biological, but both, depending on the assumptions inherent in the description of the problem to be addressed and the nature of any data available for comparison.

Funding: This study was supported by NSF RII Track-2 FEC: Leveraging Big Data to Improve 224 Prediction of Tick-Borne Disease Patterns and Dynamics (Award Number 2019609).

Data Availability Statement: No data was collected or created in this study.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Petroni, N.C.; De Martino, S.; De Siena, S. Logistic and θ-logistic models in population dynamics: General analysis and exact results. *J. Phys. A Math. Theor.* **2020**, *53*, 445005. [CrossRef]
2. Rahmandad, H.; Sterman, J. Heterogeneity and network structure in the dynamics of diffusion: Comparing agent-based and differential equation models. *Manag. Sci.* **2008**, *54*, 998–1014. [CrossRef]
3. Macal, C.M. To agent-based simulation from system dynamics. In Proceedings of the 2010 Winter Simulation Conference, Baltimore, MD, USA, 5–8 December 2010; pp. 371–382.
4. Srivastava, R.; You, L.; Summers, J.; Yin, J. Stochastic vs. deterministic modeling of intracellular viral kinetics. *J. Theor. Biol.* **2002**, *218*, 309–321. [CrossRef] [PubMed]
5. Bull, J.A.; Byrne, H.M. The hallmarks of mathematical oncology. *Proc. IEEE* **2022**, *110*, 523–540. [CrossRef]
6. Simpson, M.J.; Browning, A.P.; Warne, D.J.; Maclaren, O.J.; Baker, R.E. Parameter identifiability and model selection for sigmoid population growth models. *J. Theor. Biol.* **2022**, *535*, 110998. [CrossRef] [PubMed]
7. Bekker, R.A.; Kim, S.; Pilon-Thomas, S.; Enderling, H. Mathematical modeling of radiotherapy and its impact on tumor interactions with the immune system. *Neoplasia* **2022**, *28*, 100796. [CrossRef] [PubMed]
8. Boucharas, D.G.; Anastasiadou, C.A.; Karkabounas, S.; Antonopoulou, E.; Manis, G. Mathematical Modeling and Growth Model Analysis for Preventing the Cancer Cell Development. In Proceedings of the 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), Ioannina, Greece, 27–30 September 2022; pp. 1–4.
9. Chuprin, J.; Buettner, H.; Seedhom, M.O.; Greiner, D.L.; Keck, J.G.; Ishikawa, F.; Shultz, L.D.; Brehm, M.A. Humanized mouse models for immuno-oncology research. *Nat. Rev. Clin. Oncol.* **2023**, 20, 192–206. [CrossRef] [PubMed]
10. Wallace, D. Parts of the Whole: Error Estimation for Science Students. *Numer. Adv. Educ. Quant. Lit.* **2017**, *10*, 11. [CrossRef]
11. Kermack, W.O.; McKendrick, A.G. A contribution to the mathematical theory of epidemics. *Proc. R. Soc. London. Ser. A Contain. Pap. A Math. Phys. Character* **1927**, *115*, 700–721.
12. Prodanov, D. Asymptotic analysis of the SIR model and the Gompertz distribution. *J. Comput. Appl. Math.* **2023**, *422*, 114901. [CrossRef]
13. Kalachev, L.; Landguth, E.L.; Graham, J. Revisiting classical SIR modelling in light of the COVID-19 pandemic. *Infect. Dis. Model.* **2023**, *8*, 72–83. [CrossRef] [PubMed]
14. Amaro, J.E. Systematic description of COVID-19 pandemic using exact SIR solutions and Gumbel distributions. *Nonlinear Dyn.* **2023**, *111*, 1947–1969. [CrossRef] [PubMed]
15. Dong, S.; Xu, L.; Lan, Z.Z.; Xiao, D.; Gao, B. Application of a time-delay SIR model with vaccination in COVID-19 prediction and its optimal control strategy. *Nonlinear Dynamics* **2023**, *111*, 10677–10692. [CrossRef] [PubMed]
16. Yüzbaşı, Ş.; Yıldırım, G. A Pell–Lucas Collocation Approach for an SIR Model on the Spread of the Novel Coronavirus (SARS CoV-2) Pandemic: The Case of Turkey. *Mathematics* **2023**, *11*, 697. [CrossRef]
17. Southall, E.; Ogi-Gittins, Z.; Kaye, A.; Hart, W.; Lovell-Read, F.; Thompson, R. A practical guide to mathematical methods for estimating infectious disease outbreak risks. *J. Theor. Biol.* **2023**, *562*, 111417. [CrossRef] [PubMed]
18. Wallace, D.; Chen, M. A simulation of the US influenza outbreak in 2009–2010 using a patch sir model based on airport transportation data. In *BIOMAT 2013: International Symposium on Mathematical and Computational Biology*; World Scientific: Singapore, 2014; pp. 286–297.
19. Jensen, J.L.W.V. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Math.* **1906**, *30*, 175–193. [CrossRef]

20. Zahid, M.U.; Mohamed, A.S.; Latifi, K.; Rishi, A.; Harrison, L.B.; Fuller, C.D.; Moros, E.G.; Caudell, J.J.; Enderling, H. Proliferation saturation index to characterize response to radiation therapy and evaluate altered fractionation in head and neck cancer. *Appl Radiat Oncol* **2021**, *10*, 32–39.
21. Figueredo, G.P.; Siebers, P.O.; Owen, M.R.; Reps, J.; Aickelin, U. Comparing stochastic differential equations and agent-based modelling and simulation for early-stage cancer. *PLoS ONE* **2014**, *9*, e95150. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.