Reconstructing and Forecasting the COVID-19 Epidemic in the US Using a 5-Parameter Logistic Growth Model

Ding-Geng Chen, School of Social Work, University of North Carolina, Chapel Hill, NC, USA

Email: dinchen@email.unc.edu

Xinguang Chen, Department of Epidemiology, University of Florida, Gainesville, FL, USA

Email: jimax.chen@ufl.edu

Jenny K. Chen, Department of Statistics and Data Science, Cornell University, Ithaca, NY, USA

Email: jkc229@cornell.edu

All correspondence to Professor Ding-Geng Chen at dinchen@email.unc.edu.

Abstract

Background: Many studies have modeled and predicted the epidemic of COVID-19 in the US using data that starts from the first reported cases. However, because of the shortage of test services to detect the infected, this approach is subject to error due to under-detection in the early period of the epidemic. We attempted a new approach to overcome this limitation and to provide data supporting the public policy decisions against the life-threatening COVID-19 epidemic.

Methods: Documented data by CDC were used, including daily new and cumulative cases of confirmed COVID-19 in the US from January 22 to April 6, 2020. A 5-parameter logistic growth model was used to reconstruct the epidemic. Instead of all data in the whole study period, we fitted data in a 2-week window from March 21 to April 4 (approximately one incubation period) during which massive testing services were in position. With parameters obtained from the modeling, we reconstructed and predicted the epidemic and evaluated the under-detection.

Results: The data fit the model satisfactorily. The estimated daily growth rate was 16.8% (95% CI: 15.95%, 17.76%) overall, with 4 consecutive days having a doubling growth rate. Based on the modeling result, the tipping point for new cases to decline will be on April 7th, 2020, with 32,860 new cases. By the end of the epidemic, a total of 792,548 (95% CI: 789,162-795,934) will be infected. Based on the model, a total of 12,029 cases were not detected from the first case from January 22 to April 4.

Conclusions: Study findings suggest the usage of a 5-parameter logistic growth model with reliable data that comes from a specified window period, where governmental interventions are appropriately implemented. In addition to informing decision-making, this model adds one tool for use to capture the underlying COVID-19 epidemic caused by a novel pathogen.

Key Words: COVID-19; Logistic growth model; USA; Prediction; Reconstruction; Under-detection

Introduction

COVID-19 is an infection caused by a novel pathogen named as SARS-Cov-2. The pandemic of COVID-12 is a typical example of global health issues (Chen et al., 2020), and it spread to the world only in less than five months. Since the first case reported in the US in January 22, many studies used different models to reconstruct the epidemic and to forecast the future trends, from simple growth models to classic susceptible-infectious-recovery models (Huang, et al., 2020). Since little information is available for COVID-19 during the early period of the epidemic, there is a lack of data to construct complex and classic epidemiological models, leaving the population-based ecological growth model as a preferable option.

Historically, various population-based models are available in the literature to model population dynamics in demography and disease epidemics in public health and medicine. The first is the 1-parameter exponential growth model. In this model, population growth has no upper limit and is determined by one parameter of growth rate. To reflect the upper limit of population growth, the 2-parameter logistic growth model was developed. In this model, the population growth rate is exponential in the beginning, but this growth rate gets smaller and smaller as population size approaches a maximum carrying capacity as detailed described in Richards (1959), McIntosh (1985), Renshaw (1991), Kingsland (1995), and Vandermeer (2010).

To obtain additional characteristics key to understanding population growth, the 2-parameter logistic growth model has been extended to 3-parameter, 4-parameter, and 5-parameter logistic growth models. These models have been widely used in other fields of research, including demography and analytical chemistry (Gottschalk and Dunn, 2005; Motulsky and Brown, 2006). Despite many advantages, no study employed the method to investigate the

COVID-19 epidemic in the United States and other countries in the world. One purpose of this study is to assess the utility of the 5-parameter growth model.

Unlike typical population growth, only a small number of COVID-19 cases will be detected in the early period of an epidemic. The detected cases could approach the real epidemic if the time of outbreak of the epidemic was known. Consequently, if extensive testing services were implemented, detection would be more accurate. Reported data indicate the incubation period of COVID-19 is about 14 days (Chen & Yu, 2020); and COVID-19 test services in the US started from mid-March and sustained thereafter following CDC guidance. This provides a window time of 14 days with the highes level of detection rate that will not be affected by removal of the infected on the growth curve, which is ideal for model building. In principle, a model built with such data would be much closer to the truth than those with data from the whole study period. We tested this approach in analyzing the COVID-19 in the US.

Materials and Methods

Data: Data for this study were daily cumulative cases from January 22 to April 6, 2020. This real-time data were compiled from USA CDC website which were available by the time this study was conducted (https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html, accessed on April 7, 2020).

Model: We modeled the data using the 5-parameter logistic growth model as below:

$$C(t) = C_{min} + \frac{C_{max} - C_{min}}{\left[1 + e^{-r(t - t_{mid})}\right]^{\alpha}} \quad (1)$$

where

1) $C(t)$ is the cumulative cases of COVID-19 over time, t (t=1/22,1/23,…,4/6, 2020);

2) C_{min} is the minimum number of cases at the beginning of the epidemic on January 22, 2020 when the first case was reported in the US;
3) C_{max} is the maximum number of cases when the epidemic ends, it is the model-predicted total number of Americans who would be infected with COVID-19;
4) r is the daily exponential growth rate;
5) t_{mid} is the estimated tipping point when the daily new cases start to level-off where the daily new cases would increase at the left-side and decrease at the right-side; and
6) α is an asymmetric parameter quantifying the skewness of the distribution of daily new cases. α =1 indicate a symmetric distribution centered at t_{mid}; α >1 indicates faster increases in new cases before t_{mid} and slower after t_{mid}; and the pattern will be reversed if α <1.

With Model 1 defined above, daily new cases $D(t)$ can be obtained by taking the first derivative of the model:

$$D(t) = C'(t) = \frac{\alpha r(C_{max}-C_{min})}{\left[1+e^{-r(t-t_{mid})}\right]^{\alpha+1}} \times e^{-r(t-t_{mid})} + \in (t), \qquad (2)$$

where the error term $\in (t)$ is assumed to be normally distributed with mean 0 and standard deviation of σ.

Implementation of modeling analysis

Data analysis was conducted using the software R. Daily data for a window period from March 21 to April 4 were fitted with a 5-parameter logistic grow model as shown in Model 2. Modeling analysis was implemented using a nonlinear optimization algorithm to minimize the sum of squared errors between the observed and model-estimated data. The optimization process

was achieved by calling the R function "optim". Estimates were thus obtained through the optimization process for the five parameters C_{min}, C_{max}, t_{mid}, r, and α with a significance level set at $p < 0.05$ (two-sided).

With the estimated five model parameters, model-based cumulative and new cases day by day were estimated up to April 6 and predicted beyond April 6 using Model 1 and 2, respectively. Under-detection cases in a specific period were computed as the differences between the reported and model predicted cases.

Results

The cumulative daily cases from March 21 to April 4 fit Model 2 satisfactorily and the model fit converged nicely. The estimated parameters, their standard error and 95% CI are summarized in Table 1. All model parameters were statistically significant at $p < 0.001$ level except C_{min}. The lack of significance for C_{min} appears to be reasonable given the small scale of this number relative to other parameters and practical difficulties in determining the number of cases at the beginning of the epidemic when the first few COVID-19 cases were detected and reported.

Table 1: Summary of Parameter Estimation

Parameter	Estimate	SE	p-value	Lower 95% CI	Upper 95% CI
C_{min}	29.999	2059.86	0.988	-4007.33	4067.32
C_{max}	792,548	1727.56	<0.0001	789,162	795,934
t_{mid}	76.9	0.456	<0.0001	75.952	77.739
r	0.16854	0.00463	<0.0001	0.15947	0.17761

α	0.95364	0.06194	<0.0001	0.83224	1.07504

Note: The parameters were estimated based on the daily USA cases for a window period of 14 days from March 21 to April 4, 2020.

Based on the modeling results, an estimate of 792,548 (95% CI: 789,162-795,934) Americans will be infected with COVID-19 by the time the epidemic ends. This number is slightly twice the number of infections by April 6. The estimated tipping point is on April 7, 77 (95% CI=76-78), from the beginning of the epidemic on January 22. Therefore, we expect the epidemic curve will be flattened at around April 6 to 8, 2020.

The estimated exponential daily growth rate of COVID-19 in the US population is 16.9% (95% CI: 15.9%-17.8%). This growth rate for the US is close to 17.12%, the rate observed in China (Chen & Yu, 2020). This rate suggests that the COVID-19 cases in the US will double every four days if no anti-epidemic actions are in place. The estimated asymmetric parameter α was 0.954 (95% CI: 0.832-1.075), which is not statistically different than 1.0. This result indicates that changes in COVID-19 cases before and after the tipping point of April 7 following a similar pattern.

For further illustration, Table 2 summarizes three sets of information, including the data used for the model fitting section, a smaller reconstruction section, and a prediction section, lined up by days from the beginning of the epidemic. With this fitted model, the under-detection of COVID-19 cases were substantial. By April 7 when this study was completed, a total of 395,011 detected cases were reported; with our model, we forecasted an under-detection of 19,291 cases based on data from the CDC.

Table 2: Illustration of data usage with reported, predicted and under-reported counts.

Data Usage	Days	Date	Reported		Predicted		Under-Reported
			Total Cases	Daily Cases	Daily Cases	Total Cases	
Reconstruction	54	3/15/2020	3487	1253	3108	19781	16294
	55	3/16/2020	4226	739	3623	23141	18915
	56	3/17/2020	7038	2812	4218	27054	20016
	57	3/18/2020	10442	3404	4902	31606	21164
	58	3/19/2020	15219	4777	5687	36892	21673
	59	3/20/2020	18747	3528	6584	43019	24272
Fitting	**60**	**3/21/2020**	**24583**	**5836**	**7603**	**50102**	**25519**
	61	**3/22/2020**	**33404**	**8821**	**8755**	**58269**	**24865**
	62	**3/23/2020**	**44183**	**10779**	**10047**	**67658**	**23475**
	63	**3/24/2020**	**54453**	**10270**	**11485**	**78411**	**23958**
	64	**3/25/2020**	**68440**	**13987**	**13070**	**90676**	**22236**
	65	**3/26/2020**	**85356**	**16916**	**14797**	**104598**	**19242**
	66	**3/27/2020**	**103321**	**17965**	**16656**	**120315**	**16994**
	67	**3/28/2020**	**122653**	**19332**	**18624**	**137947**	**15294**
	68	**3/29/2020**	**140904**	**18251**	**20670**	**157589**	**16685**
	69	**3/30/2020**	**163539**	**22635**	**22750**	**179298**	**15759**
	70	**3/31/2020**	**186101**	**22562**	**24810**	**203082**	**16981**
	71	**4/1/2020**	**213144**	**27043**	**26784**	**228889**	**15745**
	72	**4/2/2020**	**239279**	**26135**	**28600**	**256597**	**17318**
	73	**4/3/2020**	**277205**	**37926**	**30180**	**286010**	**8805**
	74	**4/4/2020**	**304826**	**27621**	**31453**	**316855**	**12029**
Forecast	75	4/5/2020	330891	26065	32352	348791	17900
	76	4/6/2020	374329	43438	32830	381419	7090
	77	4/7/2020	395011	20682	32860	414302	19291
	78	4/8/2020	427460	32449	32436	446987	19527
	79	4/9/2020	459165	31705	31582	479030	19865
	80	4/10/2020	NA	NA	30340	510021	NA
	81	4/11/2020	NA	NA	28773	539601	NA

Figures 1 and 2 present the information of both new cases (i.e., daily cases) and cumulative cases estimated using the model, contrasted with the observed data. Overall, it will take approximately one month from 20,000+ per day from around April 7 to 100s per day at around early May. Correspondingly, after the tipping point, the cumulative cases will continue to increase rapidly after the tipping point until early May as illustrated in Figure 2.

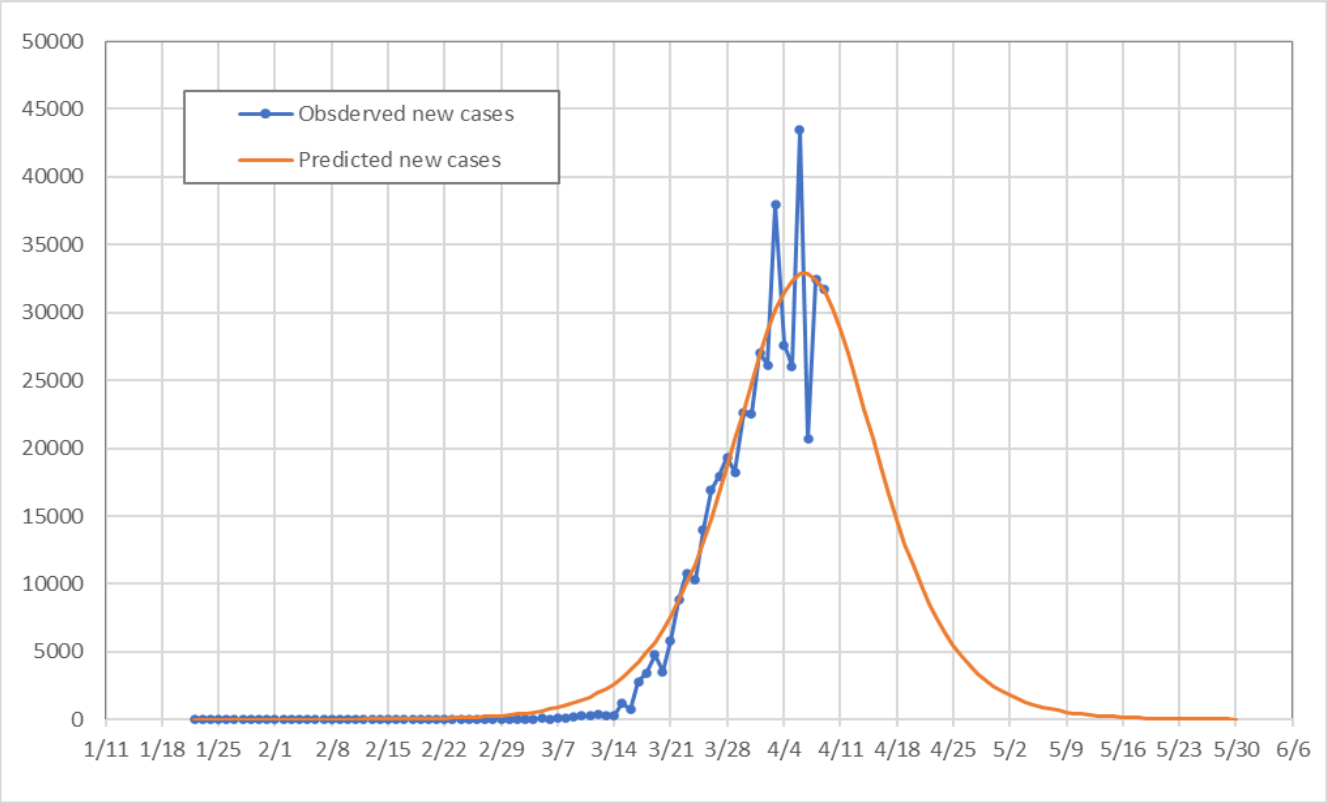

Figure 1. Observed, model-estimated and forecasted daily new cases of COVID-19 January 22 - May 30, 2020, the United States of America

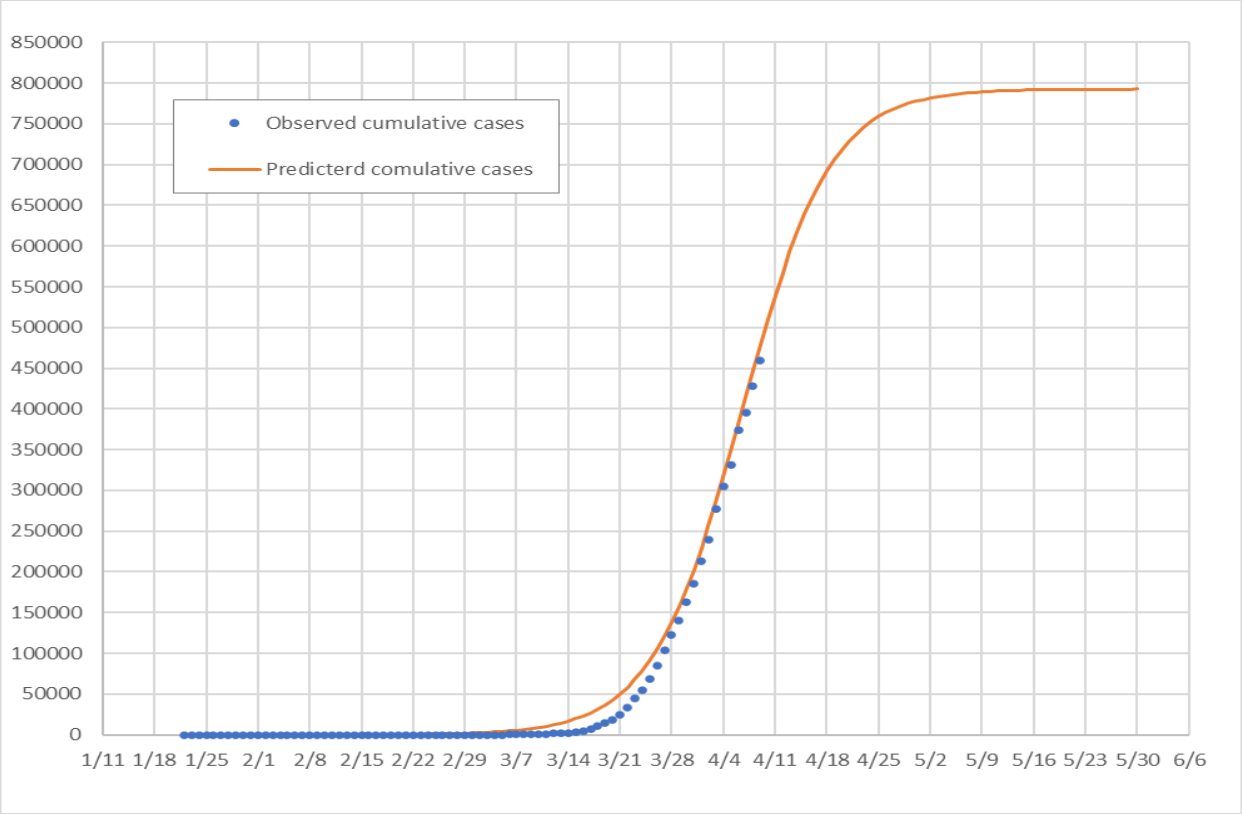

Figure 2. Observed and model-estimated and forecasted daily cumulative cases of COVID-19 January 22 - May 30, 2020, the United States of America

Discussions and Conclusions

In this study, we reported our work to model, reconstruct, and forecast the COVID-19 epidemic with a 5-parameter logistic grow model, a method widely used in demography, biology and other hard sciences. We are the first to use it in analyzing the epidemic of COVID-19 in the US. In addition, we innovatively used data from the period with more complete detection of new cases to fit the model, and then used the fitted model to reconstruct the cases before and after the study period and forecasted the future beyond the study period.

From a global health perspective, control the COVID-19 epidemic in the US is an essential part of fighting the pandemic across the globe (Chen, et al, 2020). This study provides

data much needed for public health decision-making to end the epidemic in the United States. In addition, this study demonstrates the utility and efficiency of the 5-parameter logistic growth model in understanding the epidemic of a new infection during its early period when the need for information is extremely high but not much data is available. Our modeling method provides a tool to overcome the challenge.

Based on findings from our modeling analysis, the likelihood would not be high for the new cases to increase continuously after the tipping point. However, by the end of the epidemic, an estimate of approximately 800,000 Americans will be infected. This number is lower than those by others that can be as high as 2.2 million in Ferguson et al. (2020) and 214 million from New York Intelligencer (accessed on April 12, 2014, from https://nymag.com/intelligencer/2020/03/cdcs-worst-case-coronavirus-model-210m-infected-1-7m-dead.html). At this moment, no one can tell which estimates are more reliable. The accuracy of our estimation will be tested along with the ongoing development of the epidemic in the United States.

The daily exponential growth rate of COVID-19 was 16.85% for the US population. This is very close to 17.12%, the rate estimated for the same COVID-19 in China (Chen and Yu, 2020). In the early period of an epidemic, this growth rate can be obtained with limited data, and the importance of growth rate is more information than other parameters such as R_0 (the reproduction number) to guide anti-epidemic actions. Growth rates provide a dynamic measure of instantaneous change, duration of doubling based on growth rate is practically quite useful to guide and evaluate anti-epidemic measures while R_0 is a static measure with timing not included, and it may be of great value for research, but can hardly be determined accurately at the early stage of an new epidemic with rather limited data.

There are limitations to this study. First, selection of data from a window is more subjective than objective. Caution is needed when the same method is used in different countries/regions with different anti-epidemic strategies implemented in different ways. Second, additional work is needed to improve the confidence of C_{min}, the minimum number of cases at the beginning of an epidemic. It is a challenge to improve the estimation given the large range of different measures in the model. For example, the differences between C_{min} and C_{max} in our analysis is from about 30 to 800,000. Furthermore, the reported cases at the beginning of the epidemic are highly unreliable and thus will lead to an unreliable estimation of C_{min}.

Despite the limitations, findings from this study provided timely data, much needed for public health decision-making to end the epidemic. We will continue to update our model as more data become available with the evolvement of the COVID-19 epidemic in the United States.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The data that support the findings of this study are available from https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html. Data accessed at 4:30pm, April 7, 2020.

Competing interests

The authors declare no competing interests associated with this study.

Funding

Not applicable.

Authors' contributions

All three authors participated in data validation, data analysis and manuscript preparation.

Acknowledgements

Not applicable.

Authors' information (optional)

Dr. Ding-Geng Chen is the Wallace H. Kuralt distinguished professor at School of Social Work, University of North Carolina at Chapel Hill. Dr. Xinguang Chen is a professor at the Department of Epidemiology, University of Florida. Ms. Jenny K. Chen is a student at the Department of Statistics and Data Science at Cornell University.

References

Chen, X. & Yu, B. (2020). First two months of the 2019 coronavirus disease (COVID-19) epidemic in China: real-time surveillance and evaluation with a second derivative model. Global Health Research and Policy. 2020 Mar 2; 5:7. doi: 10.1186/s41256-020-00137-4.

Chen, X., Li, H., Lucero-Prisno, D., Abdullah, A., Huang, J., Laurence, C., Laing, X., Ma, Z., Mao, Z., Ren, R., Wu, S., Wang, N., Wang, P., Wang, T., Yan, H., Zhou, Y. (2020). What is global health? Key concepts and clarification of misperceptions: Report of the 2019 GHRP editorial meeting. Global Health Research and Policy. 2020, Apr 7; 5:14. doi: 10.1186/s41256-020-00142-7

Ferguson, N. L., Laydon, D., Nedjati-Gilani, G., Imai, N., Ainslie, K., Baguelin, M., Bhatia, S., Boonyasiri, A., Cucunubá, Z., Cuomo-Dannenburg, G., Dighe, A., Dorigatti, I., Fu, H., Gaythorpe, K., Green, W., Hamlet, A., Hinsley, W., Okell, L., van Elsland, A., Thompson, H., Verity, R., Volz, E., Wang, H., Wang, Y. and Walker, P., Walters, C., Winskill, P., Whittaker, C., Donnelly, C. A., Riley, S., Ghani, A.C. (2020). Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand. On behalf of the Imperial College COVID-19 Response Team. WHO Collaborating Centre for Infectious Disease Modelling, MRC Centre for Global Infectious Disease Analysis, Abdul Latif Jameel Institute for Disease and Emergency Analytics, Imperial College London (accessed on April 12, 2020 from https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-College-COVID19-NPI-modelling-16-03-2020.pdf).

Huang, Y., Yang, L., Dai, H., Tian, F. & Chen, K. (2020). Epidemic situation and forecasting of COVID-19 in and outside China. [Submitted]. Bull World Health Organ. E-pub: 16 March 2020. doi: http://dx.doi.org/10.2471/BLT.20.255158.

Kingsland, S. E. (1995). Modeling Nature: Episodes in the History of Population Ecology. University of Chicago Press.

McIntosh, R. P. (1985) The Background of Ecology. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511608537.

Gottschalk, P. G., Dunn, J.R. (2005). The five-parameter logistic: a characterization and comparison with the four-parameter logistic. Analytical Biochemistry, 343(1):54-65.

Motulsky, H.J. and Brown, R.E. (2006) Assessing the (a)symmetry of concentration-effect curves: empirical versus mechanistic models. BMC Bioinformatics 9:7–123.

Renshaw, E. (1991). Modeling Biological Populations in Space and Time. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511624094.

Richards, F.J. (1959) A flexible growth function for empirical use. Journal of Experimental Botany, 10(2): 290–301. (https://doi.org/10.1093/jxb/10.2.290).

Vandermeer, J. (2010) How Populations Grow: The Exponential and Logistic Equations. Nature Education Knowledge 3(10):15.