Revista Colombiana de Estadística - Applied Statistics
January 2025, volume 48, issue 1, pp. 195–213
http://doi.org/10.15446/rce.v48n1.115842

Defective Survival Modeling and Cure Rate Analysis of COVID-19: A Cross-Location Comparative Study Using Parametric and Non-Parametric Approaches with Demographic Insights

Modelado de supervivencia defectuosa y análisis de la tasa de curación del COVID-19: un estudio comparativo entre localizaciones usando enfoques paramétricos y no paramétricos con perspectivas demográficas

Tasnime Hamdeni[1,a], Mohamed Toumi Nasri[2,b], Frederick Tshibasu[3,c], Rihab Loukil[4,d], Soufiane Gasmi[4,e]

[1]Statistics Department, Higher School of Statistics and Information Analysis (ESSAI), University of Carthage, Tunis, Tunisia

[2]LMPE, National Higher School of Engineers of Tunis, University of Tunis, Tunis, Tunisia

[3]Division of Diagnostic Imaging, University Hospital of Kinshasa, School of Medicine, University of Kinshasa, Kinshasa, Democratic Republic of the Congo

[4]Department of Mathematics, National Higher School of Engineers of Tunis, University of Tunis, Tunis, Tunisia

Abstract

The COVID-19 pandemic has inflicted substantial global morbidity and mortality since December 2019. This study endeavors to model the survival and cure rates of COVID-19 patients using advanced defective modeling techniques and leveraging sophisticated machine learning methods to enhance prediction accuracy. We applied a range of statistical approaches—including parametric, semi-parametric, and non-parametric methods—to fit established and novel models to COVID-19 survival data, with a particular focus on the Defective Gompertz Distribution. To our knowledge, this study represents the pioneering use of defective modeling techniques for estimating

[a]Ph.D. E-mail: tasnim.hamdeni@essai.ucar.tn
[b]Ph.D. E-mail: nasri.medtoumi@gmail.com
[c]Ph.D. E-mail: fredtshibasu@gmail.com
[d]M.Eng. E-mail: rihabloukil01@gmail.com
[e]Ph.D. E-mail: soufiane.kasmi3@gmail.com

cure rates in COVID-19 research. Furthermore, we conducted a comparative analysis across different locations and countries using geographical and demographic data from our dataset. This exploration aimed to uncover variations in survival and cure rates influenced by factors such as socioeconomic status (SES), urban versus rural residence, and healthcare accessibility. Our findings revealed significant disparities in survival and cure rates associated with demographic variables such as age, gender, SES, urbanicity, and healthcare access. Additionally, the study assessed the impact of various public health interventions and identified best practices implemented by different countries. Overall, our results contribute valuable insights to ongoing efforts aimed at comprehending and mitigating the impact of COVID-19 through robust statistical and machine learning modeling techniques. These findings are crucial for informing public health policies and interventions worldwide.

***Key words*:** Cure rate; Cross-location; Defective modeling; Survival analysis.

Resumen

La pandemia de COVID-19 ha causado una morbilidad y mortalidad sustancial a nivel global desde diciembre de 2019. Este estudio tiene como objetivo modelar las tasas de supervivencia y curación de pacientes con COVID-19 utilizando técnicas avanzadas de modelado defectuoso y métodos sofisticados de aprendizaje automático para mejorar la precisión de las predicciones. Aplicamos una variedad de enfoques estadísticos, incluyendo métodos paramétricos, semiparamétricos y no paramétricos, para ajustar modelos establecidos y novedosos a los datos de supervivencia del COVID-19, con un enfoque particular en la Distribución de Gompertz Defectuosa.

Según nuestro conocimiento, este estudio representa el uso pionero de técnicas de modelado defectuoso para estimar tasas de curación en investigaciones relacionadas con COVID-19. Además, realizamos un análisis comparativo entre diferentes ubicaciones y países utilizando datos geográficos y demográficos de nuestro conjunto de datos. Esta exploración buscó identificar variaciones en las tasas de supervivencia y curación influenciadas por factores como el nivel socioeconómico (NSE), la residencia urbana frente a rural y el acceso a la atención médica.

Nuestros hallazgos revelaron disparidades significativas en las tasas de supervivencia y curación asociadas con variables demográficas como la edad, el género, el NSE, la urbanización y el acceso a los servicios de salud. Adicionalmente, el estudio evaluó el impacto de diversas intervenciones de salud pública e identificó mejores prácticas implementadas por diferentes países. En general, nuestros resultados aportan información valiosa a los esfuerzos en curso para comprender y mitigar el impacto del COVID-19 mediante técnicas sólidas de modelado estadístico y aprendizaje automático. Estos hallazgos son cruciales para informar políticas e intervenciones de salud pública a nivel mundial.

***Palabras clave*:** Análisis de supervivencia; Comparación entre ubicaciones; Modelado defectuoso; Tasa de curación.

1. Introduction

In a number of patients, COVID-19 is deadly, with survival outcomes influenced by factors such as age, gender, underlying medical conditions, and biomarker levels. For example, an analysis by Ruan et al. (2020), of 150 COVID-19 patients from Wuhan, China, identified age and pre-existing cardiovascular conditions as significant predictors of mortality, underscoring the impact of patient characteristics on survival rates.

Kundu et al. (2021) investigated the variability in survivorship of coronavirus patients according to age group and sex. Multilevel mixed-effects survival models, along with Kaplan-Meier and the Cox proportional hazard model, were utilized. It should be noted that even though the population studied in Kundu et al. (2021) is restricted to patients in India, the obtained results are in total harmony with the results obtained by the present work.

The cured fraction of COVID-19 patients is estimated using the proportional hazards mixture cure model by Sreedevi and Sankaran (Sreedevi & Sankaran, 2021), and the effect of covariates such as gender and age on lifetime is also considered. Interesting results are reported by Liu et al. (2021), where the authors provided a comparative study between statistical models and their performance in describing COVID-19 data.

Descriptive statistics showed that the cure rate from COVID-19 disease is lower for elderly patients and those with pre-existing health burdens. For instance, studies have shown that age and underlying health conditions significantly affect both hospitalization duration and recovery outcomes in COVID-19 patients (Zhao et al., 2020). Furthermore, cure rate estimations have highlighted the need for targeted healthcare approaches for vulnerable populations (Diao et al., 2020). Yet, inferential statistics remain essential, extending beyond immediate data to provide broader insights and generalizable conclusions that guide public health strategies.

To the best of our knowledge, this study is the first to utilize defective modeling techniques for cure rate estimation. Defective modeling is a contemporary concept. It is used to describe survival data with a proportion of survivors (also called cure rate, cured fraction, long-term survivors, and proportion of immune). Inherently, the survival function of a distribution converges to zero as time goes by. The medical interpretation of this fact is that all the patients in the study are susceptible to the event of interest, which is usually death. This definition does not consider the existence of a proportion of survivors. To consider this proportion, (Haybittle, 1959) put forward the idea of changing the domain of the model's unknown parameters. The modification is such that the survival function no longer converges to zero as time approaches infinity. Instead, it converges to the cure rate. Thus, the cure rate is modeled without having to add an extra parameter like traditional cure rate models such as mixture models (Boag, 1949) and non-mixture models (Lambert et al., 2010). Besides, with defective models, there is no need to assume the existence of a cure fraction since it can be concluded from the range of the estimated parameters of the model.

A few defective models have been introduced in recent years, aimed at enhancing survival analysis by accounting for long-term survivors and populations with a surviving fraction. For instance, (Hamdeni & Gasmi, 2020) developed the Marshall–Olkin generalized defective Gompertz distribution, which provides flexibility in modeling survival data where a portion of the population does not experience the event. Similarly, (Hamdeni & Gasmi, 2022) proposed a proportional-hazards model specifically designed for datasets with long-term survivors, as demonstrated in their application to amyotrophic lateral sclerosis (ALS) data. These models contribute significantly to the field by offering improved approaches to capture complex survival patterns and support accurate prognosis in chronic and terminal diseases. Our goal from this manuscript is to provide a clear and accessible overview of the defective modeling concept that caters not only to experts in the field but also to readers from diverse backgrounds. By elucidating these concepts, we aim to facilitate a deeper understanding and broader applicability of these models beyond their traditional domain.

Although the most statistically significant characteristics of COVID-19 include vaccination status, genetic factors, behavioral and lifestyle aspects (such as physical activities and smoking), and comorbidities like chronic respiratory diseases (Hamdeni et al., 2024), in this study, we aim to explore and highlight the contributions of less significant features.

Geographic disparities play a significant role in COVID-19 outcomes, as evidenced by a population-based study comparing urban and rural areas in Germany and Italy (Assche et al., 2024). This study found notable differences in hospitalization and mortality rates, with rural populations often facing higher mortality despite lower population density. Contributing factors include limited healthcare access, differences in demographic composition, and variations in public health infrastructure. These findings underscore the importance of accounting for geographic factors in survival analysis and support the need for location-specific strategies to mitigate COVID-19's impact. Such insights are crucial in developing models that address both demographic and geographic influences on survival outcomes.

Therefore, we conducted a cross-location comparative study using location and country information within the dataset to explore variations in survival and cure rates across different regions. Socioeconomic status (SES), urban vs. rural residence, and healthcare access were integrated into the analysis to provide a comprehensive understanding of how these factors influence COVID-19 outcomes. We observed significant differences in survival and cure rates influenced by demographic factors such as age, gender, SES, urban vs. rural residence, and healthcare access. The study also evaluates the impact of public health interventions and highlights best practices from different countries.

The rest of the paper is organized as follows: Section 2 we introduce the collected datasets and develop the pre-processing techniques. Section 3 is reserved for a comparative study of well-founded distributions to fit COVID-19 survival data. Section 4 presents the parametric, non-parametric, and semi-parametric approaches to assess the significance of some demographic and locational explanatory variables. Finally, we devote Section 5 to the conclusion.

2. Data Collection and Pre-Processing

In this study, we use data collected by Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) on 1084 cases of COVID-19. Original data sets are publicly available (Johns Hopkins University, n.d.). The dataset includes symptom onset dates for each subject and death dates if the subject is deceased. The authors have converted these dates into durations. The observations were treated as survival data with right-censoring.

Nevertheless, the dataset available from Johns Hopkins University does not include information about location characteristics. Since our study focuses on the impact of location characteristics on COVID-19 survival, we sought additional datasets related to income levels, education, and healthcare access. We found the necessary information in the World Bank Open Data (World Bank, n.d.). By merging both sources, we were able to incorporate these additional features into our analysis.

We preprocessed the additional features, which included various categorical variables, by encoding them into numerical codes to facilitate computational analysis. This encoding process ensures that our data is suitable for statistical analysis and machine learning model training, which are essential for understanding the factors influencing COVID-19 survival rates.

We encoded Socioeconomic Status (SES) with high, medium, and low levels represented as 3, 2, and 1, respectively. This categorization reflects the varying degrees of income, education, and access to resources among individuals. Similarly, the Urban/Rural variable was encoded with urban areas represented as 1 and rural areas as 0. This distinction helps in analyzing the impact of population density, infrastructure, and access to services on COVID-19 outcomes.

Healthcare Access was another crucial variable that we encoded to reflect the level of medical services available to individuals. High access to hospitals and healthcare facilities was encoded as 3, moderate access with limited healthcare workers as 2, and limited access with few healthcare facilities as 1. This encoding allows us to quantitatively assess how different levels of healthcare access affect survival rates.

Encoding these categorical variables into numerical codes offers several advantages. It facilitates statistical analysis by allowing the application of various statistical techniques that require numerical input. Additionally, it enables the use of machine learning models, which generally require numerical data for processing and learning. This numerical representation simplifies data handling and improves computational efficiency, making it easier to manage and analyze large datasets. Furthermore, encoding ensures consistency and standardization, reducing the risk of errors and inconsistencies in data interpretation.

By preprocessing our dataset in this manner, we enhance our ability to perform comprehensive and accurate analyses, thereby gaining deeper insights into the demographic and locational factors that influence COVID-19 survival rates.

3. Optimal Distribution Selection for COVID-19 Survival Analysis

3.1. Overview of Selected Lifetime Distributions

Different distributions have been proposed by statisticians over the years. Here, we select 12 well-founded distributions. We propose to determine the distribution that best represents COVID-19 survival data. Names of the distributions, the corresponding Probability Density Function (PDF), and the range and type of the parameters are given in Table 6 in the Appendix.

It should be noted that the estimation procedure for Gompertz distribution (GD) (Gompertz, 1825) has yielded negative values of the scale parameter β (see details in Subsection 3.4). Therefore, their defective versions Modified Gompertz Distribution (MGD) (Haybittle, 1959) is introduced in Table 6 instead of the proper distributions GD.

The defective version of the Gompertz distribution was originally introduced in 1959 by Haybittle (1959) who has input some slight modifications to the Gompertz distribution so that it models a cure rate in the data. The modification consists in changing the range of the scale parameter from $]0, +\infty[$ to $]-\infty, 0[$. That is to say, if β is the scale parameter, then for the proper Gompertz distribution, β can only have strictly positive values, and for the defective Gompertz distribution, β can only have strictly negative values. This modification of the range of β was made to allow for a cure rate to be considered. We note that the cure rate is the limit of the survival function of the estimated parameters when $t \to \infty$.

MGD was also brought out afterward by Cantor & Shuster (1992). Gieser et al. (1998) has added covariate information to the Modified Gompertz Distribution and used it as a regression model to fit pediatric cancer data.

The probability density function of the MGD is:

$$f(t;\alpha,\beta) = \alpha e^{-\beta t} e^{\frac{\alpha}{\beta}(e^{-\beta t}-1)}$$

with shape parameter α strictly positive and scale parameter $\beta \in \mathbb{R}^*$. The hazart rate function of the MGD is given by:

$$h(t;\alpha,\beta) = \alpha e^{-\beta t}.$$

The survival $S(t;\alpha,\beta)$ of MGD and its corresponding cure rate is given respectively by:

$$\begin{aligned} S(t;\alpha,\beta) &= e^{\frac{\alpha}{\beta}(e^{-\beta t}-1)}, \\ \theta &= \lim_{t\to\infty} S(t) = e^{-\frac{\alpha}{\beta}}, \end{aligned} \tag{1}$$

3.2. Statistical Inference Methodology

The maximum likelihood approach is a prevalent statistical method used to infer the probability distribution parameters for given data. The estimation procedure of the model parameters by the maximum likelihood technique requires:

First, the expression of the likelihood function from the model taking into account the assumptions. Second, applying the logarithm to the likelihood function. Then, determine the parameter values that maximize the log-likelihood function.

In survival analysis, observations are usually censored. The contribution of complete observation to the likelihood function is with the probability density function. Instead, when the observation is censored, the survival function is used to represent the patient who did not experience the event of interest, which is death in this study. In other words, if the patient is dead (during the time of data collection) then the censoring indicator $\delta_i = 1$ (i.e. the patient's contribution to the likelihood is the PDF). Otherwise, the censoring indicator $\delta_i = 0$ (i.e. the patient's contribution to the likelihood is the survival function $S(t)$). Each datum consists of a duration or survival data (in days) (t_i, δ_i) where i is an integer in $[1, n]$, n is the number of patients in the study, t_i the independently observed duration from symptom onset to the death date of the i^{th} subject, δ_i a binary value indicating censorship. Survival of the i^{th} patient is censored at time T_i. We shall consider that the censoring time is fixed to be the last day of data collection. If Θ is the vector of unknown parameters of the model, the likelihood function is given by:

$$L(t_i;\Theta) = \prod_{i=1}^{n} f(t_i;\Theta)^{\delta_i} S(t_i;\Theta)^{1-\delta_i} \tag{2}$$

where $f(t)$ is the PDF and $S(t)$ is the survival function. We apply the logarithm to Equation 2 to get the log-likelihood function:

$$\begin{aligned} l = &\ \ln(\alpha)\sum_{i=1}^{n}\delta_i - \beta\sum_{i=1}^{n}\delta_i t_i + \frac{\alpha}{\beta}\sum_{i=1}^{n}\delta_i(e^{-\beta t_i}-1) \\ &+ \sum_{i=1}^{n}(1-\delta_i)\frac{\alpha}{\beta}\left(e^{-\beta t_i}-1\right). \end{aligned} \tag{3}$$

We derive the obtained function with respect to each of the model parameters in Θ. We obtain 2 non-linear equations, which is equal to the number of parameters in the model:

$$\frac{\partial l}{\partial \alpha} = \frac{1}{\alpha}\sum_{i=1}^{n}\delta_i + \frac{1}{\beta}\sum_{i=1}^{n}\delta_i(e^{-\beta t_i}-1) + \frac{1}{\beta}\sum_{i=1}^{n}(1-\delta_i)\left(e^{-\beta t_i}-1\right).$$

$$\begin{aligned} \frac{\partial l}{\partial \beta} = &-\sum_{i=1}^{n}\delta_i t_i - \frac{\alpha}{\beta^2}\sum_{i=1}^{n}\delta_i(e^{-\beta t_i}-1) - \frac{\alpha}{\beta}\sum_{i=1}^{n}\delta_i t_i e^{-\beta t_i} \\ &-\frac{\alpha}{\beta^2}\sum_{i=1}^{n}(1-\delta_i)\left(e^{-\beta t_i}-1\right) - \frac{\alpha}{\beta}\sum_{i=1}^{n}(1-\delta_i)t_i e^{-\beta t_i}. \end{aligned}$$

To find maximum likelihood estimates, we set the non-linear equations to zero and numerically solve the system of equations using the Newton-Raphson method. The above-described procedure is effectuated to each one of the models introduced earlier.

3.3. Model Selection Criteria and Comparative Evaluation

The relative quality of the fitted model is checked through model selection measures. Five commonly used information criteria as utilized in this paper to compare between fitted models: The AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), CAIC (Consistent Akaike Information Criterion), AICc (corrected AIC), and HQIC (Hannan–Quinn information criterion). These information criteria are not a measure of quality by themselves but they are a tool allowing the comparison between fitted models (Bozdogan, 1987). They are defined as follows:

$$AIC = 2j - 2\ln(L)$$

$$BIC = j\ln(n) - 2\ln(L)$$

$$CAIC = -2\ln(L) + j(\ln(L) + 1)$$

$$AICc = AIC + \tfrac{2(j+1)(j+2)}{n-j-2}$$

$$HQIC = -2L + 2j\ln(\ln(L))$$

where j is the length of Θ or the number of parameters in the model, n is the size of the given data sample and L is the maximized likelihood of the parameter vector Θ.

3.4. Empirical Results and Interpretation

The maximum likelihood estimation results for each of the selected models, the corresponding standard errors, and 95% confidence intervals are presented in Table 1. It should be mentioned that since the shape parameter α only allows positive values, negative values of the lower bound of the 95% confidence interval are replaced by zero.

Table 2 gives the log-likelihood function of the estimated parameters as well as the information criteria values. The preferred model is the one with the smallest information criteria. The AIC, BIC, CAIC, AICc, and HQIC values show that MGD markedly outperformed the other models.

Figure 1 illustrates the Kaplan-Meier estimator for the survival function from the COVID-19 dataset, as well as the survival curves of the defective MGD and MGGD models, respectively. For this type of model, this step allows us to estimate the cure rate in the data, as it is the value towards which the estimated survival function converges. The cure rate estimated by the Modified Gompertz distribution, using Equation 1, is 0.9222.

This estimation is in harmony with the recent literature (Cao et al., 2020). Yet, other factors may also influence the cure rate.

Table 1: Maximum Likelihood Estimates (MLE), the corresponding Standard Error (SE), 95% CI's Lower Bound (LB) and Upper Bound (UB) for each distribution.

Distribution	Parameter	MLE	SE	LB	UB
Weibull	$\hat{\alpha}$	1388.4400	659.159	547.55	3520.7
	$\hat{\beta}$	0.7448	0.0886	0.5899	0.9406
Exponential	$\hat{\alpha}$	540.2380	68.0636	428.2145	703.0437
Burr	$\hat{\alpha}$	6.4544	3.1913	2.4490	17.0107
	$\hat{\beta}$	1.3652	0.2844	0.9074	2.0538
	$\hat{\gamma}$	0.0271	0.0115	0.0117	0.0627
Gamma	$\hat{\alpha}$	0.7388	0.0914	0.5797	0.9416
	$\hat{\beta}$	1667.3700	909.2830	572.5798	$4.8554 \quad 10^3$
Extreme value	$\hat{\alpha}$	112.1063	8.8073	94.8442	129.3684
	$\hat{\beta}$	26.9675	3.0084	21.6711	33.5582
Log-logistic	$\hat{\alpha}$	7.1059	0.4627	6.1989	8.0129
	$\hat{\beta}$	1.3134	0.1553	1.0417	1.6562
Logistic	$\hat{\alpha}$	108.4490	8.5430	91.7048	125.1929
	$\hat{\beta}$	26.1034	2.8974	20.9997	32.4474
Log-normal	$\hat{\alpha}$	8.1139	0.5688	6.9990	9.2289
	$\hat{\beta}$	2.9829	0.3255	2.4085	3.6942
Nakagami	$\hat{\alpha}$	0.6345	0.0437	0.2881	0.4612
	$\hat{\beta}$	$1.2604 \quad 10^6$	$1.1216 \quad 10^6$	$2.2032 \quad 10^5$	$7.2111 \quad 10^5$
Normal	$\hat{\alpha}$	119.753	9.9627	100.2268	139.2799
	$\hat{\beta}$	53.9710	5.7397	43.8163	66.4790
MGD	$\hat{\alpha}$	0.0038	0.0007	0.0024	0.0053
	$\hat{\beta}$	-0.0469	0.0121	-0.0706	-0.0232

Table 2: Negative log-likelihood value L and information criteria for each distribution.

Distribution	L	AIC	BIC	CAIC	AICc	HQIC
Weibull	-455.9490	915.8980	925.8748	927.8748	912.9091	919.6750
Exponential	-459.3970	920.7940	925.7824	926.7824	918.7977	922.6825
Burr	-451.3390	908.6780	923.6432	926.6432	904.7001	914.3435
Gamma	-456.1220	916.2440	926.2208	928.2208	913.2551	920.0210
Extreme value	-503.7201	1011.4400	1021.4168	1023.4168	1008.4511	1015.2170
Log-logistic	-455.6590	915.3180	925.2948	927.2948	912.3291	919.0950
Logistic	-502.963	1009.9260	1019.9028	1021.9028	1006.9371	1013.7030
Log-normal	-453.5080	911.0160	920.9928	922.9928	908.0271	914.7930
Nakagami	-456.24	916.4568	926.4568	928.4568	913.4911	920.2570
Normal	-497.584	999.1680	1009.1448	1011.1448	996.1791	1002.9450
MGD	-450.2200	902.4400	914.4168	916.4168	901.4511	908.2170

4. Impact of Demographic and Cross-Location Covariates on COVID-19 Survival

To visualize the effect of some demographic as well as cross-location covariates on COVID-19 patients' overall survival, parametric, semi-parametric, and non-parametric estimation approaches have been conducted.

The regression model was evaluated using multiple statistical diagnostics to assess its fit, robustness, and adherence to the assumptions of linear regression.

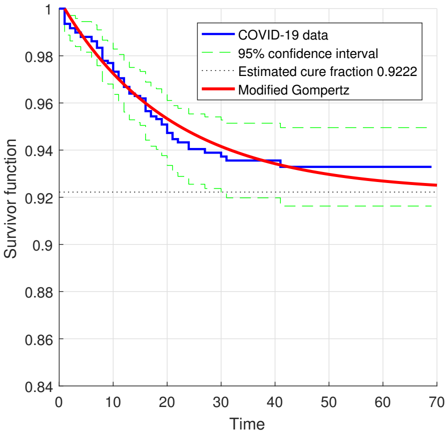

Figure 1: Kaplan-Meier, 95% confidence interval and parametric estimation of the survival function of the MGD.

Skewness measures the asymmetry of the probability distribution of a real-valued random variable. For residuals in a regression model, skewness evaluates the symmetry of the errors around the mean. Mathematically, it is calculated as the third standardized moment:

$$\text{Skewness} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma}\right)^3,$$

where x_i are the residuals, $\bar{x}$ is their mean, and σ is their standard deviation. A skewness value of zero indicates perfect symmetry, while positive skewness indicates a right tail and negative skewness a left tail. In regression diagnostics, significant skewness in residuals suggests that the model's errors are not symmetrically distributed, which can violate the normality assumption and affect inference. Kurtosis measures the "tailedness" of a distribution, specifically the propensity for outliers. It is the fourth standardized moment of a distribution:

$$\text{Kurtosis} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma}\right)^4.$$

We have also employed the Jarque-Bera (JB) test which is a goodness-of-fit test for normality, evaluating whether the sample data have skewness and kurtosis matching a normal distribution. The test statistic is given by:

$$\text{JB} = \frac{n}{6}\left(\text{Skewness}^2 + \frac{(\text{Kurtosis}-3)^2}{4}\right),$$

where n is the sample size. The JB test is asymptotically chi-square distributed with 2 degrees of freedom. We have also calculated the R-squared, or the coefficient of determination, measures the proportion of variance in the dependent variable that is predictable from the independent variables. It is calculated as:

$$R^2 = 1 - \frac{\mathrm{SS}_{\mathrm{res}}}{\mathrm{SS}_{\mathrm{tot}}},$$

where $\mathrm{SS}_{\mathrm{res}}$ is the residual sum of squares, and $\mathrm{SS}_{\mathrm{tot}}$ is the total sum of squares. R-squared ranges from 0 to 1, where a higher value indicates that more of the variance in the dependent variable is explained by the model. However, R-squared alone does not assess the model's accuracy; a high R-squared does not imply causation or guarantee a well-fitted model. Therefore we have also calculated the Adjusted R-squared which is a modified version of R-squared that accounts for the number of predictors in the model. It is defined as:

$$\text{Adjusted } R^2 = 1 - \left(\frac{(1 - R^2)(n - 1)}{n - k - 1} \right),$$

where n is the sample size, and k is the number of predictors. The F-statistic tests the null hypothesis that all regression coefficients are zero, comparing the model to one with no predictors. The F-statistic is calculated as:

$$F = \frac{\text{Explained variance}}{\text{Unexplained variance}},$$

These diagnostics are utilized to collectively provide insights into the model's fit and robustness. The Modified Gompertz Distribution (MGD) that presented the best-fitted model, according to results in Table 2 is used to estimate the cure rate for each demographic stratification: Age (≤ 60, > 60) and gender, as well as cross-location information.

4.1. Semi-Parametric Approach

We conduct a multivariate demographic and locational regression analysis to examine the relationships between various factors and the dependent variable. Tables 3 and 4 summarize the regression results, presenting coefficients, standard errors, t-values, p-values, and confidence intervals for each predictor variable. Additionally, measures of skewness, kurtosis, and model fit (R-squared and adjusted R-squared) are provided to assess the overall goodness-of-fit and statistical significance of the model.

The regression resutls show that age above 60 years old and gender were associated with the overall survival of coronavirus patients. The negative coefficients for the variable gender (-0.8662 and -0.9113) indicate that females have a higher survival rate than males. The positive coefficients (2.0656 and 2.0857) for the variable age indicate that the more the patient is old the higher the risk of death is (and thus the lower the cure rate is). In essence for the demographic characteristics, the Cox proportional hazard regression model showed a significantly higher risk of death in elderly patients and less significantly, male patients.

Table 3: Results of multivariate demographic and locational regression.

	Coef.	Std. err.	t	$P > \|t\|$	[0.025	0.975]
const	34.8218	0.738	47.153	0.000	33.371	36.272
age	2.0656	0.3004	6.8768	6.1198×10^{-12}	1.478	2.654
gender	-0.8662	0.3069	-2.8229	0.0048	-1.468	-0.265
SES	-2.2709	1.032	-0.953	0.341	-6.946	2.407
Urban/Rural	-0.4737	0.973	-0.487	0.627	-2.385	1.438
Healthcare Access	-2.6398	2.792	-0.946	0.345	-8.123	2.844
Skew	-0.110			Prob (JB)	0.634	
Kurtosis	2.513			Prob (F-statistic)	7.49×10^{-22}	
R-squared	0.144			Adj. R-squared	0.138	

Now focusing on the specifics of each table. Table 3 displays the results of the linear regression, taking into account multivariate demographic and locational coefficients. Although in this study we aim to explore and highlight the contributions of less significant features, the obtained p-value (F-statisic) is extremely low, indicating that the model is statistically significant and that at least one of the predictors is significantly related to the dependent variable.

This obtained R-squared value indicates that approximately 14.4% of the variance in the dependent variable (Survival Days) is explained by the independent variables in the model. This relatively low value suggests that the model explains only a small portion of the variability in survival days. This is consistent with our earlier discussion regarding the selection of less significant characteristics for the study.

The weaknesses of the model are apparent in some of the obtained results. For example, the coeffecient for the SES feature indicate that for each one-unit increase in socioeconomic status (SES), Survival Days decrease by approximately 4.1415 days. This effect is not statistically significant (p-value = 0.308) and contradicts previous studies. Added to that, being in an urban or rural area has a coefficient of -0.8801, which is also not statistically significant (p-value = 0.809). Healthcare Access coefficient value shows that for each one-unit increase in healthcare access, Survival Days decrease by approximately 5.0788 days. This effect is not statistically significant (p-value = 0.233).

Furthermore, the model suffers from non-normality in its residuals, as indicated by the Jarque-Bera (JB) Test value of 7.515 with a significant p-value (0.0233), suggesting that the residuals are not normally distributed. Specifically, the Skew value is -0.155, indicating a slight left skew in the residuals, despite the Kurtosis value of 2.600, which is close to 3, the kurtosis of a normal distribution.

Another issue with the model is that the condition number is higher than 10, indicating moderate multicollinearity (see Figure 2). In conclusion, the diagnostics suggest issues with normality and multicollinearity. This suggests that further model refinement and exploration of additional variables or transformations may be necessary to improve the model. Therefore, we have opted for Principal Component Analysis (PCA) to attempt to address these issues.

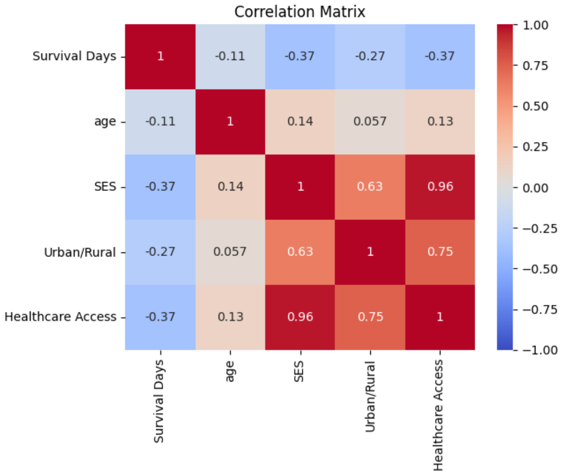

FIGURE 2: Correlation matrix of the model features.

The results after applying Principle Component Analysis are shown in Table 4.

TABLE 4: Results after applying Principal Component Analysis.

	Coef.	Std. err.	t	$P > \|t\|$	[0.025	0.975]
const	35.2900	0.659	53.531	0.000	33.996	36.584
PC1	3.1486	0.310	10.152	0.000	2.540	3.757
age	2.0857	0.3006	6.9406	3.9775×10^{-12}	1.496	2.676
gender	-0.9113	0.3070	-2.9717	0.0030	-1.516	-0.307
Skew	-0.110			Prob (JB)	0.00808	
Kurtosis	2.471			Prob (F-statistic)	7.49×10^{-22}	
R-squared	0.144			Adj. R-squared	0.139	

The R-squared remains the same in the previous model. So the model significance wasn't reduced after the PCA step. The p-value associated with the F-statistic remains extremely low, indicating that the model is statistically significant.

The first principal component (PC1) is highly significant with a p-value less than 0.001, suggesting it is a strong predictor in the model. The coefficients for age and gender didn't change much. A p-value of 0.00808 for the JB test indicates that we reject the null hypothesis of normality at typical significance levels (0.05), suggesting that the data significantly departs from a normal distribution. Additionally, the condition number of 3.61 is significantly lower than in the previous model, indicating that multicollinearity has been reduced.

The improvement in the model through PCA suggests that the multicollinearity issue has been addressed. The F-statistic and associated p-value indicate that the overall model fit has improved. The inclusion of a principal component (PC1) in the model provides statistically significant predictors, unlike the original autocorrelated variables.

Overall, the results suggest an improvement in the model through PCA, making the regression coefficients more stable and reliable.

4.2. Non-Parametric and Parametric Approaches

Further exploration of the model is conducted in this final section. In Figures 3 and 4, non-parametric Kaplan–Meier and parametric survival curves of overall survival function according to the gender and the age of the coronavirus patients. The parametric curves are based on the Modified Gompertz Distribution, with parameters estimated by the maximum likelihood approach. As the figures show, the survival curves reached a stable plateau at the right tail. Interestingly, it is recommended then to use cure rate models because they lead to more accurate results (Kim et al., 2013). The plateau of the Kaplan-Meier curve is lower for the male population and the elderly population of COVID-19 data and, hence, the associated estimated cure rate.

Maximum likelihood estimation of MGD parameters that allowed the estimated survival curves for each considered stratification are depicted in Figures 3 and 4 are given in Table 5. The corresponding standard errors and 95% confidence intervals are also shown in the same table. The cure rate Θ_{est} for each subpopulation is estimated based on the Modified Gompertz distribution, using Equation 1. Empirical cure rate Θ_{emp} is naively calculated as the proportion of patients who are still alive, or the right-censoring level in the sub dataset, which is also the survival rate plateau. The estimated and empirical cure rates have close values.

Kaplan-Meier and the parametric survival analysis demonstrated that female patients and those who are younger than 60 years old have a significantly higher chance of survival.

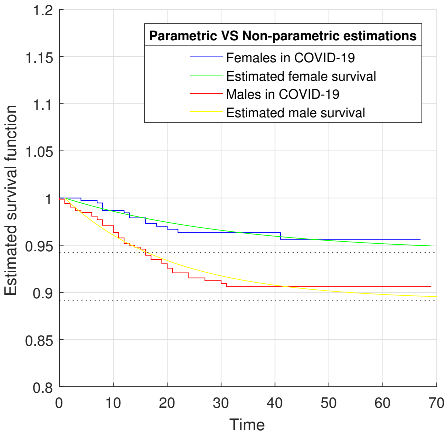

FIGURE 3: Example of non-parametric Kaplan–Meier and parametric curves of overall survival function according to gender.

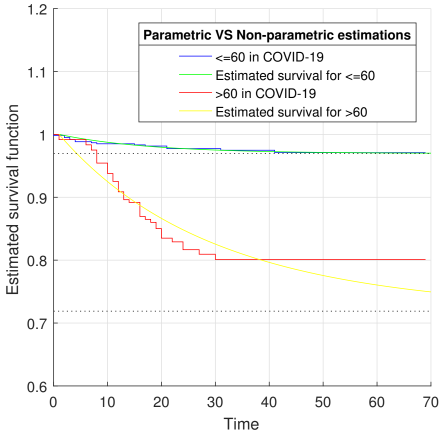

FIGURE 4: Example of non-parametric Kaplan–Meier and parametric curves of overall survival function according to age.

TABLE 5: MLE of MGD parameters, the corresponding SE and 95 % CI, empirical and estimated cure rates for each of the considered stratifications.

	Age		Gender	
	≤ 60	> 60	*Female*	*Male*
$\hat{\alpha}$	0.0018	0.0099	0.0018	0.0055
$SE_{\hat{\alpha}}$	0.0007	0.0024	0.0008	0.0013
$95\%CI_{\hat{\alpha}}$	$[0.0004, 0.0031]$	$[0.0053, 0.0143]$	$[0.0003, 0.0034]$	$[0.0030, 0.0080]$
$\hat{\beta}$	-0.0583	-0.0300	-0.0301	-0.0480
$SE_{\hat{\beta}}$	0.0253	0.0136	0.0229	0.0141
$95\%CI_{\hat{\beta}}$	$[-0.1079, -0.0088]$	$[-0.0568, -0.0033]$	$[-0.0751, 0.0148]$	$[-0.0755, -0.0204]$
Θ_{emp}	0.9751	0.8216	0.9634	0.9154
Θ_{est}	0.9696	0.7189	0.9420	0.8917

5. Conclusion

This study has comprehensively examined various survival and cure rate models for COVID-19, employing a combination of parametric, semi-parametric, and non-parametric techniques. Our analysis, anchored in defective modeling and utilizing data from diverse geographic locations, underscores the significant variation in survival and cure rates across different regions. By integrating data from multiple sources, we enhanced the robustness and applicability of our findings, providing a cross-location comparative perspective.

The defective Modified Gompertz Distribution emerged as the superior model for our data, demonstrating its effectiveness in fitting the survival rates of COVID-19 patients while accommodating variations due to demographic factors such as age and gender. The inclusion of cross-location data allowed for a richer analysis and helped identify regional differences that might influence survival outcomes and public health strategies. Applying Principal Component Analysis was beneficial in this context as it helped address multicollinearity, reduced dimensionality, and improved the stability and reliability of the regression coefficients by transforming the original variables into a set of uncorrelated principal components.

These insights pave the way for future research to further dissect the impact of demographic, environmental, and healthcare access variables on COVID-19 outcomes. Additionally, the methodologies applied in this study can be adapted to other epidemiological datasets, potentially offering a valuable tool for pandemic response and preparedness efforts worldwide.

6. Overview of the Distributions Used

Table 6 provides the names of the distributions, their corresponding Probability Density Functions (PDFs), and the range and type of their parameters. Figure 6 presents the plot of the Kaplan-Meier non-parametric estimator curve as well as the survival curves of all the fitted parametric models. The closer the parametric model is to the Kaplan-Meier curve, the better the fit.

Table 6: Probability density functions of the studied distributions.

Distribution	PDF	Parameters
Weibull	$f(t) = \frac{\alpha}{\beta}(\frac{t}{\beta})^{\alpha-1}e^{-(\frac{t}{\beta})^{\alpha}}$	$\alpha > 0$: shape, $\beta > 0$: scale
Exponential	$f(t) = \alpha e^{-\alpha t}$	$\alpha > 0$: rate
Burr	$f(t) = \frac{\alpha\beta}{\gamma}(\frac{t}{\gamma})^{\alpha-1}(1+(\frac{t}{\gamma})^{\alpha})^{-\beta-1}$	$\alpha > 0$: scale, $\beta, \gamma > 0$: shape
Gamma	$f(t) = \frac{\beta^{\alpha}t^{\alpha-1}e^{-\beta t}}{(\alpha-1)!}$	$\alpha > 0$: shape, $\beta > 0$: rate
Extreme value	$f(t) = \frac{1}{\beta}e^{\frac{t-\alpha}{\beta}}e^{-e^{\frac{t-\alpha}{\beta}}}$	$\alpha > 0$: location, $\beta > 0$: scale
Log-logistic	$f(t) = \frac{(\frac{\beta}{\alpha})(\frac{t}{\alpha})^{\beta-1}}{(1+(\frac{t}{\alpha})^{\beta})^2}$	$\beta > 0$: shape, $\alpha > 0$: scale
Logistic	$f(t) = \frac{e^{-(\frac{t-\alpha}{\beta})}}{\beta(1+e^{-\frac{t-\alpha}{\beta}})^2}$	$\alpha > 0$: location, $\beta > 0$: scale
Log-normal	$f(t) = \frac{1}{t}\frac{1}{\beta\sqrt{2\pi}}e^{-\frac{1}{2\beta^2}(\ln t-\alpha)^2}$	$\alpha \in R$: location, $\beta > 0$: scale
Nakagami	$f(t) = \frac{2\alpha^{\alpha}}{(\alpha-1)!\beta^{\alpha}}t^{2\alpha-1}e^{-\frac{\alpha}{\beta}t^2}$	$\alpha \geq \frac{1}{2}$: shape, $\beta > 0$: scale
Normal	$f(t) = \frac{1}{\beta\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{1}{\beta}(t-\alpha))^2}$	$\alpha \in R$: location, $\beta > 0$: scale
MGD	$f(t) = \alpha e^{\beta t}e^{-\frac{\alpha}{\beta}(e^{\beta t}-1)}$	$\alpha > 0$: shape, $\beta < 0$: scale

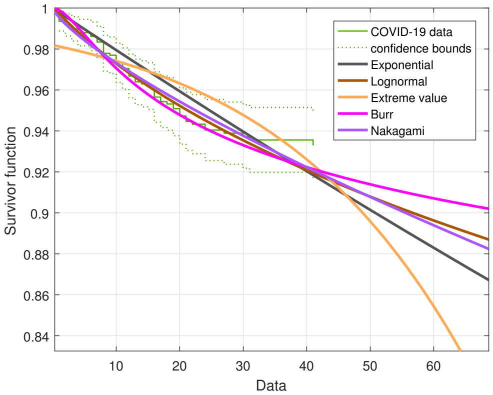

FIGURE 5: Kaplan-Meier and parametric estimation of the survival function of some proper distributions for COVID-19 data (Part 1).

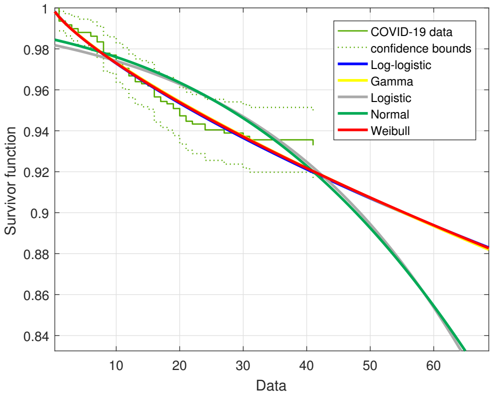

FIGURE 6: Kaplan-Meier and parametric estimation of the survival function of some proper distributions for COVID-19 data (Part 2).

[Received: February 2024 — Accepted: June 2024]

References

Assche, S. B.-V., Ferraccioli, F., Riccetti, N., Gomez-Ramirez, J., Ghio, D. & Stilianakis, N. I. (2024), 'Urban-rural disparities in COVID-19 hospitalisations and mortality: A population-based study on national surveillance data from Germany and Italy', *Plos one* **19**(5), e0301325.

Boag, J. W. (1949), 'Maximum likelihood estimates of the proportion of patients cured by cancer therapy', *Journal of the Royal Statistical Society. Series B (Methodological)* **11**(1), 15–53. https://www.jstor.org/stable/2983684

Bozdogan, H. (1987), 'Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions', *Psychometrika* **52**(3), 345–370.

Cantor, A. B. & Shuster, J. J. (1992), 'Parametric versus non-parametric methods for estimating cure rates based on censored survival data', *Statistics in Medicine* **11**(7), 931–937.

Cao, L., Huang, T.-t., Zhang, J.-x., Qin, Q., Liu, S.-y., Xue, H.-m., Gong, Y.-x., Ning, C.-h., Shen, X.-t., Yang, J.-x. et al. (2020), 'Estimation of instant case fatality rate of COVID-19 in Wuhan and Hubei based on daily case notification data', *medRxiv* .

Diao, Y., Liu, X., Wang, T., Zeng, X., Dong, C., Zhou, C., Zhang, Y., She, X., Liu, D. & Hu, Z. (2020), 'Estimating the cure rate and case fatality rate of the ongoing epidemic COVID-19', *medRxiv* . https://www.medrxiv.org/content/10.1101/2020.02.26.20028189v1

Gieser, P. W., Chang, M. N., Rao, P., Shuster, J. J. & Pullen, J. (1998), 'Modelling cure rates using the Gompertz model with covariate information', *Statistics in Medicine* **17**(8), 831–839.

Gompertz, B. (1825), 'On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. In a letter to Francis Baily, Esq. FRS & c', *Philosophical Transactions of the Royal Society of London* **115**, 513–583.

Hamdeni, T., Frederick, T., Asma, K. & Soufiane, G. (2024), 'CRP, PCT, and D-dimer as Biomarkers for Disease Severity in COVID-19 Patients: A Retrospective Study in Kinshasa, Democratic Republic of Congo', *Journal of Biostatistics and Epidemiology* .

Hamdeni, T. & Gasmi, S. (2020), 'The Marshall–Olkin generalized defective Gompertz distribution for surviving fraction modeling', *Communications in Statistics-Simulation and Computation* pp. 1–14.

Hamdeni, T. & Gasmi, S. (2022), 'A proportional-hazards model for survival analysis and long-term survivors modeling: Application to amyotrophic lateral sclerosis data', *Journal of Applied Statistics* **49**(3), 694–708.

Haybittle, J. (1959), 'The estimation of the proportion of patients cured after treatment for cancer of the breast', *The British Journal of Radiology* **32**(383), 725–733.

Johns Hopkins University (n.d.), 'COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE)', https://github.com/CSSEGISandData/COVID-19. Accessed: June 2024.

Kim, S., Zeng, D., Li, Y. & Spiegelman, D. (2013), 'Joint modeling of longitudinal and cure-survival data', *Journal of statistical theory and practice* **7**(2), 324–344.

Kundu, S., Chauhan, K., Mandal, D. et al. (2021), ‘Survival Analysis of Patients With COVID-19 in India by Demographic Factors: Quantitative Study’, *JMIR Formative Research* **5**(5), e23251. https://formative.jmir.org/2021/5/e23251/

Lambert, P. C., Dickman, P. W., Weston, C. L. & Thompson, J. R. (2010), ‘Estimating the cure fraction in population-based cancer studies by using finite mixture models’, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **59**(1), 35–55.

Liu, X., Ahmad, Z., Gemeay, A. M., Abdulrahman, A. T., Hafez, E. H. & Khalil, N. (2021), ‘Modeling the survival times of the COVID-19 patients with a new statistical model: A case study from China’, *PLOS One* **16**(7), e0254999. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0254999

Ruan, Q., Yang, K., Wang, W., Jiang, L. & Song, J. (2020), ‘Clinical predictors of mortality due to COVID-19 based on an analysis of data of 150 patients from Wuhan, China’, *Intensive Care Medicine* pp. 1–3. https://link.springer.com/article/10.1007/s00134-020-05991-x

Sreedevi, E. P. & Sankaran, P. G. (2021), ‘Statistical methods for estimating cure fraction of COVID-19 patients in India’, *Model Assisted Statistics and Applications* **16**(1), 59–64. https://content.iospress.com/articles/model-assisted-statistics-and-applications/mas210508

World Bank (n.d.), ‘World Bank Open Data’, https://data.worldbank.org. Accessed: June 2024.

Zhao, W., Yu, S., Zha, X., Wang, N., Pang, Q., Li, T. & Li, A. (2020), ‘Clinical characteristics and durations of hospitalized patients with COVID-19 in Beijing: a retrospective cohort study’, *medRxiv* . https://www.medrxiv.org/content/10.1101/2020.03.13.20035498v1