AJMAS
https://journal.utripoli.edu.ly/index.php/Alqalam/index eISSN 2707-7179

Original article

Applying Time Series Analysis (Box-Jenkins) to Predict the Number of Patients with Cancer at Benghazi Medical Center

Kamilah Othman[1]*, Naeima Abdelati [2], Sara Al warrad[1]

[1]*Department of Statistic, Faculty of Arts and Science Al–Marj, University of Benghazi, Benghazi, Libya.*
[2]*Department of Statistic, Faculty of Science, University of Benghazi, Benghazi, Libya*

ARTICLE INFO

Corresponding Email. *:kamlah.alabd@uob.edu.ly*

Received*: 05-11-2023*
Accepted*: 29-11-2023*
Published*: 30-11-2023*

Keywords*. Time Series Analysis, Box-Jenkins Methodology, ARIMA Models, Forecasting, Cancer.*

ABSTRACT

Background and objective. *Cancer has a negative impact on human health in the world, leading to illness and death. This study designed to apply time series analysis using Box-Jenkins methodology to describe the behavior and find an appropriate model for predicting the time series of patients with cancer in Benghazi City.* ***Methods.*** *The data were collected from the medical records of the Oncology department at Benghazi Medical Center (BMC) from January 2011 to December 2022, of 11266 patients distributed on 144 time series of observations.* ***Results.*** *The study found that the time series of data was non-stationary and it had a fluctuation pattern around the mean that changed into a stationary series by performing the first difference. We generated several ARIMA models and compared them based on different criteria, including the R-squared=0.423, root mean square error (RAMSE) =14.045 and Bayesian information criterion (BIC) = 5.447. The suitable model chosen to represent the data series was ARIMA (2,1,3) which, also used for predicting new cases in the next four years. As a result, the estimated model established a similarity between the predicted values and the real values of the time series. In addition, the results indicated that a progressive increase in the total number of people with cancer from January 2023 to December 2026, reaching up to 5344 patients.* ***Conclusion.*** *The ARIMA (2,1,3) model is a good tool for predicting the number of patients with cancer in Benghazi Medical Center. Finally, the importance of this study depending on results to raise awareness and knowledge of risks the cancer in the Libyan community.*

Cite this article. *Othman K, Abdelati N, Al warred S. Applying Time Series Analysis (Box-Jenkins) to Predict the Number of Patients with Cancer at Benghazi Medical Center. Alq J Med App Sci. 2023;6(2):756-765.*
https://doi.org/10.5281/zenodo.10225164

INTRODUCTION

Cancer has a negative impact on human health in the world, leading to illness and death. In 134 out of 183 countries, it is the primary or secondary cause of premature mortality. By 2040, the estimates indicate that there will be over a million new cases each year. Especially underdeveloped countries will be affected by this increase [1]. However, "Libya is not an exception", after cardiovascular disease, cancer ranks as the second-leading cause of death in Libya. Its situation become more complicated because of the fragility of the healthcare system since political conditions [2]. Moreover, Libya's contribution to this data is unreliable due to the absence of a comprehensive national or local cancer registry that can accurately track cancer incidence, types, disease, and mortality in the country [1]. According to 2015 reports based on the Benghazi City Cancer (BMC) Registry, there was an age-standardized incidence of all site cancers in males 135.4 and 107.1 of females per 100,00 [3]. As a result of this, according to the rise in the number of people who suffered cancer in Libya, especially since the civilian war in 2011[4], it requires the application of the time series predicting

methods, including the ARIMA model, to forecast future cancer development and study the trends of time series of disease. The methods of time series forecasting depend on historical data analysis, which assumes that patterns in the data can be used to predict future events [5]. Autoregressive integrated moving average (ARIMA) models are methods of time series modeling that have become increasingly popular recently in the healthcare sector and disease events prediction [6]. Also, its primary objective is to thoroughly analyze previous observations of a time series and generate an appropriate model that can predict future values for the series [5].

Time series models have been applied in several study papers to forecast cancer and some diseases. The Box-Jenkins methodology (1976), which is founded on studies dealing with forecasting cancer using ARIMA models has been widely used to predict the future depending on variables of any disease [7]. Duong et al., studied the spread of Coronavirus disease (COVID-19) in Vietnam using daily reports data obtained from the World Health Organization between 21st January and 16th March 2020. The results found that the ARIMA (1,2,1) model is capable of identifying and estimating the total number of new cases in the region and making it a powerful tool for analyzing the epidemiological progress of the disease globally [8]. Earnest et al., used data from Australian patients with prostate cancer, collected by the Institute of Health and Welfare, from 1982 to 2013 to forecast the incidence and mortality rates for the future period of 2014 to 2022. The findings revealed that the ARIMA (1,1,0) predicted an increase in incidence cases of patients with prostate cancer to 25283 by 2022 and the forecasted model had a higher level of accuracy in prediction [9].

An earlier used monthly data on the brucellosis epidemic in the City of Jinzhou, Chain was analyzed for the period from 2004 to 2013. The research aimed to predict the monthly incidence of brucellosis in 2015 and examine the specific features and conditions of the disease in 3078 patients. The research findings showed that the ARIMA $(0,1,2)(1,1,1)_{12}$ model was appropriate for forecasting the brucellosis incidence in Jinzhou [10]. Wei et al., used ARIMA to predict the incidence of hepatitis epidemic data between January 2005 and December 2012 from the Heng County City of Disease Center for Control and Prevention. The study showed that the ARIMA $(0,1,2)$ $(1,1,1)_{12}$ was the most suitable, and the combined model of ARIMA-GRNN revealed better forecasting of hepatitis incidence in Heng County City, China, whereas the GRNN model was unsuitable as fitting for disease [11]. A study performed by Pan et al., applied real data from the Disease Control Center and Prevention (CDC) between January and August 2014. The study utilized ARIMA models to improve the predictive accuracy. The study found that the nominated model achieved 92.1% of prediction accuracy in the CDC [12].

Another study employed data from a health center that relied on 30 mammogram images of patients with breast cancer in India. The researchers advanced a method called CDC (Computer-Aided Diagnosis) that created EIT (Electrical Impedance Tomography) to identify and categorize breast cancer. The study found that the ARIMA models achieved better accuracy and sensitivity in detection [13]. Moreover, analyzed onchocerciasis outbreak data from 1988 to 2011 from two regions Chiapas and Oaxaca in Mexico to predict the pattern of onchocerciasis in the period between 2012 and 2013. Based on the findings, the ARIMA $(1,1,1)$ $(1,0,1)_{12}$ mixed models predicted an enormously low level of onchocerciasis cases for the next two years in Mexico [14].

One of the purposes of reveal the effect of ARIMA models on the incidence of cancer development prediction in the future and discuss the advantages and potential challenges related to the applications. Through this study, we will highlight the main study objectives. Firstly, describe and study the pattern of ARIMA models of patients with cancer in Benghazi Medical Center (BMC). Next, find an optimal model for the number of patients with cancer in BMC. Lastly predicting the number of new cancer cases at BMC for the next 4 years (from 2023 to 2026 per month) based on the period from 2011 to 2022 monthly.

METHODS

Study design and setting

The study methodology relied on the theoretical side, which focused on Boxes and Jenkins models in time series analysis. This involved diagnosis, estimation, testing the suitability of the diagnosed model, and predicting future trends. The theoretical side was supported by the application of real data on the number of people with cancer from the Benghazi Medical Center to obtain the best model for predicting the number of cases of cancer in the future. The study presents significant results, discussion, conclusion, and recommendations, as well as sources.

Data collection procedure

The data were collected from medical records of the Oncology department at (BMC) Benghazi Medical Center, between January 1st, 2011, and December 30th, 2022. The dataset consists of 11266 patients distributed on 144 time series of observations and satisfies the Box-Jenkins method of time series forecasting. This method recommended that the number of observations should be more than 100 observations [15]. Based on the data, we suggest a suitable ARIMA model and use it to forecast new cases for the next 4 years, from January 2023 to December 2026 [16].

Autoregressive integrated Moving-Average (ARIMA) Method

The autoregressive integrated moving average model reflects a variable as the weighted average of its previous values. ARMA is a grouping of autoregressive (AR), integrated (I), and moving average (MA), however, most are combinations of AR and MA. ARIMA models are generally useful for data series of stationary where mean function, variance, and autocorrelation stay unchanged over time [17][18]. The static process called the autoregressive process, involves the dependent variable Z_t being expressed as a function of the previous values of the dependent variable. The AR method of p order is given by: $Z_t = \psi + \alpha_1 Z_{t-1} + \alpha_2 Z_{t-2} + \cdots + \alpha_p Z_{t-p} + \xi_t$ (1)

In case of the ψ is equal to zero, $\Rightarrow Z_t = \alpha_1 Z_{t-1} + \alpha_2 Z_{t-2} + \cdots + \alpha_p Z_{t-p} + \xi_t$ (2)

The Z_t represent the static depend variable with time (t), $Z_{t-1}, \dots, Z_{t-p}$ are depend variables with time series at $t-1, \dots, t-p$. The ξ_t is white noise satisfies: $\xi_t \sim N(0, \sigma^2)$, $\psi = \mu(1 - \alpha_1 -, \dots - \alpha_p)$ then $\alpha_1, \dots, \alpha_p$ indicated to the AR parameters model. Also from equation (2) through applying the operator of Lag, as a result of this: $(1 - \alpha_1 \pi - \cdots - \alpha_p \pi^p) Z_t = \xi_t$, then $\alpha(\pi) Z_t = \xi_t$

The moving-average method is a static process conveyed as random errors in previous values. The MA method of q order is given by:

$Z_t = \mu + \xi_t - \Omega_1 \xi_{t-1} + \Omega_2 \xi_{t-2} - \cdots - \Omega_q \xi_{t-q}$ (3)

In case of the ψ is equal to zero, $\Rightarrow Z_t = \xi_t - \Omega_1 \xi_{t-1} + \Omega_2 \xi_{t-2} - \cdots - \Omega_q \xi_{t-q}$ (4)

Also from equation (4) through applying the operator of the backshift as a result of this:

$\mu + (1 - \Omega_q \pi - \cdots - \Omega_q \pi^q) \xi_t = Z_t$ then $\mu + \Omega(\pi) \xi_t = Z_t$

The ξ_t is white noise and $\Omega_1, \dots, \Omega_q$ indicate the weights of MA.

The grouping between AR and MA is the result of that grouping ARMA(p, q), and the resulting model is:

$Z_t = \psi + \alpha_1 Z_{t-1} + \alpha_2 Z_{t-2} + \cdots + \alpha_p Z_{t-p} + \xi_t - \Omega_1 \xi_{t-1} + \Omega_2 \xi_{t-2} - \cdots - \Omega_q \xi_{t-q}$ (5)

There are some situations where ARIMA models cannot be used. In the order to achieve stationarity, the data needs to be comparing periods when the time series data are non-stationary. In this instance, the dependent variable is subjected to the ARIMA model. The ARIMA model is indicated ARIMA (p, d, q) and described as [17].

$\dot{Z}_t = \psi + \alpha_1 \dot{Z_{t-1}} + \alpha_2 \dot{Z_{t-2}} + \cdots + \alpha_p \dot{Z_{t-p}} + \xi_t - \Omega_1 \xi_{t-1} + \Omega_2 \xi_{t-2} - \cdots - \Omega_q \xi_{t-q}$ (6)

By using the Lag operator: $\alpha(\pi)(1 - \pi)^d Z_t = \psi + \Omega(\pi) Z_t$

The $\dot{Z}_t$ represent the difference d of the dependent variable Z_t and n is the observations number, the equation can be denoted by: $\Delta^d Z_t = \dot{Z}_t (1 - \pi)^d Z_t = \sum_{n=0}^{d} \binom{d}{n} (-1)^n Z_{t-n}$ (7)

The Box-Jenkins methodology

The four Box-Jenkins steps are: model identification, model estimation, model diagnostic checking, and forecasts (Young, 1977) we can summarize the four steps as follows:

1. *Model identification*: verify that the variables are stable, identify seasonality in the series, or non-seasonality, and use the series' autocorrelation functions (ACFs) and partial autocorrelation functions (PCFs) graphs to determine which autoregressive or moving-average components could be used in the model.
   The autocorrelation function of the sample (ACFs) is known as:

$$R_k = \frac{\sum_{t=1}^{T-k} (Z_t - \bar{Z}_t)(Z_{t+k} - \bar{Z}_t)}{\sum_{t=1}^{T} (Z_t - \bar{Z}_t)^2}$$

   The partial autocorrelation function measures the correlation degree of two variables (PCFs) and is denoted by the formula:

$$P_{k,k} = \begin{cases} R_1 & \text{If } k = 1 \\ \frac{R_k - \sum_{j=1}^{k-1} (P_{k-1,j} * R_{k-j})}{1 - \sum_{t=1}^{k-1} (P_{k-1,j} * R_j)} & \text{If } k > 1 \end{cases}$$

   The $P_{k,j} = P_{k-1,j} - P_{k,k} P_{k-1,j-1}$ for $j = 1,2, \dots, k-1$
   The order of moving- average process will be determined by ACF, whereas the autoregressive process order will be defined by PACF.
2. *Model estimation*: depends on calculation algorithms to obtain the coefficients best suited to the chosen ARIMA model. The most important approaches use the maximum likelihood estimation or nonlinear least square estimation.

3. *Model diagnostic checking*: To verify whether the assessed model meets the specifications of the stationary univariate method. Principally, the residuals must be independent of each other and be constant in average and variance over time; graphing residual ACF and PACF helps to determine inappropriate specifications. If the estimates are inadequate, it is necessary to return to the first stage and attempt to construct a better model. Moreover, it is important to compare the estimated model with other ARIMA models to select the best model for the data. The residual formula of ARIMA models is represented by: $\hat{\xi}_t = \hat{Z}_t - (\hat{\sigma} + \sum_{i=1}^{p} \hat{\alpha}_i Z_{t-1} - \sum_{i=1}^{q} \hat{\delta}_i \widehat{\xi_{t-i}})$

In addition, the two common criteria for selecting a model are Akaik's Information Criteria (AIC), Bayesian Information Criteria (BIC), and the Box-Ljung Criteria defined as the following criteria. $AIC = 2m - 2\ln(\hat{L})$

$$BIC = \ln(n)\, m - 2\ln(\hat{L})$$

$$Q = n(n+2)\sum_{k=1}^{t} \frac{R_k^2}{n-k}$$

The $\hat{L}$ is the model value of the maximum likelihood function, m represents the parameters number of the estimated model, n is the observation number of samples, also the R_k^2 is the sample autocorrelation. AIC, BIC, and Q are utilized in the standard criterion; mean square error.

4. *Forecasting model*: in case of, the nominated ARIMA model matches the specification of the stationary univariate procedure; the model can be used for forecasting [16][20] . To find accurate expectations, the model must have the least value of the mean square error, which is considered through the following procedure: $MSE = \frac{1}{n}\sum_{t=1}^{n} {\xi_t}^2$ .When the value of MSE approaches to zero, this indicates that the estimated values series are close to the original values [20] Similarly many other different statistical methods to measure the accuracies of forecast which are: Mean absolute error (MAE), root mean square error (RMSE), mean square error (MSE) and mean absolute percentage error (MAPE) to fit model:

$$MAE = \frac{1}{n}\sum_{t=1}^{t=n}\left|Z_t - \hat{Z}_t\right|, \quad RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{t=n}(Z_t - \hat{Z}_t)^2}$$

$$Where \;\; q_t = \frac{|Z_t - \hat{Z}_t|}{\frac{1}{n-1}\sum_{t=2}^{t=n}|Z_t - Z_{t-1}|}, \quad MASE = \frac{1}{n}\sum_{t=1}^{t=n}|q_t|$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{t=n} 100 * \frac{\left|Z_t - \hat{Z}_t\right|}{|Z_t|}$$

Where Z_t is the dependent variable at t of time, Z_{t-1} is the dependent variable at $t-1$ of time, $\hat{Z}_t$ is the predicted value, the q_t is the measured error, and n is the sample size.

RESULTS

The first step in data analysis was to represent the data series graphically, to study the data behavior, as shown in Figure 1, which indicates that the series is non-stationary. Furthermore, Figure 2 clarifies the pattern data through the correlograms of the partial autocorrelation and autocorrelation functions, with confidence intervals. Besides that, the graphical examination of Figure 1 explains that the data series includes the oscillation behavior. As an initial procedure, change the series into a linear trend and stationary state by applying the difference. The value of the test of Dickey-Fuller was -3.0762, with lag order 5 and p-value = 0.1283 more than 0.05 to confirm the non-stationary series.

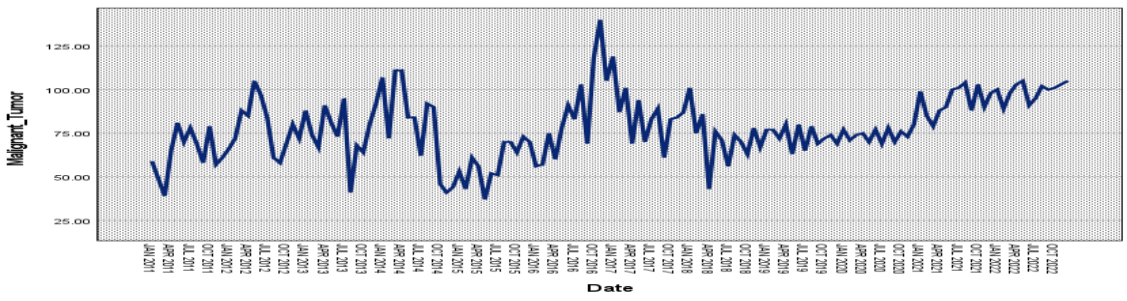

Figure 1. The plot represents the time series of cancer in the period(2011-2022)

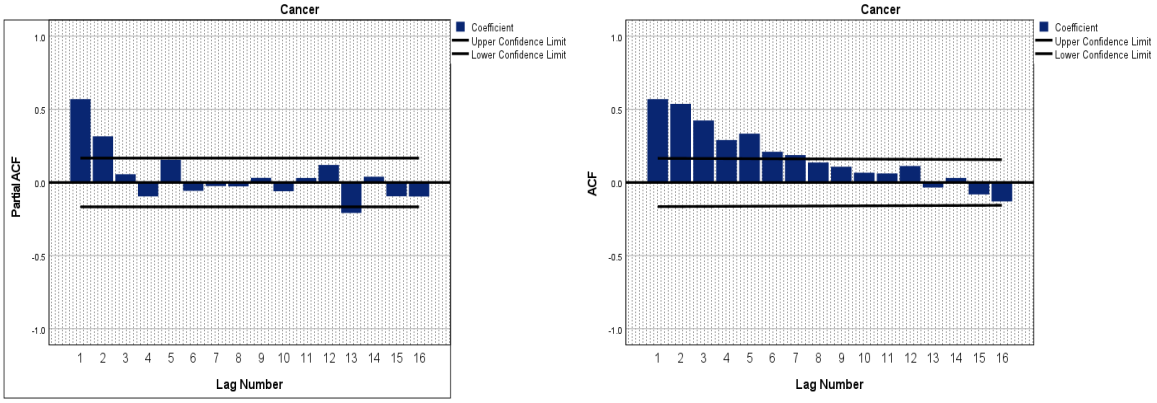

Figure 2. Graphs show the CAF and PACF for cancer series

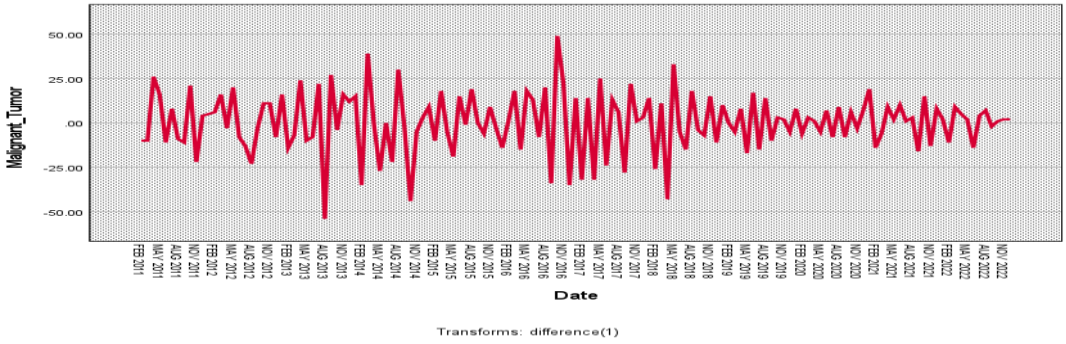

Figure 3. Curve of time series stationary in the first difference around the mean

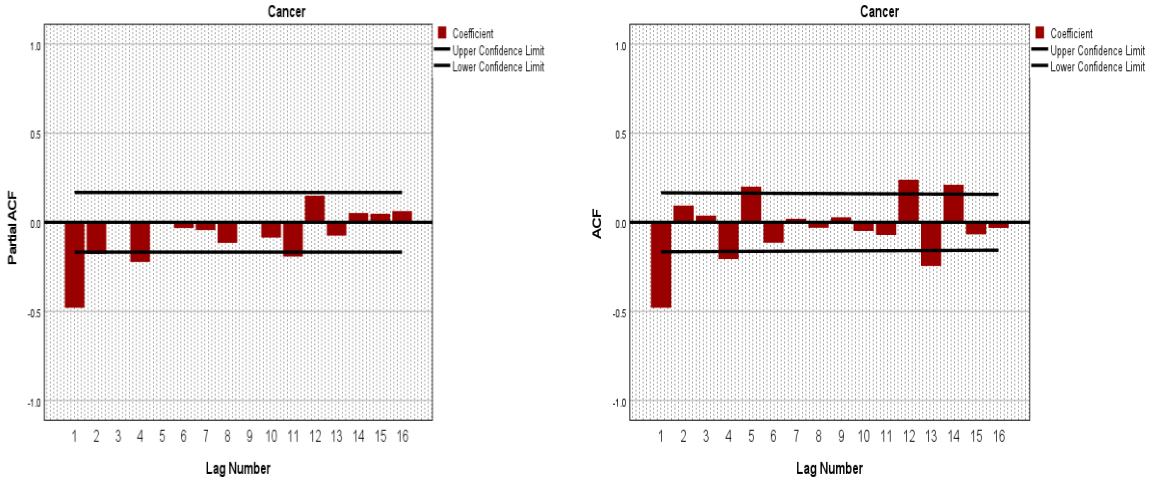

Figure 4. Different parameters (p, q) plot of CAF and PACF in first difference

Figure 3 explains when data series converted to the stationary state at d=1 around mean and variance by the equation (7) $(\sum_{n=0}^{d}\binom{d}{n}(-1)^{n}Z_{t-n})$ and Dickey-Fuller test = -6.0022, Lag order = 5, p-value = 0.01 less than 0.05, which means t hat the data series is stationary. After performing the difference in the series, the next step is building a time series m odel. The building stage model relies on identifying the parameters (p and q) in the ARIMA model with autocorrelatio n and partial autocorrelation functions of the sample as revealed in Figure 4. Furthermore, shows that the cancer series partial autocorrelation function is not significantly different from zero when the order of spikes is greater than 1, 4, and 11, hence the proposed values are AR (0,1, 2, 3). Similarly, the autocorrelation function is not significantly different f rom zero when the order of lag is greater than 1,4,5,12,13, and 14, therefore the suggested values of CFA are MA (0,1, 2, 3).

Table 1. Comparison of different ARIMA Models of (p, q)

ARIMA Models	Stationary R-squared	R-squared	RMSE	MAPE	MaxAPE	MAE	MaxAE	BIC
ARIMA (1,1,1)	0.269	0.38	14.46	15.21	107.66	10.71	44.14	5.45
ARIMA (0,0 ,1)	0.202	0.202	16.29	18.30	108.07	12.89	46.95	5.65
ARIMA (0,1,0)	2.22E-16	0.16	16.69	17.91	132.49	12.86	54.32	5.66
ARIMA (0,1,1)	0.257	0.37	14.43	15.17	109.22	10.69	44.78	5.41
ARIMA (2,1,1)	0.265	0.38	14.46	15.22	103.29	10.76	42.35	5.48
ARIMA (1,1,2)	0.260	0.38	14.51	15.19	107.42	10.70	44.04	5.49
ARIMA (2,1,2)*	0.267	0.38	14.49	15.16	105.75	10.69	43.36	5.52
ARIMA (2,1,3)*	0.316	0.423	14.045	14.85	88.016	10.53	40.10	5.49
ARIMA (3,1,2)*	0.313	0.42	14.08	14.718	98.918	10.41	47.29	5.50
ARIMA (3,1,3)*	0.319	0.426	14.07	14.648	96.944	10.37	45.92	5.53
ARIMA (0,1,3)*	0.275	0.388	14.362	14.948	100.877	10.52	49.80	5.47
ARIMA (0,1,2)	0.259	0.375	14.469	15.18	108.418	10.69	44.45	5.45
ARIMA (2,1,0)	0.255	0.371	14.506	15.358	106.411	10.83	43.63	5.45

The result in Table 1, displays the different parameters of ARIMA models with different statistics, R-squared value, RMSE value, MAPE value, and BIC value. All estimated statistics are fundamental measures for choosing the optimal ARIMA models. The ARIMA (2,1,2)*, ARIMA (2, 1, 3)*, ARIMA (3,1,2)*, ARIMA(3,1,3)* and ARIMA(0,1,3)* are a good model to fit the data series because contains the approximately minimum values of RMSE, BIC, and also a high value of determination coefficient R-squared which indicates the good fit model.

Table 2. Preference Criteria to Select Order ARIMA Models

ARIMA Models	Number of Estimated Parameters	Significance Test of estimated parameters
ARIMA (2,1,2)	4	2
ARIMA (2,1,3)	5	4
ARIMA (3,1,2)	5	3
ARIMA (3,1,3)	6	3
ARIMA (0,1,3)	3	2

From table 2, it can be concluded that the best model to describe the time series data is the ARIMA (2, 1, 3) *, since, most of its parameters approve the statistical significance of t-test coefficients, also accepted among all other ARIMA models in all values of preference criteria and used tests.

Table 3. Estimated statistics of Model fit

Model Fit										
Fit Statistic	**Mean**	**Minimum**	**Maximum**	**Percentile**						
				5	10	25	50	75	90	95
Stationary R-squared	0.316	0.316	0.316	0.316	0.316	0.316	0.316	0.316	0.316	0.316
R-squared	0.423	0.423	0.423	0.423	0.423	0.423	0.423	0.423	0.423	0.423
RMSE	14.05	14.05	14.05	14.05	14.05	14.05	14.05	14.05	14.05	14.05
MAPE	14.85	14.85	14.85	14.85	14.85	14.85	14.85	14.85	14.85	14.85
MaxAPE	88.02	88.02	88.02	88.02	88.02	88.02	88.02	88.02	88.02	88.02
MAE	10.53	10.53	10.53	10.53	10.53	10.53	10.53	10.53	10.53	10.53
MaxAE	40.10	40.10	40.10	40.10	40.10	40.10	40.10	40.10	40.10	40.10
Normalized BIC	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49

Table 3 defines the result of model fit data statistics, which reflect the expected accuracy of the model; the result can be summarized as the R-squared value around 42.3%, which indicates that the model and the value of RMSE, which explains 14.045 of the error average between observed and predicted values.

Table 4. The Ljung Box test for the ARIMA Model

Model Statistics							
Malignant Tumor monthly - Model_1	Model Fit Statistics			Ljung-Box Q (18)			Number of Outliers
	Number of Predictors	Stationary R-squared	Normalized BIC	Statistics	DF	Sig.	
	0	0.316	5.493	21.593	13	0.061	0

Table 4 reports the statistic of the Ljung-Box (Q-test) is 21.593 with p-value=0.061which proves the residuals are insignificantly different from white noise and the model is acceptable to forecast.

Table 5. Estimation of ARIMA Model Parameter

ARIMA Model Parameters					
		Estimate	**SE**	**T**	**Sig**
AR	Constant	0.325	0.499	0.652	0.516
	Lag 1	-0.363	0.076	-4.772	0.000
	Lag 2	-0.783	0.078	-10.079	0.000
	Difference - 1				
MA	Lag 1	0.190	0.108	1.751	0.082
	Lag 2	-0.726	0.255	-2.844	0.005
	Lag 3	0.619	0.183	3.387	0.001

Table 5 shows the constant and t-test values for the estimated parameters of the selected model. The t-test values have the most statistically significant coefficients at 0.05, whereas the MA coefficients of lags 5 are not statistically significant at 0.05. We could express of estimated model ARIMA (2,1,3) for time series data of patients with cancer at BMC in the following equation:

$$Z_t = 0.325 - 0.363\, Z_{t-1} - 0.783\, Z_{2-1} + 0.190\xi_{t-1} - 0.726\, \xi_{t-2} + 0.619\xi_{t-3} + \xi_t (8)$$

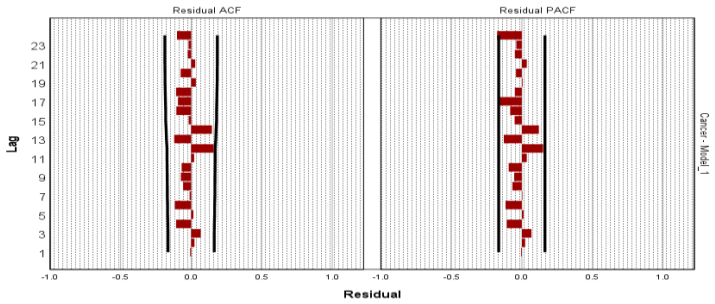

Figure 5. Plots of residual series of autocorrelation and partial autocorrelation functions

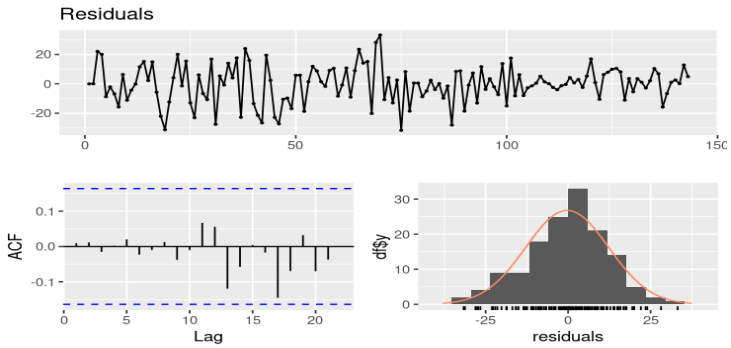

Figure 6. Represent residuals of the ARIMA (2,1,3)

The diagnostic step requires the adequate of the estimated model to fit the data series, the diagnoses involved on a test of randomized residuals, independence residuals, normality of residuals, and general adequate of the estimated model. The random test of residuals of ARIMA (2,1,3) applied on confidence intervals for generated residuals from series data that expressions from Figure 5, all the points of autocorrelations and partial autocorrelations functions between confidence boundaries at (95%) that denoting the residuals series of candidate model represents random variables and the selected model is valid.

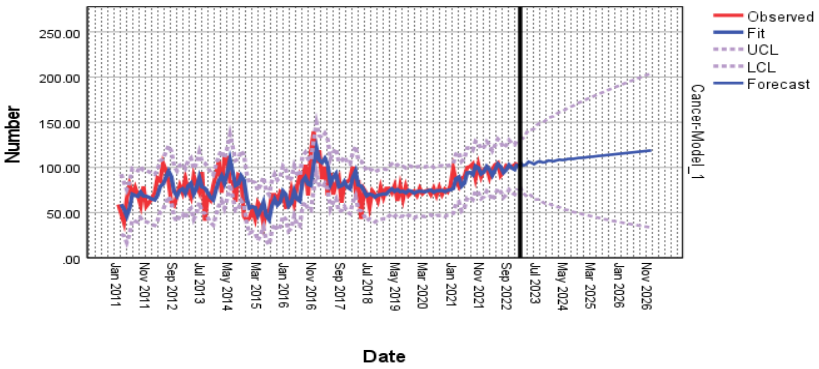

Figure 7. Forecast cancer cases from 2023 to 2026 per month

Finally, in the forecasting step, the time series of cancer cases of observed values and predicted values were obtained by the use of ARIMA (2,1,3) in Figure 5 by using the equation (8). The figure illustrates that the effect of fitting is an approximate pattern to the original data series. The result of predication values with upper and lower limits from 2023 to 2026 monthly is revealed in subsequent Table 6. Table 6 shows a gradual rise in the cancer cases at Benghazi Medical Center for the next 4 years, it can be seen the total of months for each year 2023,2024,2025, and 2026 are 1266,1314,1358 and 1406 respectively beside that, the total for the next years will reach to 5344 patients.

Table 6. Number of cancer cases Forecasted in the period (January 2023- December 2026)

Date	Forecast	Upper limit	Lower limit	Date	Forecast	Upper limit	Lower limit
JAN 2023	106	133	79	JAN 2025	112	176	47
FEB 2023	102	132	72	FEB 2025	112	177	46
MAR 2023	103	137	69	MAR 2025	112	178	46
APR 2023	106	142	71	APR 2025	112	180	45
MAY 2023	105	142	69	MAY 2025	113	181	44
JUN 2023	104	143	65	JUN 2025	113	182	44
JUL 2023	106	147	65	JUL 2025	113	184	43
AUG 2023	107	149	65	AUG 2025	114	185	42
SEP 2023	106	149	62	SEP 2025	114	186	42
OCT 2023	106	152	60	OCT 2025	114	188	41
NOV 2023	108	155	60	NOV 2025	114	189	40
DEC 2023	107	156	59	DEC 2025	115	190	40
JAN 2024	107	157	57	JAN 2026	115	191	39
FEB 2024	108	159	56	FEB 2026	116	193	39
MAR 2024	109	161	56	MAR 2026	116	194	38
APR 2024	109	162	54	APR 2026	116	195	38
MAY 2024	109	164	53	MAY 2026	116	196	37
JUN 2024	109	166	53	JUN 2026	117	197	37
JUL 2024	110	167	52	JUL 2026	117	199	36
AUG 2024	110	168	51	AUG 2026	118	200	36
SEP 2024	110	170	50	SEP 2026	119	201	35
OCT 2024	111	172	50	OCT 2026	118	202	35
NOV 2024	111	173	49	NOV 2026	119	203	34
DEC 2024	111	174	48	DEC 2026	119	204	34

DISCUSSION

Time series analysis approved the effectiveness of handling with historical data to predict disease events, through using the statistical technique the ARIMA models depend on several stages (identification, estimation, diagnostic checking, and forecasts), which classified them as powerful tool methods. Our study amid to study the pattern and determine an appropriate model for forecasting the number of patients with cancer at Benghazi Medical Center by applying an autoregressive integrated moving-average (ARIMA) model.

Based on our knowledge, this the second study to apply and classify behavior of ARIMA model cancer of 11266 divided on (144 months) time series of individuals for 12 years in Libya. The findings of our study approximately adapt with some of the previous studies that have employed ARIMA models for forecasting patient numbers and incidence of disease effectively applied an ARIMA model to forecast new cases of the COVID-19 pandemic per day for the entire world [8].

Likewise, researchers successfully implemented an ARIMA model to predict the yearly incidence and mortality rate for prostate cancer in Australia [9,10]. Also, they utilized a successful ARIMA model to predict the epidemic situation of brucellosis incidence per month in the City of Jinzhou, China which developed an ARIMA-based model for forecasting the patient number of epidemic diseases, the importance of the effectiveness of this approach [12]. Moreover, previous study used time series analysis, including ARIMA modeling, to analyze onchocerciasis data from Mexico, representing a trend of the disease [14]. In the current study, we found that the data series of patients with cancer exhibited an oscillation pattern i.e., non-stationary series as exhibited in prior study [9]. In addition, the condition of stationary satisfied of data series by a performance of the first-order difference which is different from [8].

Our findings have shown that the ARIMA (2,1,3) model was the most appropriate model to fit the data series and most of the estimated parameters model was statically significant of t-statistic [8,9]. The residual test confirmed that all the points of (AC) and (PAC) were between confidence intervals [12]. Finally, the selected model of the predictive values showed approximately the actual values of the data series to predict the number of patients with cancer at Benghazi Medical Center from January 2023 to December 2026.

CONCLUSION

The results of this study suggest that the ARIMA (2,1,3) model is a good tool for predicting the number of patients with cancer in Benghazi Medical Center. The results also suggest that, the model revealed acceptable accuracy and reliability in forecasting numbers of patients through statistical tests (significance of estimated parameters and residuals of autocorrelation function). The results indicated that a gradually growth in the total number of people with cancer from January 2023 to December 2026, reaching up to 5344 patients.

Conflict of Interest

There are no financial, personal, or professional conflicts of interest to declare.

REFERENCES

1. Zarmouh A, Almalti A, Alzedam A, Hamad M, Elmughrabi H, Alnajjar L, et al. Cancer incidence in the middle region of Libya: Data from the cancer epidemiology study in Misurata. Cancer Rep. 2022;5(1):1–7.
2. Attia A, Siala I, Azribi F. General Oncology Care in Libya. Cancer Arab World. 2022;133–48.
3. El Mistiri M, Salati M, Marcheselli L, Attia A, Habil S, Alhomri F, et al. Cancer incidence, mortality, and survival in Eastern Libya: Updated report from the Benghazi Cancer Registry. Ann Epidemiol. 2015;25(8):564–8.
4. Erashdi M, Al-Ani A, Mansour A, Al-Hussaini M. Libyan cancer patients at King Hussein Cancer Center for more than a decade, the current situation, and a future vision. Front Oncol. 2023;12(January):1–12.
5. Murat M, Malinowska I, Gos M, Krzyszczak J. Forecasting daily meteorological time series using ARIMA and regression models. Int Agrophysics. 2018;32(2):253–64.
6. Villani M, Earnest A, Nanayakkara N, Smith K, Courten B De. Time series modelling to forecast prehospital EMS demand for diabetic emergencies. 2017;1–9.
7. Dritsakis N, Klazoglou P. Forecasting Unemployment Rates in USA Using Box-Jenkins Methodology. Int J Econ Financ Issues [Internet]. 2018;8(1):9–20.
8. Duong NQ, Thao LP, Thi D, Nhu Q, Binh LT, Thi C, et al. Predicting the Pandemic COVID-19 Using ARIMA Model. 2020;36(4):46–57.
9. Earnest A, Evans SM, Sampurno F, Millar J. Forecasting annual incidence and mortality rate for prostate cancer in Australia until 2022 using autoregressive integrated moving average (ARIMA) models. BMJ Open. 2019;9(8):1–7.
10. Wang L, Liang C, Wu W, Wu S, Yang J, Lu X, et al. Epidemic situation of brucellosis in jinzhou city of china and prediction using the ARIMA Model. Can J Infect Dis Med Microbiol. 2019;2019.
11. Wei W, Jiang J, Liang H, Gao L, Liang B, Huang J, et al. Application of a combined model with autoregressive integrated moving average (arima) and generalized regression neural network (grnn) in forecasting hepatitis incidence in heng county,

China. PLoS One. 2016;11(6):1–13.
12. Pan Y, Zhang M, Chen Z, Zhou M, Zhang Z. An ARIMA Based Model for Forecasting the Patient Number of Epidemic Disease. 2015;31–4.
13. Kumar N, Kumari P, Ranjan P, Vaish A. ARIMA model based breast cancer detection and classification through image processing. SCES 2014 Inspiring Eng Syst Glob Sustain. 2014;
14. Lara-Ramírez EE, Rodríguez-Pérez MA, Pérez-Rodríguez MA, Adeleke MA, Orozco-Algarra ME, Arrendondo-Jiménez JI, et al. Time Series Analysis of Onchocerciasis Data from Mexico: A Trend towards Elimination. PLoS Negl Trop Dis. 2013;7(2):1–8.
15. Box GEP, Tiao GC. Intervention analysis with applications to economic and environmental problems. J Am Stat Assoc. 1975;70(349):70–9.
16. Abonazel MR, Abd-Elftah AI. Forecasting Egyptian GDP using ARIMA models. Reports Econ Financ. 2019;5(1):35–47.
17. Ugoh CI, Echebiri UV, Temisan GO, Iwuchukwu JK, Guobadia , Emwinloghosa Kenneth. On Forecasting Nigeria's GDP: A Comparative Performance of Regression with ARIMA Errors and ARIMA Method. Int J Math Stat Stud. 2022;10(4):48–64.
18. Mohamed J. Time Series Modeling and Forecasting of Somaliland Consumer Price Index: A Comparison of ARIMA and Regression with ARIMA Errors. Am J Theor Appl Stat. 2020;9(4):143.
19. W. L. Young, The Box-Jenkins Approach to Time Series Analysis and Forecasting: Principles and Applications, RAIRO-Operations Research- Recherche Opérationnelle, 11 (1977), 129-143. https://doi.org/10.1051/ro/1977110201291.
20. Fakhreddine. Munich Personal RePEc Archive GDP Forecast of the Biggest GCC Economies Using ARIMA. 2021;(108912).

تطبيق تحليل السلاسل الزمنية (بوكس-جينكينز) للتنبؤ بعدد مرضى السرطان في مركز بنغازي الطبي

كاملة العبد عثمان[1]*، نعيمة نصر عبد العاطي[2]، سارة عطية الوراد[1]

[1]قسم الإحصاء، كلية الآداب والعلوم المرج، جامعة بنغازي، بنغازي، ليبيا.
[2]قسم الإحصاء، كلية العلوم، جامعة بنغازي، بنغازي، ليبيا

المستخلص

الخلفية والاهداف. للسرطان تأثير سلبي على صحة الإنسان في العالم، حيث يؤدي إلى المرض والوفاة. صممت هذه الدراسة لتطبيق تحليل السلاسل الزمنية باستخدام منهجية بوكس جنكينز لوصف السلوك وإيجاد نموذج مناسب للتنبؤ بالسلاسل الزمنية لمرضى السرطان في مدينة بنغازي. **طُرق الدراسة**. تم جمع البيانات من السجلات الطبية لقسم الأورام بمركز بنغازي الطبي في الفترة من يناير 2011 إلى ديسمبر 2022، لـ 11266 مريضاً موزعين على 144 سلسلة زمنية من المشاهدات. **النتائج**. وجدت الدراسة أن السلسلة الزمنية للبيانات كانت غير ثابتة ولها نمط تذبذب حول الوسط الذي تحول إلى سلسلة ثابتة عن طريق إجراء الفرق الأول. لقد أنشأنا العديد من نماذج ARIMA وقمنا بمقارنتها بناءً على معايير مختلفة، بناء علي قيم: R-squared=0.423 ، وجذر متوسط مربع الخطأ 14.045= (RAMSE) ومعيار المعلومات البايزية 5.447 = (BIC), النموذج المناسب الذي تم اختياره لتمثيل سلسلة البيانات هو ARIMA(2,1,3) والذي يستخدم أيضًا للتنبؤ بالحالات الجديدة في السنوات الأربع القادمة. ونتيجة لذلك، أثبت النموذج المقدر وجود تشابه بين القيم المتوقعة والقيم الحقيقية للسلسلة الزمنية. بالإضافة إلى ذلك، أشارت النتائج إلى ارتفاع تدريجي في إجمالي عدد المصابين بالسرطان من يناير 2023 إلى ديسمبر 2026، ليصل إلى 5344 مريضا. **الخاتمة**. يعد نموذج ARIMA(2,1,3)، أداة جيدة للتنبؤ بعدد مرضى السرطان في مركز بنغازي الطبي. وأخيرا تأتي أهمية هذه الدراسة بالاعتماد على النتائج في رفع مستوى الوعي والمعرفة بمخاطر مرض السرطان في المجتمع الليبي.

الكلمات الدالة. تحليل السلاسل الزمنية، منهجية بوكس جنكينز، نماذج أريما، التنبؤ، السرطان.