Clinical Trial/Experimental Study

Medicine®

OPEN

HCV extinction analysis in district Gujrat, Pakistan by using SARIMA and linear regression models

Muhammad Rashid, MPhil, Hammad Ismail, PhD*

Abstract

Background: To investigate the track of Gujrat, a District of Pakistan is very essential, either it follow-up World Health Organization (WHO) Hepatitis C Virus (HCV) elimination plan or not. This study aimed to find out HCV extinction analysis by time series forecast from District Gujrat, Pakistan.

Methods: From January 1, 2016 to December 31, 2020 total n-5,111 numbers of HCV real-time polymerase chain reaction (RT-PCR) tests were performed in Gujrat. For extinction analysis we used 2 different models, the first model was seasonal auto-regressive integrated moving average (SARIMA) and the second linear regression (LR) model. First, we fitted both models then these fitted and valid models were used to predict future HCV percentage in District Gujrat.

Results: In District Gujrat, the men HCV infected ratio is high with a higher viral load as compared with women, from year 2016 to 2020 male to female ratio was (53.75:53.19), (45.67:43.84), (39.67:39.36), (41.94:35.88), (37.70:31.38) respectively. HCV percentage is decreasing from 2016 to 2020 with an average of 4.98%. Our both fitted models SARIMAX (0,1,1)(0,1,1,6) at 95% confidence intervals and LR model $Y = -0.379X + 53.378$ at 99% confidence intervals (P-value = .00) revealed that in June 2029 and in August 2027 respectively HCV percentage will be 0 from district Gujrat, Pakistan.

Conclusions: This study concluded that both SARIMA and LR models showed an effective modeling process for forecasting yearly HCV incidence. District Gujrat, Punjab, Pakistan is on track to achieve the WHO HCV elimination plan, before 2030 HCV will be extinct from this region.

Abbreviations: ADF = augmented Dickey-Fuller, CI = confidence intervals, HCV = Hepatitis C virus, LR = linear regression, MAPE = mean absolute percentage error, SARIMA = seasonal auto-regressive integrated moving average, SPSS = Statistical Package for Social Sciences, WHO = World Health Organization.

Keywords: district Gujrat Pakistan, hepatitis C, linear regression, modeling, seasonal auto-regressive integrated moving average

1. Introduction

Hepatitis C virus (HCV) is a blood-borne pathogen, it affects hepatic cells and causes liver cirrhosis or hepatocellular carcinoma after untreated chronic hepatitis. HCV viral structural proteins are (C, E1, E2, and P7) whereas nonstructural core proteins are (NS2, NS3, NS4A, NS4B, NS5A, and NS5B).[1] Almost 200 million people are infected (~5% of the population) with HCV worldwide and about 3.5×10^5 deaths occur globally due to liver cancer caused by HCV.[1] In Pakistan 1 out of 20 people is infected with HCV, about 10 million people are infected with HCV by unscreened blood donation, contaminated injections practices in health care systems and utilization of unsterilized equipment that make Pakistan the second-highest HCV infected country after Egypt.[2] From 2016 to 2030 due to the increasing population of Pakistan up to 250 million, HCV prevalence will also increase from 3.9% to 5.1%. On account of this HCV chronic infection will also increase from 7.5 to 12.6 million, HCV incident infections will increase from 7.0×10^5 to 1.1 million annually and 1.4 (1.0–2.0) million HCV-related deaths will occur.[3] To save these people life's from HCV-related deaths until there is no successful vaccination for HCV due to its huge genotyping variation. HCV has 7 genotypes based on its sequence and phylogenetic analysis, these 7 genotypes have 67 well-recognized subtypes and 20 provisional subtypes.[4]

To combat this HCV that has vast genetic variability, World Health Organization (WHO) set a global health sector strategy to eliminate HCV infection by 2030. In this strategy, a major focus is on reducing hepatitis incidence (90%) and hepatitis mortality (65%) by 2030.[5] Unfortunately, to meet the WHO target, 76% of high-income countries are not on track and 60% are off-track by at least the next 20 years. There are just 11 countries that are on the track of the WHO plan (Australia, Canada, France, Germany, Iceland, Italy, Japan, Spain, Sweden, Switzerland, and United Kingdom).[6] To meet WHO targets Government of

Editor: Abrar Hussain Khan.

The authors have no conflicts of interest to disclose.

Supplemental Digital Content is available for this article.

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Department of Biochemistry and Biotechnology, University of Gujrat-Hafiz Hayat Campus, Gujrat, Punjab, Pakistan.

* Correspondence: Hammad Ismail, University of Gujrat - Hafiz Hayat Campus, University of Gujrat, Gujrat 50700, Punjab, Pakistan (e-mail: hammad.ismail@uog.edu.pk).

How to cite this article: Rashid M, Ismail H. HCV extinction analysis in district Gujrat, Pakistan by using SARIMA and linear regression models. Medicine 2021;100:49(e28193).

Received: 3 March 2021 / Received in final form: 17 November 2021 / Accepted: 19 November 2021

http://dx.doi.org/10.1097/MD.0000000000028193

Pakistan established a National Hepatitis Strategic Framework (2017–2021) in October 2017 with the participation of Federal and Provincial partners together with private sectors and NGOs that plans HCV testing and access to direct-acting antivirals treatment at a low cost in Pakistan.[7] To control hepatitis at the provincial level "The Punjab Hepatitis Act 2018" has been launched.[8]

In Pakistan at the Provincial and District level allots of efforts are going to be performed but unfortunately until there is no recent comprehensive study and extinction analysis of HCV has been done at any level in Pakistan. Based on emigration and overseas Pakistani out of 154 Districts of Pakistan, District Gujrat is on 11th number therefore the major aim of this study was to find out either District Gujrat is on track with WHO plan to eliminate HCV by 2030 or not.

2. Materials and methods

Gujrat is a District of Punjab that covers 1232 (sq. mi area), it consists of 3 tehsils (Gujrat, Kharian, Sarai-Alamgir) and has a 2.7×10^6 population. This research work was conducted from January 1, 2016 to December 31, 2020 with the collaboration of the Department of Biochemistry and Biotechnology, University of Gujrat and Dr Mujahid Lab between concerned persons. Blood samples n-5,111 (sample size margin error 2%, confidence level [CI] 99%)[9] of positive anti-HCV screened patients were collected in Dr Mujahid Lab (DML) Gujrat and from its all-collection points (Hospital Labs) that located within district, to perform HCV RT-PCR. Research work was performed in the Biochemistry Laboratory of the University of Gujrat and Dr Mujahid Lab, Gujrat. This research work is done with the consent of patients and study design approved by the Institutional Review Board (IRB) letter reference no. UOG/ORIC/2020/292 dated November 24, 2020 of the University of Gujrat (UOG), Gujrat.

2.1. Samples collection and storage

Blood sample (4 cm^3) from patients was collected in a Serum Gel Vacuum tube after performed the venipuncture and then incubated at 37 °C for 20 to 30 minutes. After the incubation period sample was centrifuged at 4000 rpm for 2 minutes at room temperature then the upper layer of serum was separated in a sterilized (labeled) Eppendorf tube and stored at –20 °C. Blood samples from all other collection points were delivered to DML in iceboxes, after received these samples were stored at –20 °C till further HCV RNA extraction and amplification.

2.2. RNA extraction and amplification

By using a commercially available kit (FavorPrep Viral Nucleic Acid Extraction Kit by Taiwan) RNA was extracted from collected serum as per manufacturer guidelines. This extracted RNA was stored at –20 °C till further amplification. The amplification of HCV extracted RNA was done by RT-PCR (Cepheid Smart cycler II by America). The master mixture was prepared from a commercially available kit (HCV Real-TM Quant by Italy) by adding 300 μL of RT-PCR-mix-1, 200 μL of RT-PCR-mix-2, 20 μL of Hot Start Taq Polymerase, and 10 μL of M-MLV revertase into dithiothreitol tube according to manufacturer protocol. For amplification, 12.5 μL prepared master mixture and 12.5 μL extracted RNA were transferred into PCR reaction tube then mixed it at vortex mixer for 10 seconds then centrifuged this mixture at 8000 rpm for 40 to 60 seconds. Amplification was done under fluorescence detections (FAM & Cy3) by using kit protocol. The results were interpreted by using FAM (Green) to detect internal control and Cy3 (Yellow) to detect HCV amplified product. Viral load was calculated by using the following formula in which HCV and Internal Control (IC) cDNA copies were attained by a standard curve that was run according to manufacturer kit protocol.

$$\frac{\text{HCV cDNA copies per reaction}}{\text{Internal Control cDNA copies per reaction} \times \text{Coefficient A} \times \text{Coefficient B}} = \text{Viral load IU/mL}$$

Coefficient $A = (100/150\,\mu\text{L}$ [serum volume])
Coefficient $B = 99{,}200$

$$\text{Viral load}\frac{\text{IU}}{\text{mL}} \times 4 = \text{Viral load copies/mL}$$

2.3. Extinction analysis

For extinction analysis, 2 approaches were applied in the first approach we used the seasonal auto-regressive integrated moving average (SARIMA) model with the help of integrated development environment Jupyter Notebook (anaconda3) a coding tool of Anaconda Navigator 3 version 2020.02 that was installed in Microsoft Window 10. The second statistical analysis approach of the linear regression model was used by using Statistical Package for Social Sciences (SPSS) for Windows Version 21.0 in Microsoft Window 10.

2.3.1. SARIMA model. The SARIMA model is the most popular and widely used time series forecasting method that was proposed by Box and Jenkins in 1970.[10] The general form of SARIMA model is (p,d,q)(P,D,Q)S. Here p, d, and q are the orders of autoregressive (AR), differences (I), and moving average (MA) whereas P, D, and Q are the seasonal orders of the autoregressive, difference and moving average. S represents the length of the seasonal periods.

We adopted the following steps to process the SARIMA model.[11,12]

2.3.1.1. Import library. To start our analysis we created the following library dataset coding in Jupyter Notebook.

```
import warnings
import pmdarima as pm
import itertools
import numpy as np
import matplotlib
import matplotlib.pyplot as plt
import pandas as pd
from statsmodels.tsa.arima.model import ARIMA
import statsmodels.api as sm
from statsmodels.tsa.stattools import adfuller
from statsmodels.tools.eval_measures import rmse, aic
```

2.3.1.2. Data pre-analysis. To build and apply SARIMA model first preconditions like linearity, seasonality, and stationarity of the dataset were checked linearity patterns and seasonality of the dataset were checked from the Additive Decomposition model by using the following coding. To check out the stationarity of data

we used 2 different approaches, first from graphically visualized by "rolling statistics" and second from the Augmented Dickey-Fuller (ADF) test. In ADF our null hypothesis (H_o) was data are not stationary and the alternative hypothesis (H_a) was data are stationary. If $P < .05$ we reject H_o and accept H_a and if $P > .05$ then we accept H_o and reject H_a.

```
from pylab import rcParams
from statsmodels.tsa.seasonal import seasonal_decompose
```

2.3.1.3. Model Identification and estimation. To find the best parameters for our model we used the Pyramid auto-ARIMA model also called "auto-arima" or "pmdarima" which automatically generated the best and suitable model for forecasting based on grid search. To functionalize and install it we generated the "pmdarima" library by using the following command. From Pyramid auto-ARIMA best model was selected based on minimum Akaike Information Criterion value this selected model was then fitted for prediction analysis.

```
"pip install pmdarima"
```

2.3.1.4. Model acceptability and diagnosis. Model diagnostic test was performed to find out the validity, acceptability, and white noise of our model. It was graphically represented by Standardized residues, Histogram, Quantile-Quantile (Q-Q) plot, and correlogram.

2.3.1.5. Accuracy test of selected model. To check the accuracy of our selected model we used one step ahead forecast at 60% of testing data from the tail. Moreover, for this, we also run and checked mean absolute percentage error (MAPE) to give validation of our fitted model.

2.3.1.6. Prediction analysis. The best and fitted obtained model was used to predict the HCV percentage of the upcoming 10 years in District Gujrat at 95% CI.

2.3.2. SPSS based linear regression model. In the second approach, we used Statistical Data analysis by LR Model with the help of IBM SPSS Version 21.0, for Windows, version 8.0 in Microsoft Windows 10. In this model, 2 variables were used first HCV percentage (dependent variable) and second periods in months (independent variable). Linear regression (LR) is a very powerful statistical model for long-term analysis only when it is used appropriately based on required assumptions about the dataset. To assess the "goodness of fit" of our LR model we checked the following assumptions of our data.[13]

2.3.2.1. Linearity and outliers. To check the linearity and outliers of our dataset we graphically visualized it by a "simple scattered" plot in SPSS statistics.

2.3.2.2. Residual errors. The sum of Residual errors of the LR model was calculated by saving the residuals in SPSS "data view" window then checkout it by calculating its sum.

2.3.2.3. Autocorrelation. Independence of values that can be defined as error associated with one value is not correlated with the error of any other values was calculated by the Durbin Watson test that is more reliable for time series data.

2.3.2.4. Homoscedasticity. Homoscedasticity (constant variance) was graphically visualized by scatterplot and also checked by the Breusch–Pagan test (analog way) in this test we take a square of residual values, then apply the LR model in which we take the square of residues as a dependent variable and months as an independent variable, in Breusch-Pagan Test our null hypothesis (H_o) data are homoscedasticity and alternative hypothesis (H_a) data are Heteroscedasticity. If analysis of variance (ANOVA) $P < .05$ we reject H_o and accept H_a if $P > .05$ then we accept H_o and reject H_a.

2.3.2.5. Normality. The normality of the residual was graphically checkout by the "Normal probability plot" also called the P-P plot. Statically it was checked by the normality test in this our null hypothesis (H_o) data are normality distributed and alternative hypothesis (H_a) data are not normality distributed. If $P < .05$ we reject H_o and accept H_a if $P > .05$ then we accept H_o and reject H_a.

3. Results

3.1. HCV distribution in district Gujrat

In district Gujrat, the annual percentage of HCV-infected men patients is high with a higher viral load as compared to women. From January 1, 2016 to December 31, 2020 annually HCV infected men percentage was (53.75%), (45.67%), (39.67%), (41.94%), (37.70%) respectively along with average viral load 1.0×10^5 IU/mL, 1.0×10^6 copies/mL. While HCV infected women annually percentage from January 1, 2016 to December 31, 2020 was (53.19%), (43.84%), (39.36%), (35.88%), (31.38%) respectively along with average viral load 1.0×10^4 IU/mL, 1.0×10^5 copies/mL. The average pervasiveness of HCV infection in both men and women was at the age of 40 to 49 (Table S1, Supplemental Digital Content, http://links.lww.com/MD2/A740).

The results showed that overall annually percentage of HCV in 2016 was (53.38%), 2017 (44.45%), 2018 (39.47%), 2019 (38.04%), and in 2020 (33.45%). This percentage is continuously decreasing from 2016 to 2020 with an average of 4.98%. As in 2016 HCV percentage was 53.38% and in 2020 it was 33.45% in these years total of 19.93% HCV infection decrease from District Gujrat. To find out at which month and year this percentage will be zero SARIMA and linear regression models used in this study.

3.2. SARIMA model

In the first time-series data analysis strategy we used the SARIMA model.

3.2.1. Data visualization. After importing, install and running our library we generally visualized the dataset by plotting a graph (Fig. 1). It showed that our time series data has a linear pattern because the set of data cluster less round a straight line due to this the trend is linear. Therefore, our time series data has a linear pattern.

3.2.2. Additive decomposition. To identify the hidden pattern in our time series data an additive decomposition model was selected to decompose the data because data have a linear pattern. It consisted of 3 systematic components including trend, seasonality, and 1 non-systematic component called noise

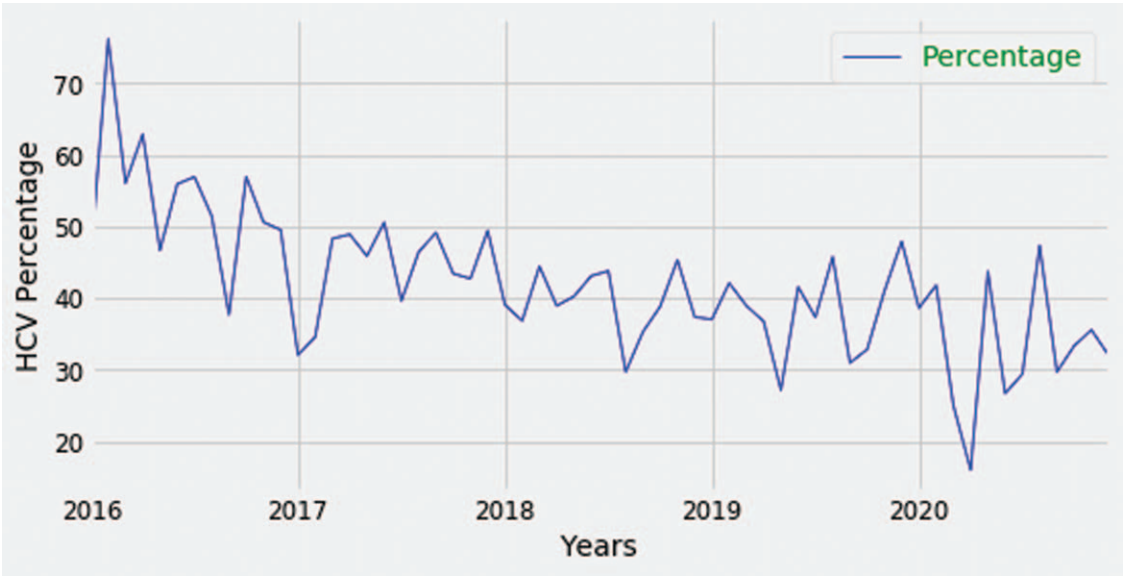

Figure 1. Dataset visualization: January 2016 to December 2020 HCV percentage dataset was visualized by Jupyter Notebook (anaconda3) that showed a linear trend. HCV = hepatitis C virus.

(Fig. 2). This gives us more insight into our data that it has a downward linear trend and also has seasonality where the percentage value maximum twice every year. Therefore SARIMA model was selected as compared with ARIMA model.

3.2.3. Stationary tests. Before modeling it was necessary to check our data are stationary or not because most time series models only work on the assumption if the time series data are stationary. Time-series data are said to be stationary if its statistical properties like mean and variance remain constant over time. We used 2 methods to determine data stationarity, in the first method rolling statistics was used to visualize our data for this we plotted the data with the rolling mean and rolling standard deviation for any case of trends (Fig. 3). It showed that it is constant with time there is no sharp increasing or decreasing trend in our data and rolling mean and rolling standard deviation lines are approximately parallel to the x-axis or horizontal and this data was stationary. In the second method ADF test was imposed on data to check the stationarity. After performing the ADF test at 95% CI, the P-value is $<.05$ so we rejected null hypothesis H_o so the data were stationary. Also, the test statistics are less than the critical values as shown in Table 1. So, the dataset was stationary.

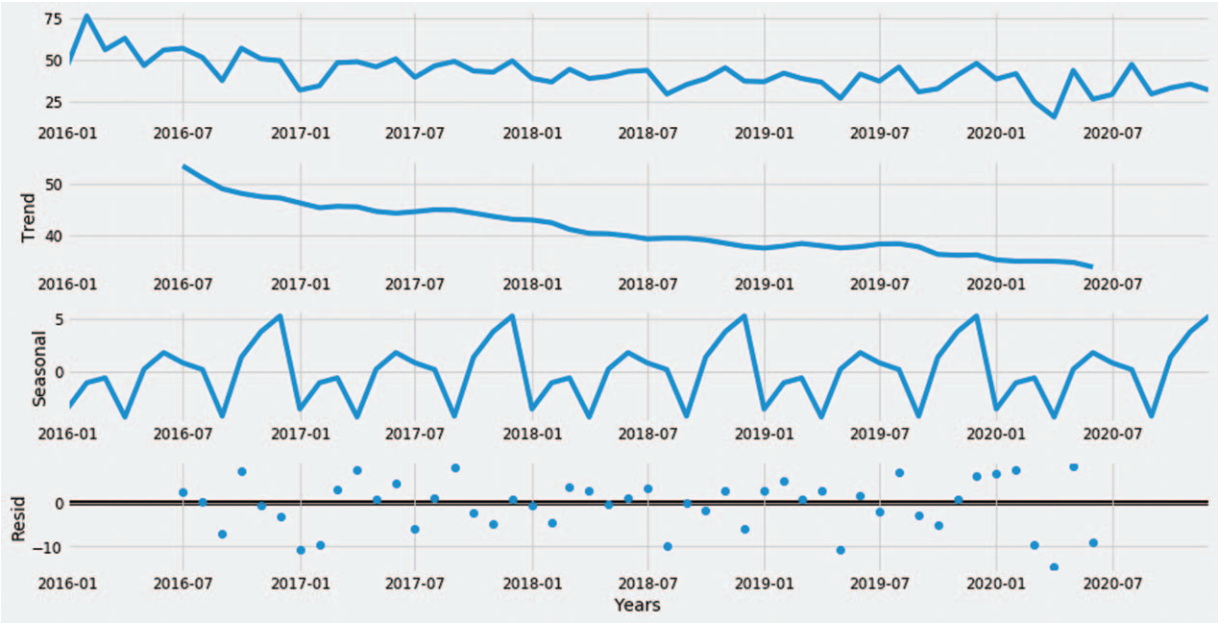

Figure 2. Additive decomposition: this plot gives us a clear idea that our data has linear downward trend and seasonality.

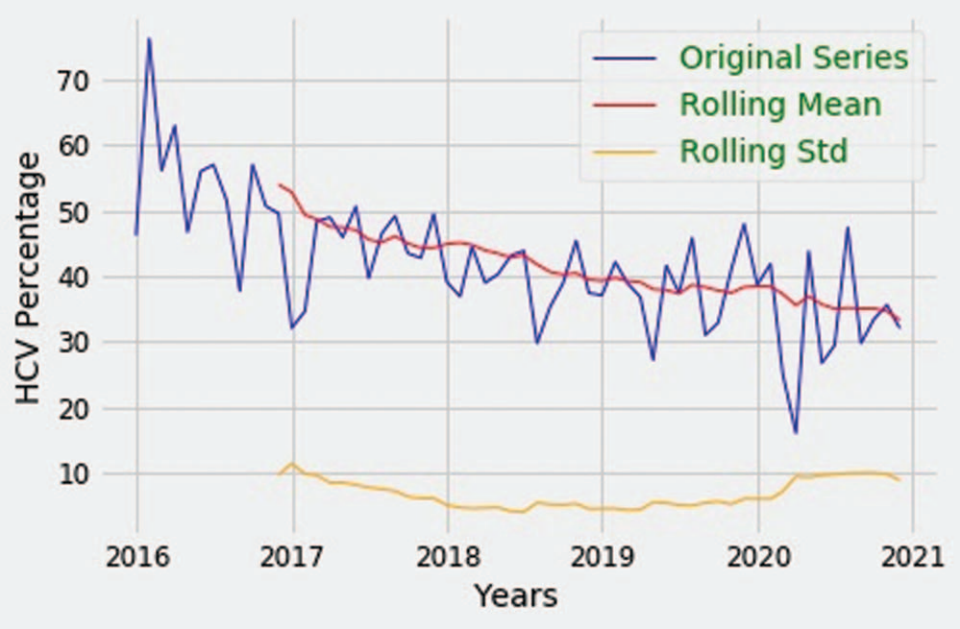

Figure 3. Rolling means and rolling standard deviation: this plot represented that our dataset was stationary.

3.2.4. Model selection by pyramid auto-ARIMA. To find the best parameters for our model we used the Pyramid auto-ARIMA model also called "auto-arima" or "pmdarima." This package automatically or itself generated the best and optimal parameters SARIMAX (0,1,1)(0,1,1,6) by Grid search. This model was selected based on the minimum values of Akaike Information Criterion metrics which was 386.473 (Figure S1, Supplemental Digital Content, http://links.lww.com/MD2/A739). This fitted SARIMAX (0,1,1)(0,1,1,6) model summary as elaborated in Table 2 have significant ma.L1, ma.S.L6, sigma2 (moving average Lag 1, seasonal moving average Lag 6 and variance respectively).

Table 1

Augmented Dickey Fuller (ADF) test.

Test statistics	–4.706185
P-value	.000068
No. of lags used	0.000000
Number of observations used	59.000000
Critical value (1%)	–3.546395
Critical value (5%)	–2.911939
Critical value (10%)	–2.593652

In ADF test $P < .05$ and test statistics value is smaller than all critical values showed dataset is stationary.

3.2.5. Model acceptability. A diagnostic test was performed to check out acceptability and white noise to fit our model (Fig. 4). The white noise process is defined as a sequence of residues that should be uncorrelated and normally distributed around a zero mean. From the diagnostic test, we observed that our fitted model residuals satisfied the requirements of the white noise process and have all properties like, mean close to zero, constant variance, normally distributed and no correlation. In standardized residues, residuals have a mean around zero and constant variance. In the histogram plot, we have seen that red Kernal density estimate (KDE) line followed closely with the standard notation N(0,1) line (here 0 is mean and 1 is standard deviation), and residuals stay much the same across the historical data which indicated that residuals are normally distributed. Q-Q plot showed that the ordered distribution of the residues (blue dots) followed the linear trend of samples taken from a standard normal distribution with N(0,1). This Q-Q plot also indicated that residuals are normally distributed. Correlogram shows that there is no significant correlation in the residual series. From all these, it was suggested that the fitted model was valid and acceptable. We selected this model SARIMAX(0,1,1)(0,1,1,6) to forecast future percentages of HCV.

3.2.6. Validating forecast. We applied the accepted fitted model to predict the value of the test data set at 60% from tail to check its accuracy. It showed (Fig. 5) that our predicted values are so

Table 2

Fitted model summary.

	Coeff	Std. Err.	*z*	$P > \|z\|$	[0.025	0.975]
ma.L1	–0.9051	0.116	–7.817	.000	–1.132	–0.678
ma.S.L6	–0.6345	0.158	–4.007	.000	–0.945	–0.324
sigma2	69.0055	17.349	3.978	.000	35.003	103.008

Fitted SARIMAX model have significant moving average Lag 1, seasonal moving average Lag 6, and variance respectively.

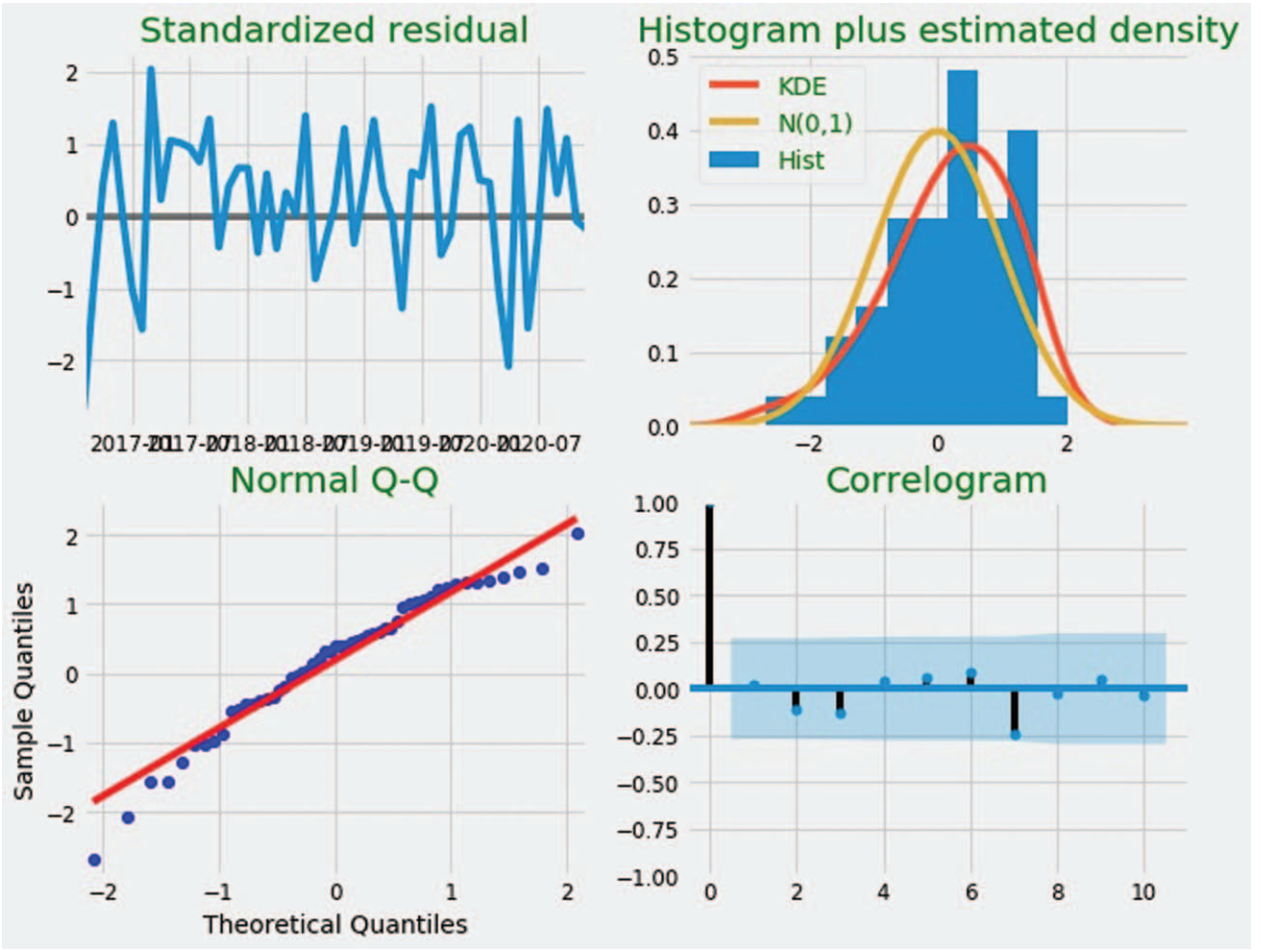

Figure 4. Diagnostic test: it showed that our dataset has a mean close to zero, no correlation, constant variance, and normal distribution properties.

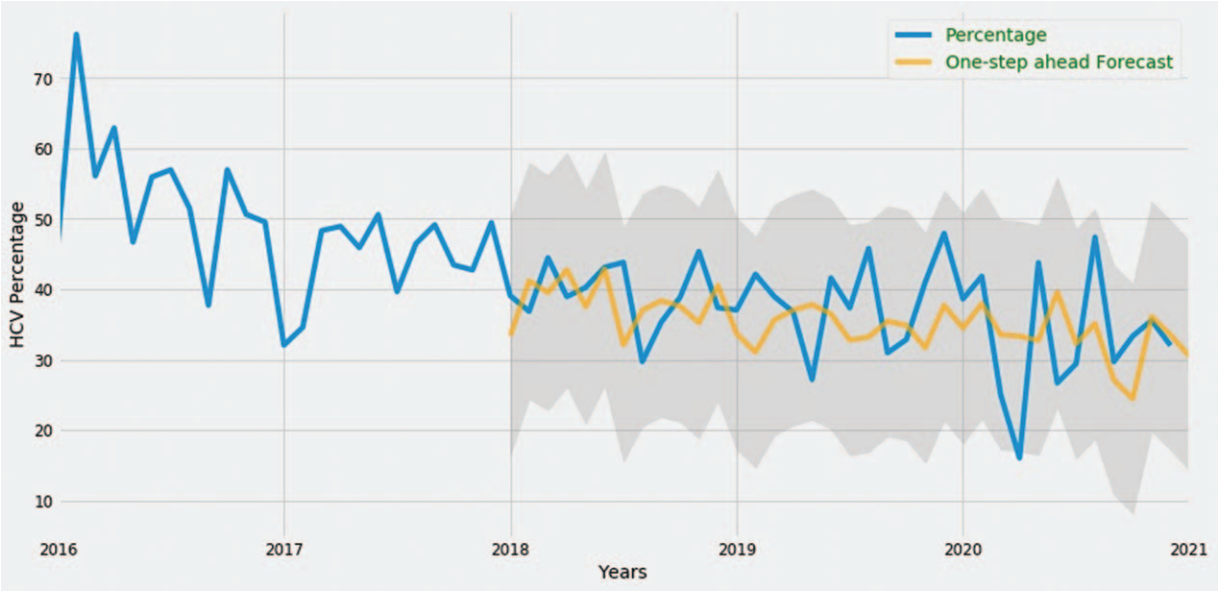

Figure 5. One step ahead forecast: this analysis showed that our predicted values followed the same pattern and have the same similarity as our historical dataset.

close to our original values and predicted values also follow the decreasing seasonal trend same as our original values. To check the accuracy of our model we also find out the MAPE which was 18.39%. This 18.39% MAPE implies the model is 81.61% accurate and valid in predicting the future HCV percentage values.

3.2.7. Forecast analysis. We used our fitted and validate model SARIMAX (0,1,1)(0,1,1,6) to predict future HCV percentage values. Blue lines (Fig. 6) show our original percentage values and the red line show forecasted percentage values for the next 10 years. This forecast has captured the seasonal pattern seen in the historical data and replicated it for the next 10 years. The greyed-out area consisted of 95% CI that means the predictions will not cross that area. From this forecasting, the HCV percentage will be zero in June 2029. It means that District Gujrat is on track and following the WHO plan (HCV elimination) and HCV will extinct from District Gujrat before 2030 if the same efforts continue for its elimination.

3.3. Linear regression model goodness of fit

Our dataset adopted all assumptions of the linear regression model as the goodness of fit.

3.3.1. Linearity and outliers. Simple scattered was plotted that showed the linear relationship present between HCV percentage and months and there is no specific outlier (Fig. 7A).

3.3.2. Residual errors. In our dataset, the sum of residual errors was zero that not violated this assumption.

3.3.3. Autocorrelation. In Durbin Watson (DW) test, if the tested result value of DW fall 1.5 to 2.5 then there is no autocorrelation present in data and if the tested result value is below 1.5 then autocorrelation is present. In our analysis, the Durbin Watson value was 2.054 it means that there was no autocorrelation present in our time series dataset.

3.3.4. Homoscedasticity. To check out the homoscedasticity "Scatterplot" (Fig. 7B) was plotted, it was shown in the figure that our data meet the assumption of homoscedasticity because observed values were not dispersed and these were along the line of best fit. In Breusch–Pagan (analog) test our ANOVA $P > .05$ ($P = .461$) so we were failed to reject the null hypothesis and our data have homoscedasticity instead of heteroscedasticity.

3.3.5. Normality. The normality of the residual was visualized by the P-P plot (Fig. 7C). In this data are almost fall and fit the line and residues were normally distributed. Statically to check out the normality we used Kolmogorov Smirnov (KS) and Shapiro Wilk (SW) tests. In our analysis, KS test $P = .200$ and SW test $P = .357$ in these both tests our residues are not significant due to this we accepted the null hypothesis (H_o) which mean data were normality distributed.

3.4. Linear regression model outputs

In the second approach, we used statistical data analysis by linear regression model. First, we fit our model and then used it for prediction analysis. In this model based on 2 variables (months and HCV monthly percentage), we applied linear regression with 99% CIs level. We obtained highly significant model (P-vale = .00) that was $Y = -0.379X + 53.378$, in this model "Y" was dependent variable (HCV percentage), "X" was independent variable (months), "–0.379" was slope (P-value = .00) and "53.378" was intercept (P-value = .00). In our model $R = 0.666$ showed that there was a good relationship between dependent and independent variables. By using this model we analyzed that in August 2027 HCV percentage will be zero in District Gujrat, it is on track to WHO plan of HCV elimination before 2030.

4. Discussion

This study revealed that in District Gujrat concerning the male and female ratio, the male's ratio is always high as compared with females. From year 2016 to 2020 male to female ratio was (53.75:53.19), (45.67:43.84), (39.67:39.36), (41.94:35.88) (37.70:31.38) respectively. In this region, HCV pervasiveness range is 30 to 59 age with an average of 40 to 49 same as previous studies mentioned, in addition to this percentage of HCV infected

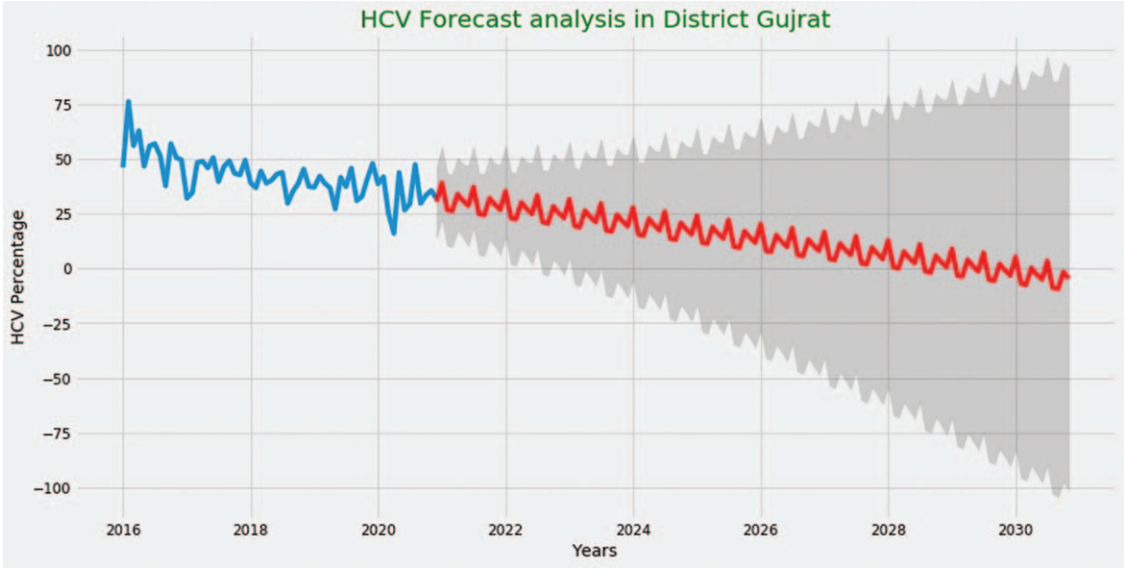

Figure 6. Forecasting: forecast analysis indicated that HCV percentage will become zero in June 2029 (95% confidence interval).

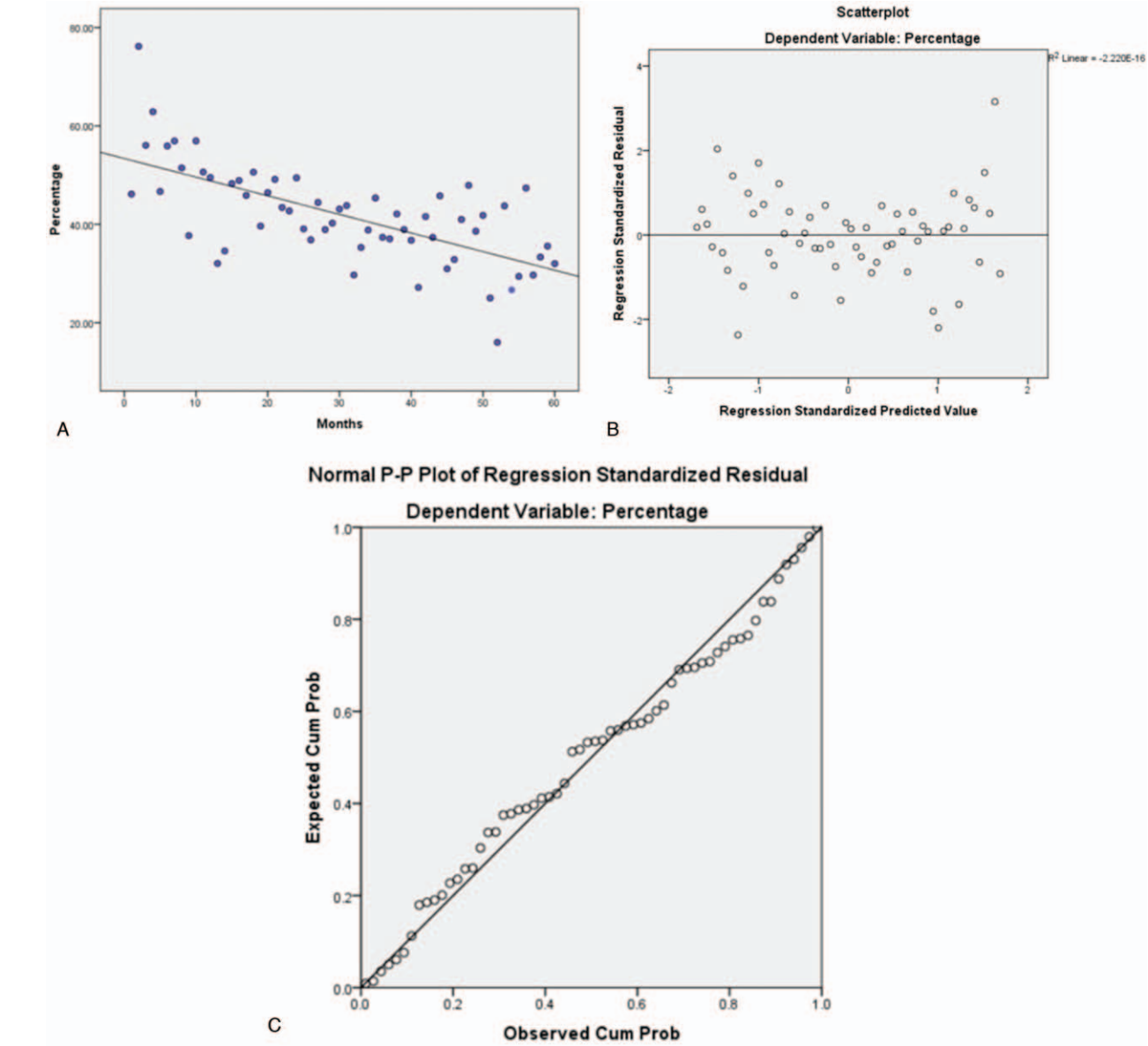

Figure 7. Assumptions of linear regression model. (A) The "simple scattered plot" showed that our data were not violating the assumption of linearity. (B) Scattered plot showed a graphical representation of our homoscedasticity data that were adopted most important assumption of our linear regression model. (C) P-P plot illustrated that residues were around about fitted the line and these were normally distributed.

men is high with a higher viral load as compared with women from the last 5 years. In Pakistan, higher HCV prevalence has been recorded in 40 to 50 years of age people and the anti HCV antibodies ratio in men (15.09%) is higher than in women (12.3%).[14] In Punjab, a province of Pakistan prevalence of anti-HCV is 8.9%.[8] In this province HCV infected female to male ratio was 18.7:17.5 respectively in which maximum patients were at the age of 41 to 61 years (33%), HCV seroprevalence in District Gujrat was 7.4% and in Northern Area (25.7%), Rawalpindi (2.45%), Multan (4.06%), Karachi (4–6%), NWFP (5%), Islamabad (5.31%), Faisalabad (20.89%), and Mardan (9%).[15,16] In this study, it was found that in District Gujrat annual percentage of HCV in 2016 was (53.38%), 2017 (44.45%), 2018 (39.47%), 2019 (38.04%), and in 2020 (33.45%). This percentage is continuously decreasing from 2016 to 2020 with an average of 4.98%.

To eliminate this infection before 2030 WHO settled a target but 76% of countries are unable to meet this target due to low literacy rates, little awareness, and ignorable HCV screening. Pakistan cannot be achieved the WHO target until 4.9×10^5 infected patients are treated per year for a decade.[17] We can meet this target if we annually diagnose 9.0×10^5 HCV-infected persons and treat 7.0×10^5 patients.[18] Pakistan just can meet it if we expend 9.0% of health expenditures for HCV screening and treatment per year.[19] Until now there is no study to find out either Pakistan is on track to achieve the WHO target are not due to the unavailability of comprehensively proper and authenticated data from all over regions of Pakistan. But HCV elimination is possible because the virus can be easily detected by Polymerase Chain Reaction (PCR), Unlike SARS CoV-2, HCV is a blood born disease therefore it lacks a non-human reservoir, by interrupting all modes of transmission and successful HCV

treatment that revolutionized after the development of direct-acting antivirals with short duration, very high SVR, and minimum side effect.[20] Most important, easy availability of DDA for Hepatitis C treatment in Pakistan. Therefore by considering all these parameters if such analyses perform in the future at the Districts and Provinces level a clearer image will be seen that either country is on track to achieve HCV elimination before 2030 or not. Different machine learning techniques can be used to find out the HCV extinction analysis. In this study, the SARIMA and linear regression models used to predict the HCV percentage and extinction analysis in District Gujrat these models showed their effective modeling process.

SARIMA model deals with seasonality and trends and has more long-term accuracy than the ARIMA model. One another major advantage of the SARIMA model is that it can scrutinize the high accuracy for both long-term and short-term forecasts in time series, this model demanded a large amount of historical data with continuity to fit it.[21] SARIMA model has a good panorama and has been used in the forecast analysis of several diseases. When a fitted and acceptable model is obtained, it can be used to forecast any given number of future time intervals.[22] The combination of the SARIMA-NARNN model for prediction analysis did not always provide better estimates than the single model.[23] The seasonal ARIMA (SARIMA) model was used to predict dengue cases in the population of São Paulo State, Brazil. It was suggested that it is a very effective and reliable predictive tool for disease control and prevention.[24] The Artificial Neural Networks have the best accuracy performance, it is used for modeling the non-linear problems, Artificial Neural Networks specific non-linear functions within the time series data may not be explained well in practice.[25] SARIMA model has purposed a useful tool to monitor epidemics.[11] This study has possible limitations, it is a single District level study with an average number of patient's data. We recommend that such study should be conducted in all districts and provinces level of all countries along with the vast number of consistent historical data of patients to monitor HCV elimination track. Secondly, short term historical data does not give good prediction although different approaches are being used to interpret data according to the nature of datasets there should be some latest single or hybrid model approaches that can easily apply and give accuracy and validity also for short terms time-series data to predicts, monitor, and manage the epidemics. SARIMA and linear regression models can be used to predict HCV elimination if a large number of the historical dataset is available and it does not violate any assumption. In this study, we used SARIMA and Linear Regression models to find out HCV track of District Gujrat we did not do the comparison of these models.

5. Conclusion

Based on both significantly fitted SARIM and Linear Regression models it was concluded that HCV percentage will be zero before 2030 and District Gujrat is on Track to achieve the WHO plan of HCV elimination before 2030.

Acknowledgments

Author's Acknowledge Dr Hafiz Khalid Jameel (MBBS, D.Derm, FCPS) and all supporting staff of Dr Mujahid Lab and Biochemistry Lab of the University of Gujrat for their help during sample collections and storage.

Author contributions

Conceptualization: Muhammad Rashid, Hammad Ismail.
Data curation: Muhammad Rashid, Hammad Ismail.
Formal analysis: Muhammad Rashid, Hammad Ismail.
Investigation: Muhammad Rashid, Hammad Ismail.
Methodology: Muhammad Rashid, Hammad Ismail.
Project administration: Muhammad Rashid, Hammad Ismail.
Resources: Muhammad Rashid, Hammad Ismail.
Software: Muhammad Rashid, Hammad Ismail.
Supervision: Hammad Ismail.
Validation: Muhammad Rashid, Hammad Ismail.
Visualization: Muhammad Rashid, Hammad Ismail.
Writing – original draft: Muhammad Rashid, Hammad Ismail.
Writing – review & editing: Muhammad Rashid, Hammad Ismail.

References

[1] Tariq S, Batool Z, Jahanzeb V, Tariq Q, Durrani SH. Hepatitis C virus genotype distribution amongst HCV positive patients presenting at a private tertiary care hospital of Peshawar, Khyber Pakhtunkhwa. J Rehman Med Inst 2016;2:33–9.
[2] Afzal MS, Shah ZH, Ahmed H. Recent HCV genotype changing pattern in the Khyber Pakhtunkhwa province of Pakistan; is it pointing out a forthcoming problem? Braz J Infect Dis 2016;20:312–3.
[3] Aaron GL, Huma Q, Hassan M, et al. Curbing the hepatitis C virus epidemic in Pakistan: the impact of scaling up treatment and prevention for achieving elimination. Int J Epidemiol 2018;47:550–60.
[4] Messina JP, Humphreys I, Flaxman A, et al. Global distributionand prevalence of hepatitis C virus genotypes. Hepatology 2015;61:77–87.
[5] World Health Organization (WHO). Combating hepatitis B and C to reach elimination by 2030 [Internet]. Geneva: WHO; c2016. Available at: https://www.who.int/hepatitis/publications/hep-elimination-by-2030-brief/en/. Access February 5, 2021.
[6] Gamkrelidze L, Pawlotsky JM, Lazarus JV, et al. Progress towards hepatitis C virus elimination in high-income countries: an updated analysis. Liver Int 2021;41:456–63.
[7] World Health Organization (WHO). National Hepatitis Strategic Framework (NHSF) for Pakistan 2017-2021. Geneva: WHO; c2017. Available at: http://phrc.org.pk/assets/pakistan-s-national-hepatitis-strategic-framework–09-01-2018. Accessed February 5, 2021.
[8] Government of Punjab (GoP). The Punjab Hepatitis Act 2018 [Internet]. Geneva: GoP; c2018. Available at: http://punjablaws.gov.pk/laws/2704.html#_ftn1. Accessed: February 5, 2021.
[9] Survey Monkey. Sample size calculator [Internet]. Geneva: SM. Available at: https://www.surveymonkey.com/mp/sample-size-calculator/. Accessed February 5, 2021.
[10] George EP, Jenkins GM, Reinsel GC. Time Series Analysis Forecasting and Control, 4th ed. Holden: Day; 1976.
[11] Maoa Q, Zhanga K, Yanb W, Chenga C. Forecasting the incidence of tuberculosis in China using the seasonal auto-regressive integrated moving average (SARIMA) model. J Infect Public Health 2018;11:707–12.
[12] Lee-Ing T. Forecasting field failure data for repairable systems using neural networks and SARIMA model. Int J Qual Reliab 2005;22:410–20.
[13] Casson RJ, Farmer LDM. Understanding and checking the assumptions of linear regression: a primer for medical researchers. Clin Exp Ophthamol 2014;42:590–6.
[14] Muhammad N, Jan MA. Frequency of hepatitis "C" in Buner, NWFP. J Coll Physicians Surg Pak 2015;15:11–4.
[15] Ahsan A, Khan AZ, Javed H, Mirza S, Chaudhary SU, Hussan SSU. Estimation of hepatitis C prevalence in the Punjab province of Pakistan: a retrospective study on general population. PLoS One 2019;14:1–12.
[16] Muzaffar F, Hussain I, Haroon TS. Hepatitis C: the dermatologic profile. J Pakistan Assoc Dermatol 2008;18:171–81.
[17] Ayoub HH, Abu-Raddad LJ. Treatment as prevention for hepatitis C virus in Pakistan: mathematical modeling project. Br Med J 2019;9:1–9.
[18] Chhatwal J, Chen Q, Wang X. Assessment of the feasibility and cost of hepatitis c elimination in Pakistan. JAMA Netw Open 2019;2:1–12.
[19] Lim AG, Walker JG, Mafirakureva N, et al. Effects and cost of different strategies to eliminate hepatitis Cvirus transmission in Pakistan: a modelling analysis. Lancet Glob Health 2020;8:440–50.

[20] Haqqi A, Munir R, Khalid M, Khurram M, Zaid M. Prevalence of Hepatitis C virus genotypes in Pakistan: current scenario and review of literature. Viral Immunol 2019;32:402–13.
[21] Tingting F, Risto L. Evaluation of multiple linear regression model and SARIMA model in forcasting heat demand fordistrict heating system. Appl Energy 2016;179:544–52.
[22] Qinqin X, Runzi L, Yafei L, et al. Forecasting the incidence of mumps in Zibo city based on a SARIMA model. Int J Environ Res Public Health 2017;14:925–36.
[23] Zhou L, Zhao P, Wu D, Cheng C, Huang H. Time series model for forecasting the number of new admission inpatients. BMC Med Inform Decis Mak 2018;18:1–11.
[24] Martinez AZ, Silva EAS. Predicting the number of cases of dengue infection in Ribeirão Preto, São Paulo State, Brazil, using a SARIMA model. Rio de Janeiro 2011;27:1809–18.
[25] Zhang X, Liu Y, Yang M, Zhang T, Young A, Li X. Comparative study of four time series methods in forecasting typhoid fever incidence in China. PLoS ONE 2013;8:e63116–27.