Pneumonia and Influenza Prediction Model Using Time Series

Yinzhi Chen

2022-04-14

Abstract

One of the biggest pneumonia and influenza flu outbreaks in the 20th century occurred in 1968, it is one of the fastest spreading diseases which could cause death in the last century. This influenza pandemic was caused by the influenza A (H3N2) virus, which spread to the United States in September 1968. The virus is highly lethal, killing over 1 million people worldwide and 100,000 in the United States.(1) This connects to our situation of the pandemic, Covid 19. It has been a global issue in recent 3 years, it brings panic to the people and also harms human health, and has a certain fatality rate. In this case, creating a predictable table for the death rate of pneumonia and influenza from 1968 to 1978 could be helpful for doctors and specialists to prevent the virus from spreading and prepare future treatments for patients. In the report, we will use the flu data from astsa package to build a seasonal ARIMA model. Then applied the significance tests for the fitted model and used the best-fitted model for future values prediction. Through the observation of the fitted model, there exists a seasonal trend which indicates most of the monthly deaths of flu were around winter. Meanwhile, we observe that there is a slowly decreasing trend of monthly deaths in the US, which may be caused by the tendency of world medical development, or the government's effort in flu prevention. The results indicate that the monthly death rate of flu has been continuously decreasing, so the US people took reliable actions to decrease the death rate of the flu. It indicates the success of governments control and doctors' effort. Meanwhile, for generating an accurate model, other factors that could influence the fitted model should also be considered.

keywords: flu, influenza, reduce, time series, fitted model

Introduction

As the covid 19 pandemic that our society is experiencing right now and the flu happened from 1968 to 1978, it is drastically important that the virus is affecting our life severely, it threatens all human beings' health all over the world. Building a reliable and accurate model for predicting future monthly deaths caused by the flu could prevent extreme virus spreads and reduce the number of deaths caused by the flu. For building the model, upload the data set "flu" from the astsa package which collected the United States monthly number of deaths caused by pneumonia and influenza per 10,000 people from 1968 to 1978. The number of people who have died because of pneumonia and influenza has been boosted since the H3N2 influenza virus occurred in 1968 spread from Hong Kong(2). So I hope this report could be helpful for predicting the future number of deaths caused by the flu, and helps doctors and experts to prevent pneumonia and influenza and reduce the number of deaths. In the following part of the report, I fitted ARIMA model and make predictions of future values for statistical method, then conclude the results and discuss the findings at last.

Statistical method

Initially, we begin by generating a monthly time series plot of monthly pneumonia and influenza deaths in the U.S from 1968 to 1978. From the first figure, we could observe that the time series plot exists a slowly decreasing trend with slightly decreasing variance.

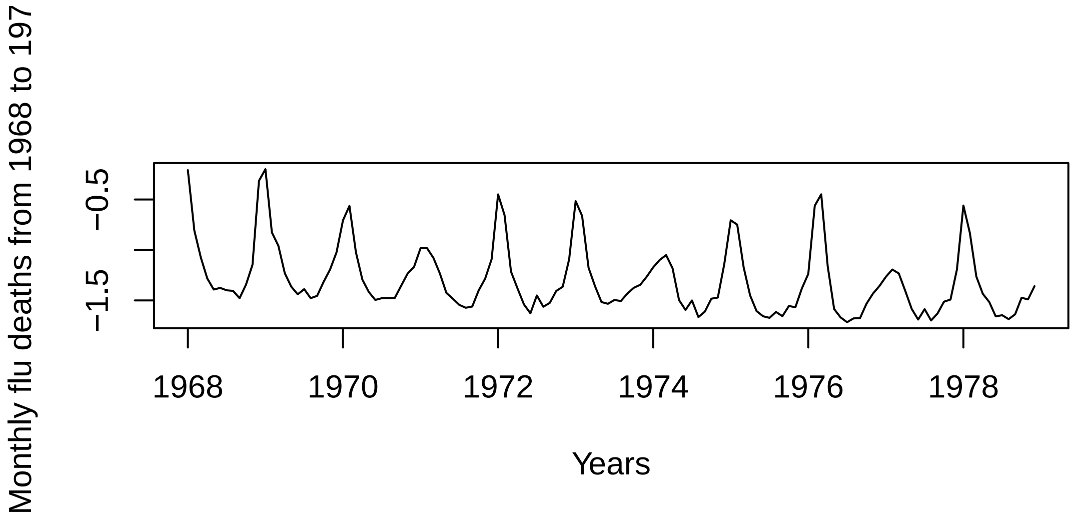


Figure 1: Monthly pneumonia and influenza death in the US fom 1968 to 1978

For creating a stationary process, certain actions must be taken in order to ensure the time series is constant, we could use transformation or difference. Applying transformation and difference to the data, we initially log transform the data to reduce and remove the data's skewness, then remove the slowly decreasing trend. Therefore, we could easily observe that the time series plot and ACF plot exists a seasonal trend through the time series as well, it is clear that there is still persistence in the seasons which follows a repeating trend every twelve months, then a difference of seasonal trend lag equals twelve is also mandatory.

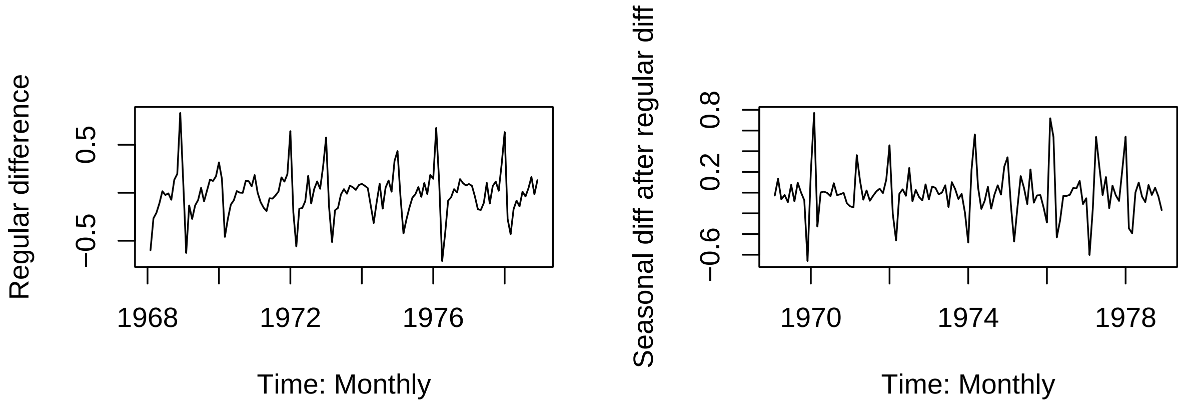
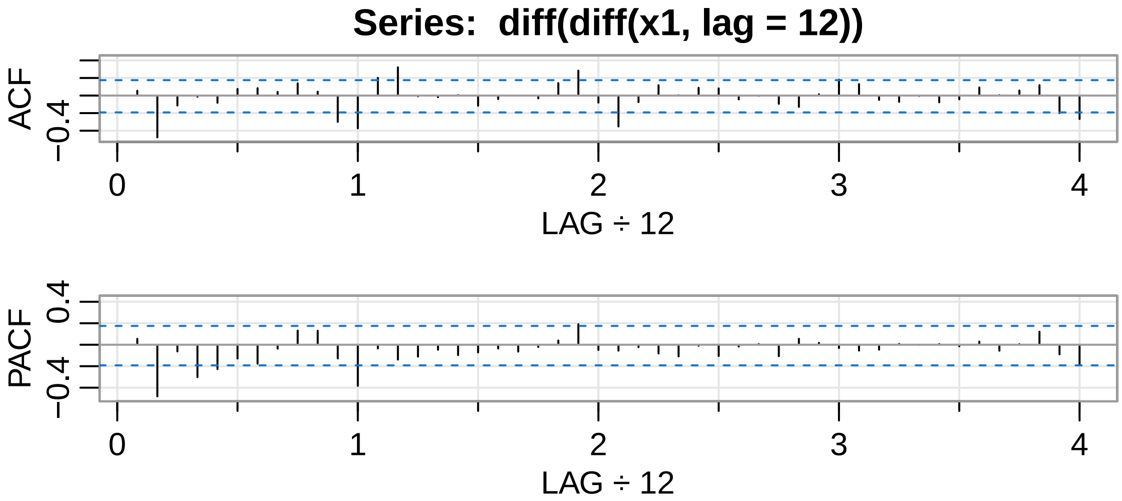


Figure 2: Monthly time series plot after regualr differencing and seasonal differencing after regular differencing

After regular and seasonal differencing, we could tell that the time series data appears stationary with a constant mean and variance. Thus, we could use the differenced data for fitting the sample ACF and PACF models.



For fitting a suitable SARIMA model for the dataset, we have to identify the cut offs of the ACF and PACF graph for both non-seasonal and seasonal components. First, observing the non-seasonal components, in the ACF plot, it cuts off after lag = 2, PACF cuts off after lag 2. Then, we could estimate p = 2 and q = 2. Since we used regular differencing once which indicates d = 1. For seasonal components, ACF cuts off after 2s which s = 12, PACF cuts off after 1s. Therefore we could estimate that P = 1, Q = 2. Since we used seasonal differencing once which indicates D = 1. Thus, we could start to fit the first model ARIMA(2, 1, 2) x (1, 1, 2)12. The below figures show the analysis of the model with its t-table.

	Estimate	SE	t.value	p.value
ar1	0.2239	0.1700	1.3175	0.1904
ar2	-0.2093	0.1337	-1.5655	0.1203
ma1	-0.4572	0.2601	-1.7576	0.0815
ma2	-0.5358	0.2114	-2.5351	0.0126
sar1	-0.7067	0.0941	-7.5093	0.0000

	Estimate	SE	t.value	p.value
sma1	0.1389	0.2098	0.6619	0.5094
sma2	-0.8607	0.1930	-4.4602	0.0000

By observing the initial model's T-table, the model has some insignificant p values, which means these variables with insignificant p values could be rejected, so we should try dropping one of the parameters from p and q in order to modify the variables to be significant. Thus, try dropping the p-value by 1 and fit the model ARIMA(1, 1, 2) x (1, 1, 2)s

Comparing the 2 fitted models, the first ARIMA model contains some insignificant p values, then drop parameters to ensure the p values are significant. The second ARIMA model drops p-value by 1 which more p values become significant and all of the p-values points for Ljung-Box statistics are above the blue line. By observing the first and third model, both of the standard residuals doesn't follow any patterns or trend; The ACF of both series don't exceed the blue dash lines; Both normal QQ plots are mostly around the residual line with very few outliers; Both of the residual diagnostics for Ljung-Box statistics are above the significance level. The above features indicate that both models' results are significant. Thus, calculating AIC, BIC and AICc will determine a better model.

	AIC	BIC	AICc
ARIMA(2,1,2)x(1,1,2)12	-2.406192	-2.219361	-2.397713
ARIMA(1,1,2)x(1,1,2)12	-2.416641	-2.253163	-2.410338

From the table above, the second model ARIMA = (1,1,2) x (1,1,2)s is chosen since it has smaller AIC, BIC and AICc values. Then the fitted model could be written as below, Xt represents monthly death of flu and Wt represents white noise has N(0,sigmaˆ2)

$$(1-\phi_1 B)(1-\phi_1 b^{12})\nabla\nabla_{12} log X_t = (1+\theta_1 B+\theta_2 B^2)(1+\Theta_1 B^{12}+\theta_2 B^{24})W_t$$

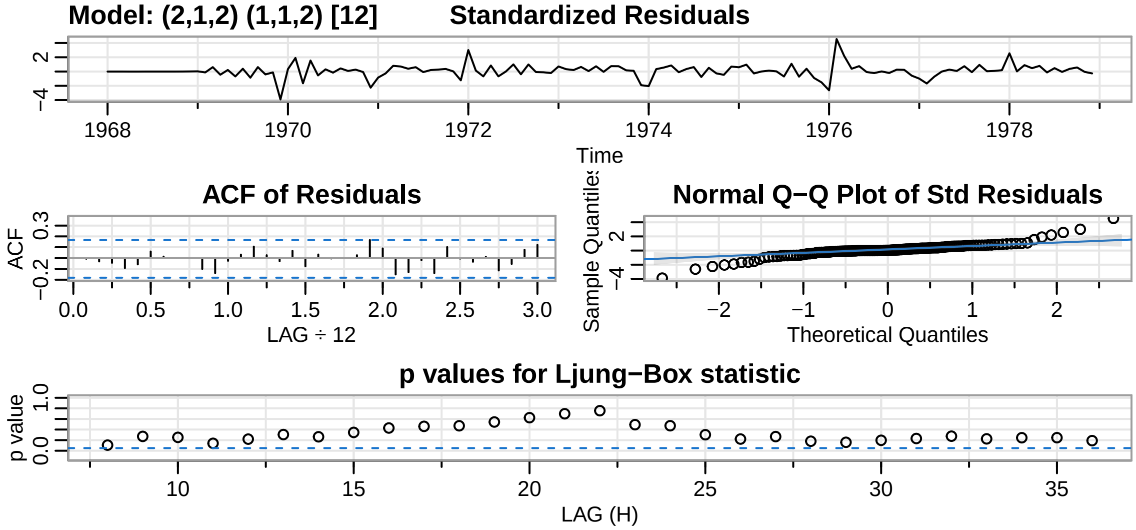


Figure 3: Residual analysis of ARIMA(2,1,2) X (1,1,2)12 for flu

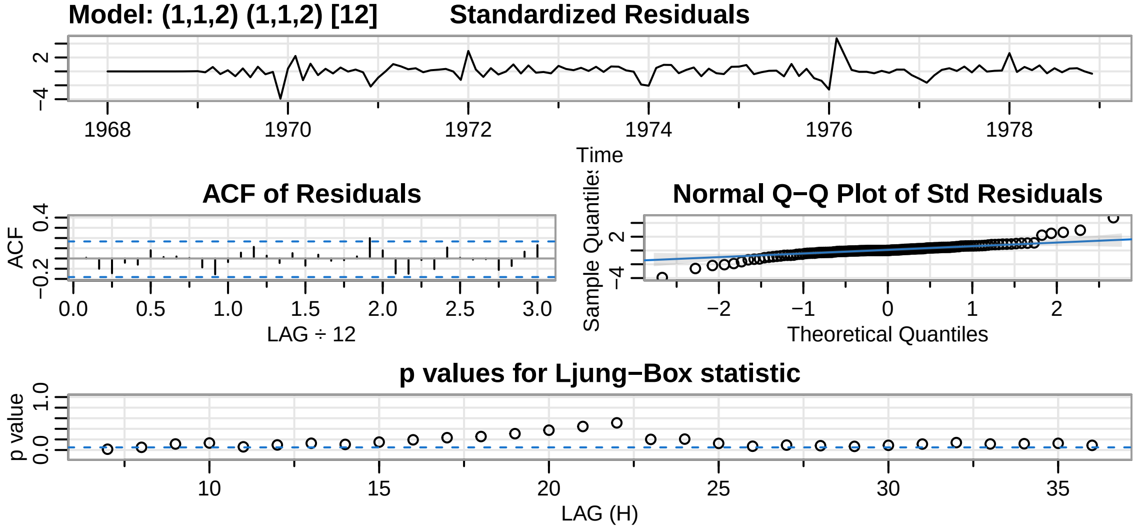


Figure 4: Residual analysis of ARIMA(1,1,2) X (1,1,2)12 for flu

Results

Thus, observing the t-table of the selected model, we could put the parameter estimations with their p-values in the following table:

	Estimate	SE	t.value	p.value
ar1	0.0619	0.1221	0.5072	0.6130
ma1	-0.3024	0.1027	-2.9447	0.0039
ma2	-0.6976	0.0927	-7.5227	0.0000
sar1	-0.7165	0.0923	-7.7594	0.0000
sma1	0.1547	0.2038	0.7593	0.4492
sma2	-0.8452	0.1838	-4.5999	0.0000

The table shows the accurate result of p-values of each parameter since our significance level is 0.05. To determine whether the parameters are significant or not, we should compare the p-values with a significance level of 0.05. Then, the p-value of parameters of AR(1) and SMA(1) is greater than 0.05 but other parameter estimations' p-value is smaller than 0.05. Thus, the final fitted ARIMA(1,1,2)X(1,1,2)12 model is:

$$(1-0.0619B)(1-0.0619B^{12})\nabla\nabla_{12}logX_t = (1+0.3024B-0.6976B^2)(1+0.1547B^{12}-0.8453B^{24})W_t$$

The formula indicates that the monthly flu are affected by above factors with p = 1 and seasonal P = 1 the order of the autoregressive model, d = 1 is the times of differencing, D = 1 is the time of seasonal differencing, q = 2 and seasonal Q = 2 refers to the order of a moving average model, s = 12 represents the seasonal circles every 12 months of the data. . Thus, the fitted model is suitable for predicting the future 10 months of monthly flu deaths in the U.S. Below figure shows the future ten-month forecasting model, the red dots refer to the future 10 months prediction value and the gray shades area refers to the prediction intervals.

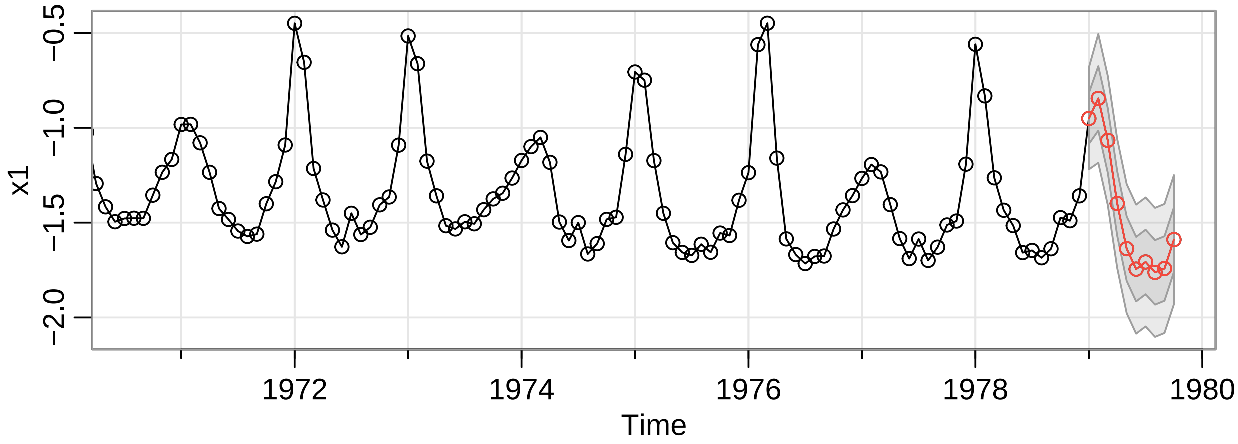


Figure 5: Future 10 months ARIMA(1,1,2)X(1,1,2)12 model forecast

Below is the table of 10 months forecast of ARIMA(1,1,2)X(1,1,2)12 model of log(flu) prediction intervals values.

Prediction	X95..PI.Lower.Bound	X95..PI.Upper.Bound
-0.9511190	-1.214103	-0.6881353
-0.8450708	-1.177775	-0.5123664
-1.0661907	-1.399456	-0.7329252
-1.3992326	-1.732520	-1.0659450
-1.6373810	-1.970670	-1.3040922
-1.7451806	-2.078470	-1.4118917
-1.7074352	-2.040724	-1.3741462
-1.7622833	-2.095572	-1.4289944
-1.7417813	-2.075070	-1.4084923
-1.5894850	-1.922774	-1.2561963

Observing the future 10-month predicted results with the predicted model, it's clear there exists a slowly decaying trend, it indicates monthly pneumonia and influenza death will keep decreasing in the next 10 months. Meanwhile, it's also clear that there exists a periodic seasonal trend with a peak during winter and a valley during summer. It's logically and scientifically normal since a colder climate is a key factor to make more people susceptible to infect pneumonia and influenza. From the 95% prediction interval, it represents the value ranges the response will fall into, as the prediction of values goes further, the prediction interval becomes larger since the further time goes, the bigger error prediction will have. But for the 10 prediction points, the prediction interval is narrow which indicates the result is relatively reliable and accurate.

Next, in order to find out the first three dominant periods of a spectrum, we could use a periodogram to identify the dominant periods of the flu time series.

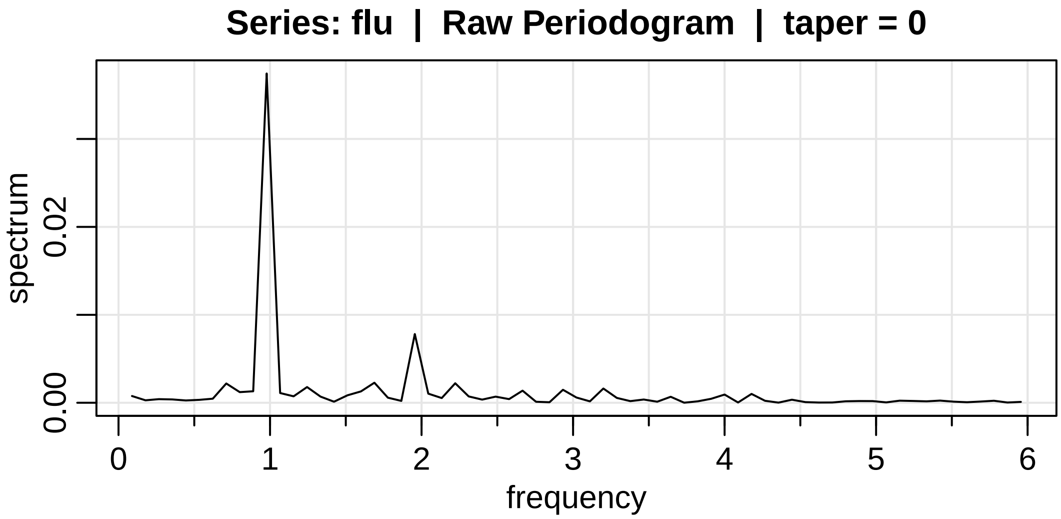


Figure 6: Perdiodogram of flu

Through the observation of periodogram, it has a peak around 0.9, we have a seasonal period s = 12, so compare 0.9 with 1/12 which is relatively close. Then the first three dominant periods could be analyzed using spectrum, the 95% confidence interval predominant spectrum is shown below.

Series	Dominant.Freq	Spec	lower	Upper
flu	0.9778	0.0374	0.0101386	1.4772211
flu	1.9556	0.0078	0.0021145	0.3080835
flu	1.6889	0.0023	0.0006235	0.0908451

Observing the 95% confidence interval predominant spectrum, the three periods all have a very small range between lower and upper intervals which could be used for determining significance. However, there exist overlaps between each period's confidence interval. The intervals of the first and second dominant periods have overlapped and the intervals of the second and third dominant periods also have overlapped. As a result, we cannot establish the significance of the first peak since the periodogram ordinate is 0.0374, which lies in the confidence interval of the second and third peaks; we cannot establish the significance of the second peak since the periodogram ordinate is 0.0078, which lies in the confidence interval of the second and third peak;

We cannot establish the significance of the first peak since the periodogram ordinate is 0.0374, which lies in the confidence interval of the second and third peak. As a result, all of the three dominant spectrum cannot be used to establish significance.

Discussion

In conclusion, combining the original data with prediction data, we observe that monthly pneumonia and influenza deaths in the U.S have a slowly decaying trend which means there will be fewer and fewer people died due to pneumonia and influenza. There might be several reasons for this. First, through reading the report of MedicalNewsTodays report about Covid 19(3), the Omicron cause less severe disease than Delta is because the increased transmissibility of virus will decrease its damage, so the same as pneumonia and influenza, the virus damage for people will keep decreasing since the virus wants to spread more. Second, the government, experts, and doctors are experienced in preventing the spread of virus such as wearing masks, decrease social interaction and develop efficient treatment. The predicted model is not completely accurate since there still exists limitations, since the sample size only has 132 observations which are relatively small, small sample size may result in inaccurate future value predictions. Additionally, there are a few outliers in the normality QQ plot which may affect the performance of the model. Lastly, different places have different amounts of flu-infected patients. If it's possible to record the flu infections number of each state in the US instead of the total number for the whole country it would be better to make predictions based on different states' situations.

Bibliography

(1)Rogers, K. (n.d.). 1968 flu pandemic. Encyclopædia Britannica. Retrieved April 13, 2022, from https://www.britannica.com/event/1968-flu-pandemic

(2)Centers for Disease Control and Prevention. (2019, January 2). 1968 pandemic (H3N2 virus). Centers for Disease Control and Prevention. Retrieved April 14, 2022, from https://www.cdc.gov/flu/pandemic-resources/1968-pandemic.html

(3)Shukla, D. (n.d.). Does Omicron cause less damage to the lungs? Medical News Today. Retrieved April 15, 2022, from https://www.medicalnewstoday.com/articles/covid-19-does-omicron-cause-less-damage-to-the-lungs