Forecasting Top 3 Cancer Rate in Each Gender for New Insurance Product Design

NTHU BAFT Team 4
104064538 Kuan-Yu Chen 104078504 I-Chun Chao
104078701 Edward Song 105078517 Yi-Wei Lai
105078510 Yen-Ju Tseng

Lecturer Prof. Galit Shmueli

Executive Summary

There are some reasons to work on this topic. Medicare doesn't compensate that much because of the strict compensation policy. Due to the policy constraint, medical report for compensation is required. The size of tumor cannot really show the severity which is contrary to the policies nowadays. And the most expensive treatment, target therapy content, will not be paid. That's why the cancer insurance products come out – try to complement the gap that Medicate doesn't cover. However, the general cancer insurance actually can't include all the details since the treatments of the cancers depend on the type of the cancer. Therefore, patients' pain points still remain and we should make some efforts on it.

Our business goal is to provide the cancer crude rate forecasts for better insurance products design which helps patients to gain compensation they deserved. Specific cancer may occur in certain gender with greater possibilities. Therefore, we decide to forecast among different genders and choose Top 3 cancer crude rate forecasts in each gender as prototypes for Insurance Companies. Finally, we set our forecasting goal for providing Insurance Companies the forecasts of TOP 3 cancer crude rate each gender for 4 years.

The yearly cancer crude rate dataset that we use comes from the government website. After trying various methods for each series, we include different external information series. However, we didn't use those series to optimize our forecast models at last because the models' performance didn't get well. We choose one best model among Naive, Holt, Regression, Neural Network, and ARIMA. Finally, we generate cancer crude rate forecasts, use empirical rolling-forward method to plot histogram and find out the forecast intervals. We then compare the best model to the ensemble method.

The recommendations for forecasts implementation are stated as four main points. First, the forecasting models we built are the prototypes. The ensemble model is for generating forecasts for all cancers and we set the best one model as premium package for Insurance Companies. Compare to ensemble model, the performance and the forecasting accuracy of choosing the best one model on each cancer would be higher. Second, Insurance Companies can design better cancer portfolio insurance. By doing so, customers may receive enough compensation they deserved. Third, as long as people know that these Insurance Companies provide the new products which are meet their needs, Insurance Companies can make profits and establish good reputation in the long run because of solving the customer's pain points. Last but not least, Insurance Companies can take some marketing strategies and CSR (Corporate Social Responsibility) actions. For instance, they can promote products to customers with low risk of getting cancer by telling them what factors will cause cancer. The project we proposed can surely benefit both Insurance Companies and customers in the future.

Problem Description

Patient diagnosed with cancer is considered one of the most devastating things in the world. It is not only because the painfulness or uncertainty of treatments but also immense costs for those treatments. Although there are government-based Medicare and other cancer-related insurances in the market already, but those products still cannot fulfil patients' need due to following reasons

i) Medicare doesn't compensate much due to strict compensation policies; meanwhile, the complicated application procedure is also time-consuming.

ii) "Cancer Insurance for all kind of cancers" has lots of limitation for applying the compensation and it does not compensate much.

In order to address this problem, we would like to provide insurance companies with Cancer Crude Rate prediction so they can better design their cancer insurance products that really solve patients' pain points.

Business Goal

We would like to provide insurance companies with our forecasts on specific cancers as prototype so they can use it to predict each incidence of cancers and design better insurance products which can really help patients to get compensation they deserved.

Forecasting Goal

We would like to provide Insurance Companies the forecasts of TOP 3 cancer crude rate in each gender for 4 years since year 2013.

- Y_t :the incidence of Top 3 Cancer in each gender of the series at time period t
- F_t :the forecasted incidence of Top 3 Cancer in each gender .
- Forecast Horizon : 4 years
- k :1,2,3, 4

Data Description

The dataset we used is from Taiwan Ministry of Health and Welfare. The dataset capture yearly cancer crude rate in Taiwan from 1979 to 2013. We chose TOP 3 cancer crude rate in each gender so we have 6 series in total and 35 records in each series.

1. Measurement

Cancer Crude Rate is measuring how many people in 100,000 people who get or die from each cancer in that year. Here is the formula of Cancer Crude Rate:

$$\text{Cancer Crude Rate} = \frac{\text{People Diagonsed or Die from Each Cancer}}{\text{Total Residents in Taiwan}} \times 100{,}000$$

2. Data Limitation

As a result of the data collecting process, there are 2.5 years of time delay which means the latest available data is year 2013. Therefore, we set our forecast horizon as 4 years.

3. Data Demonstration

Year	Gender	Colorectal	Liver	Lung
1979	M	8.32	10.63	11.43
1980	M	9.9	11.87	15.05
1981	M	9.9	13.19	14.83
1982	M	9.17	11.7	13.17
1983	M	9.68	14.96	15.16
1984	M	10.63	15.63	15.73
1985	M	10.94	16.93	16.57
1986	M	12.1	18.59	17.32
1987	M	14.7	22.33	19.96
1988	M	13.28	22.41	18.69
1989	M	8.32	27	22.74

Table 1 Sample of 10 rows of Male Cancer Rates Data

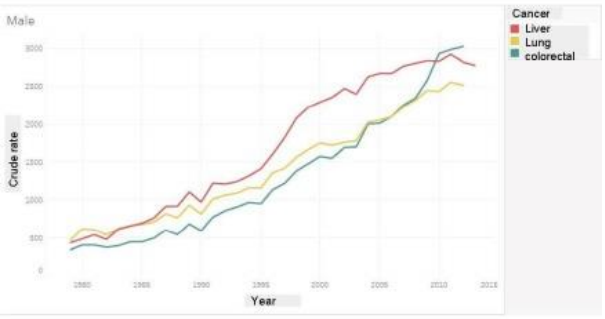


Figure 1 Time series of Male Top 3 Cancers

Year	Gender	Breast	Lung	Colorectal
1979	F	8.4	4.2	5.88
1980	F	9.23	5.08	7.36
1981	F	10.32	5.13	7.02
1982	F	10.29	4.59	7.25
1983	F	11.97	5.8	7.83
1984	F	11.69	5.75	8.36
1985	F	12.65	6.28	8.96
1986	F	12.18	6.72	9.31
1987	F	13.82	7.98	11.32
1988	F	14.8	7.5	11.31
1989	F	17.52	9.46	12.29

Table 2 Sample of 10 rows of Female Cancer Rates Data

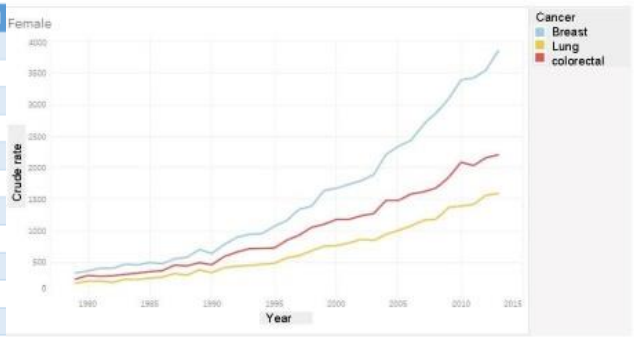


Figure 2 Time series of Female Top 3 Cancers

4. External Information

In order to optimize the performance, we include different external information series into the models, such as yearly working hours, average alcohol consumption, mean of unemployment rate and screening rate etc. (See Appendix A) For instance, Yearly Working Hours is included for colorectal and liver cancer forecasts since we infer that people may have higher chance to get those cancers under high working pressure. However, there are two main issues about implementation. First, lack of records, many of the series we found were too short compared to our original data. Second, time lag, the cancer series only available until year 2013 so we have to apply lag-1 or lag-2 into external series which makes the useful records even less.

● Forecasting Solutions

First of all, we tried different methods for each series and then included some relevant external information by using R. After the process of trial and error, we revealed that the models for 6 series including external information do not have better performance because of the data time lag. Therefore, we decided to exclude the external information.

Second, we ensemble these 4 models together with same weight for each series after building up Naïve, Holt, ARIMA, Linear/Exponential and Regressions models. (See Appendix B) We excluded Neural Networks model because of the overfitting problem. We evaluated the MAPE, RMSE, forecast error, residuals (See Appendix C) and forecast intervals with empirical rolling-forwards method. (See Appendix D) of 6 models and then generate the future forecasts, from 2014 to 2017. We use Naïve Forecast model as our benchmark.

Third, we choose the best one model and ensemble for each. Results are as followed:

Male Top 1 Cancer - Colorectal

ARIMA is the best model with lowest forecast error rate in validation period, test set in R. We generate forecasts from both ARIMA and Ensemble method for male colorectal cancer.

Male - Colorectal	Best One - Arima	Ensemble
t+1 (2014)	77.78144	74.93475
t+2 (2015)	80.37725	76.83563
t+3 (2016)	83.23855	78.80288
t+4 (2017)	86.02115	80.75046

Table 3 Male Colorectal Cancer Real Forecasts

Male - Colorectal	Naive	Holt	Arima	Regression	Neural Network	Ensemble
Training Set_MAPE	8.502046	6.105316	6.193803	17.85355	6.962791	
Test Set_MAPE	13.430617	8.039282	5.434335	19.20403	6.424405	11.52706
Training Set_RMSE	2.846477	1.776792	1.84172	3.830837	2.049433	
Test Set_RMSE	9.883008	6.229475	4.717597	13.981759	4.769468	8.460472

Table 4 Male Colorectal Cancer Forecast Error

Male Top 2 Cancer - Liver

Naive is the best model with lowest forecast error rate in validation period, test set in R. We generate forecasts from both Naive and Ensemble for male liver cancer.

Male - Liver	Best One - naive	Ensemble
t+1 (2014)	69.24	70.617268
t+2 (2015)	69.24	71.332784
t+3 (2016)	69.24	72.076884
t+4 (2017)	69.24	72.74362

Table 5 Male Liver Cancer Real Forecasts

Male - Liver	Naive	Holt	Arima	Regression	Neural Network	Ensemble
Training Set_MAPE	7.772675	5.852836	7.398157	10.74853	4.378219	
Test Set_MAPE	0.989014	7.799285	1.250918	11.89142	1.863237	7.052754
Training Set_RMSE	2.9956897	2.198122	2.891152	3.511842	1.708639	
Test Set_RMSE	0.8640312	6.023188	1.098793	8.710843	1.599094	5.358187

Table 6 Male Liver Cancer Forecast Error

Male Top 3 Cancer - Lung

Holt is the best model with lowest forecast error rate in validation period, test set in R. We generate forecasts from both Holt and Ensemble for male lung cancer.

Male - Lung	Best One - Holt	Ensemble
t+1 (2014)	70.01233	70.03043
t+2 (2015)	72.66219	72.17609
t+3 (2016)	75.31205	73.38719
t+4 (2017)	77.96192	75.04164

Table 7 Male Lung Cancer Real Forecasts

Male - Lung	Naive	Holt	Arima	Regression	Neural Network	Ensemble
Training Set_MAPE	7.465838	5.456136	5.990482	9.127266	6.21259	
Test Set_MAPE	6.245818	5.146714	4.777481	7.135316	5.702441	4.201881
Training Set_RMSE	2.561777	1.813742	1.807999	2.624546	1.962561	
Test Set_RMSE	7.408903	4.236781	4.791777	6.751147	5.160719	5.437507

Table 8 Male Lung Cancer Forecast Error

Female Top 1 Cancer - Breast

Holt is the best model with lowest forecast error rate in validation period, test set in R. We generate forecasts from both Holt and Ensemble for female breast cancer.

Female - Breast	Best One- holt	Ensemble
t+1 (2014)	100.691	103.0086
t+2 (2015)	105.0888	107.4298
t+3 (2016)	109.4866	112.024
t+4 (2017)	113.8843	116.8049

Table 9 Female Breast Cancer Real Forecasts

Female - Breast	Naive	Holt	Arima	Regression	Neural Network	Ensemble
Training Set_MAPE	7.920077	4.95717	5.249586	4.221931	0.1114828	
Test Set_MAPE	11.649576	1.926726	1.356704	8.473194	7.8234304	0.7672796
Training Set_RMSE	3.220786	1.82795	1.918869	1.464369	0.04060605	
Test Set_RMSE	11.512136	1.921466	1.348309	8.81426	7.19656716	0.8273282

Table 10 Female Breast Cancer Forecast Error

Female Top 2 Cancer - Colorectal

ARIMA is the best model with lowest forecast error rate in validation period, test set in R. We generate forecasts from both ARIMA and Ensemble for female colorectal cancer.

Female - Colorectual	Best One - Arima	Ensemble
t+1 (2014)	57.1186	55.78776
t+2 (2015)	58.9772	57.15016
t+3 (2016)	60.8358	58.51255
t+4 (2017)	62.6944	59.87495

Table 11 Female Colorectal Cancer Real Forecasts

Female - Colorectual	Naive	Holt	Arima	Regression	Neural Network	Ensemble
Training Set_MAPE	7.133572	5.169986	4.694019	16.2484	5.393481	
Test Set_MAPE	11.732706	5.534542	3.373292	15.01189	6.489156	8.913107
Training Set_RMSE	1.961618	1.267097	1.298541	2.59347	1.297379	
Test Set_RMSE	6.536232	3.200765	2.221532	8.112562	3.59661	4.860081

Table 12 Female Breast Cancer Forecast Error

Female Top 3 Cancer - Lung

Holt is the best model with lowest forecast error rate in validation period, test set in R. We generate forecasts from both Holt and Ensemble method for male colorectal cancer.

Female - Lung	Best One	Ensemble
t+1 (2014)	42.73827	40.5546925
t+2 (2015)	44.62867	41.7001575
t+3 (2016)	46.51908	42.7299525
t+4 (2017)	48.40948	44.2148675

Table 13 Female Lung Cancer Real Forecasts

Female - Lung	Naive	Holt	Arima	Regression	Neural Network	Ensemble
Training Set_MAPE	8.585232	5.740899	8.321943	2.167366	6.390976	
Test Set_MAPE	7.07409	2.589663	6.084976	6.564967	41.617345	3.471121
Training Set_RMSE	1.551885	1.025399	1.49264	17.18948	0.9542977	
Test Set_RMSE	3.446143	1.14922	2.881238	16.87776	17.5401126	1.694955

Table 14 Female Lung Cancer Forecast Error

● Conclusion

According to our data series, the main constraint we faced is 2.5 years data delay as the result of the process of data collection and confirmation of government. We set our window as 4 years because of 2.5 years data delay with one year insurance product design and implementation. This issue limits us to get the latest information and forecasts with higher accuracy. Apart from that, we sort out four recommendations for insurance companies based on our process above.

1. Top 3 Cancer forecasts as Prototypes - Ensemble vs. Best Model

We use Top3 cancer forecasts methods as prototypes, and take ensemble method as one time basic option and best model as premium. If insurance companies want to use advanced method, the best model, they need to pay more to get more accurate forecast, or they can just use ensemble method to get cancer crude rate forecast.

2. Cancer Portfolio Insurance Design

It's kind of risky for customers to buy certain cancer insurance so we would like to suggest the company to design a portfolio or bundle-like cancer insurance as well. Then, the compensation can be higher (compared to the all cancers) and the risk (of not diagnosed with certain cancer he/she buys) is lower for the customers. The insurance companies can make profits based on the actuary's computation and the bundle design.

3. Long-term Profits

Taiwanese people nowadays find out that the "cancer insurance" can't provide them enough assurance. As long as people know that these Insurance Companies provide the product which meets their needs, they would like to choose it instead of other companies' products. Although designing the products that can't probably "maximize" the company's profit, but they can earn money and establish good reputation in the long run because of solving the customer's pain points and customers know their products really helpful.

4. Marketing Strategy & CSR Campaign

Insurance companies can take some other marketing strategies. For instance, they can promote those products for those lower risky customers. They can not only educate the customers that which kind of food or lifestyle is good for them, but also take the company CSR (Corporate Social Responsibility) actions, such as building a health-education website for spreading the health-concept. These actions can not only help insurance companies to gain better reputation but also reduce the possibility for their customers to get cancers. With reduction of prevalence among customers, companies can gain more profit from it

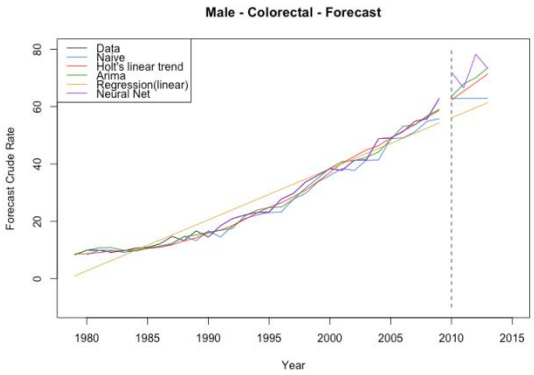
Appendix A – External Information

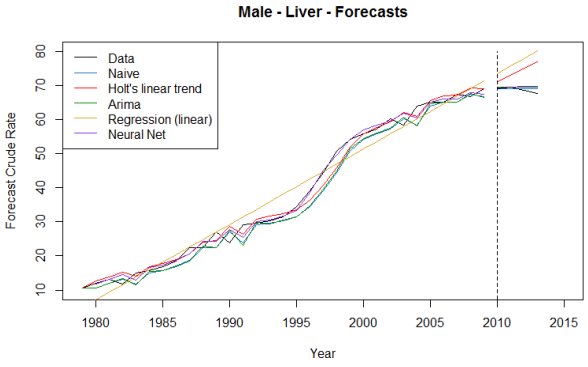
This figure shows considerable external factor on each cancer for linear regression.

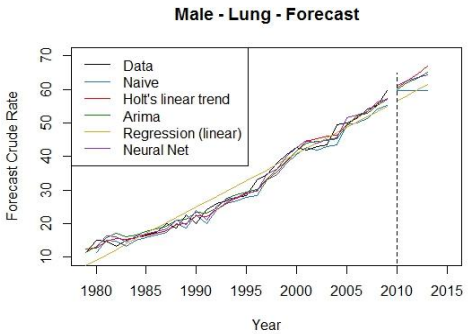
Gender	Cancer	External Factor
Male	Colorectal	Yearly working hours
	Liver	Average alcohol consumption
		Mean of Unemployment rate
		High-level education rate
		Yearly working hours
	Lung	The ratio of between richest 20%and poorest 20%
		Mean of Unemployment rate
Female	Breast	Screening rate
		Mean of Unemployment rate
	Colorectal	Working hours(yearly)
	Lung	Men's smoking rate (yearly)
		The ratio of between richest 20%and poorest 20%

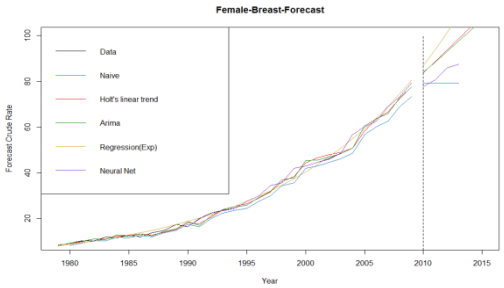
Appendix B – Forecast Models

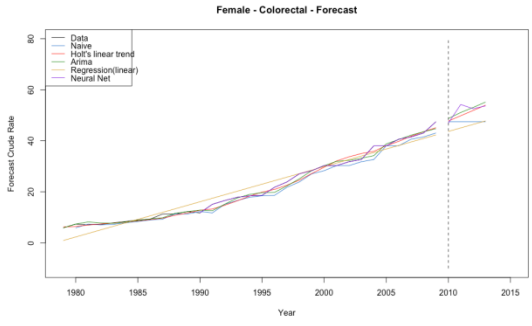
Forecast Models of TOP3 Cancer in each gender.
Each graphs show the forecast model with different methods on each cancer.

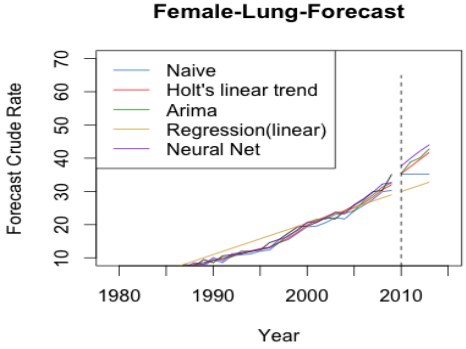






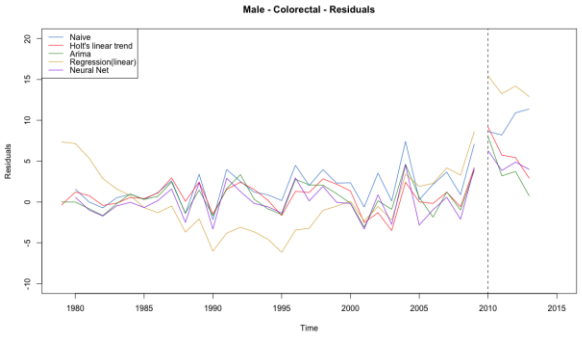


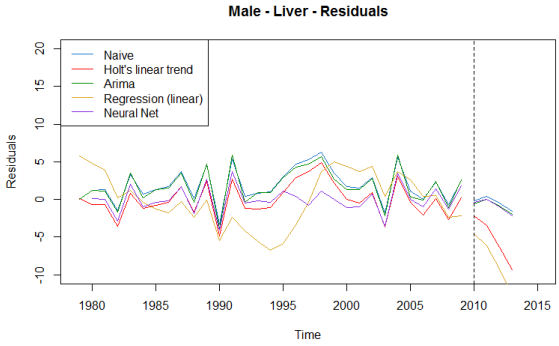


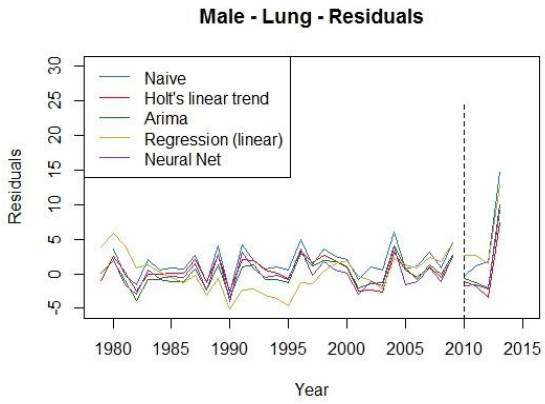


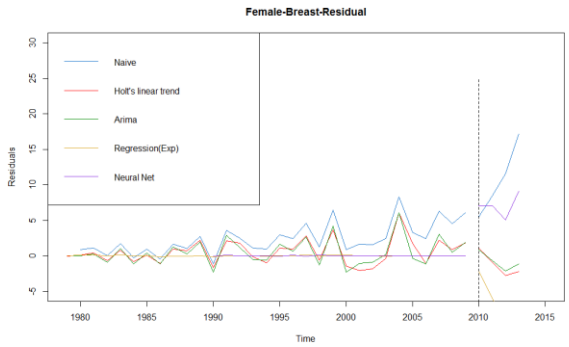
Appendix C – Residuals

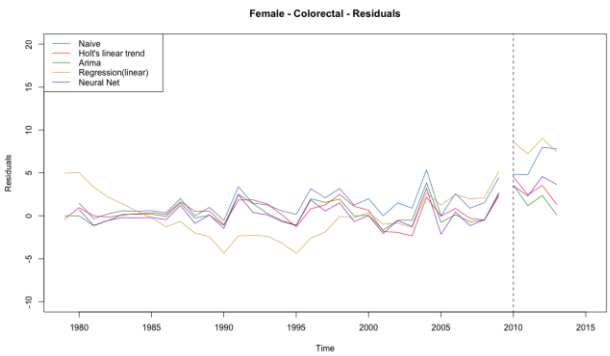
Residuals of Forecast model of TOP3 Cancer in each gender
Each graphs show the residuals of each forecast model with different methods on each cancer.

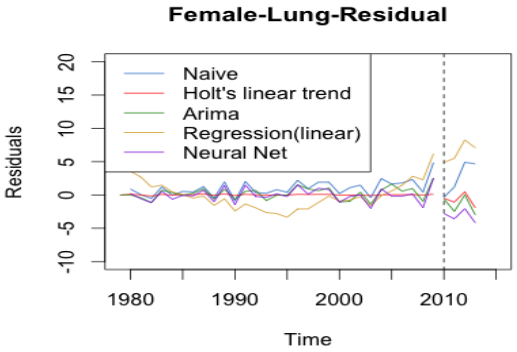








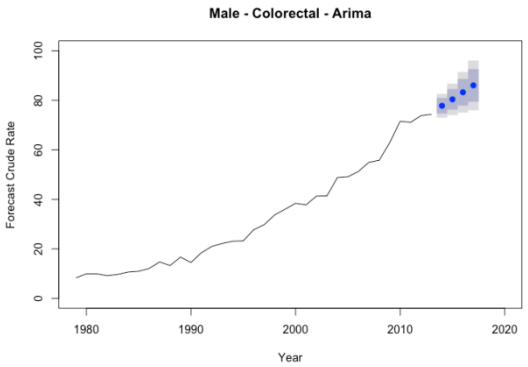


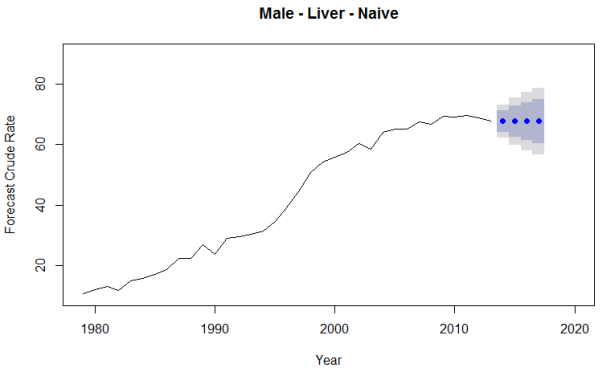


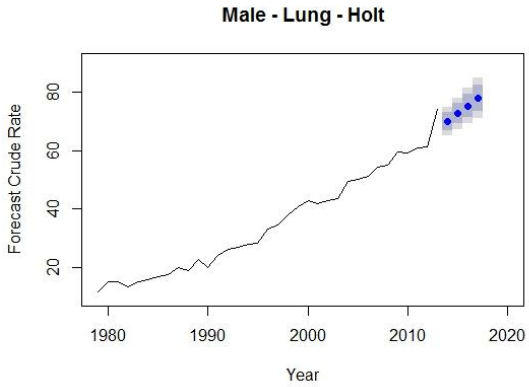
Appendix D - Prediction Interval

The graphs and table show the prediction interval of forecast from 2014 to 2017 of Best-One model on each cancer.
"Built-in" - the prediction interval comes from the R model.
"Empirical" - the prediction interval comes from Empirical Roll-Forward.

Male

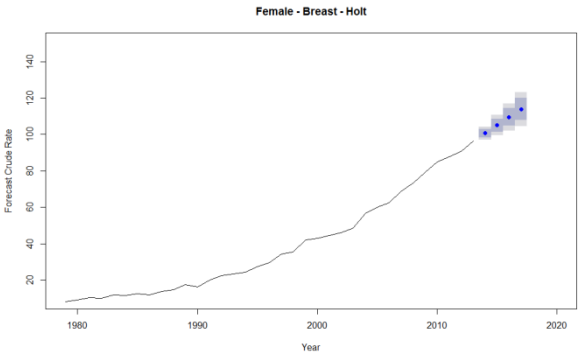


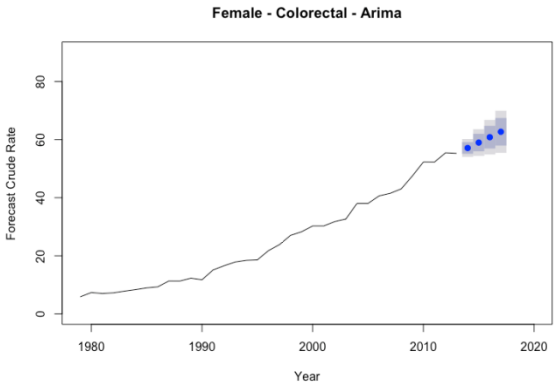


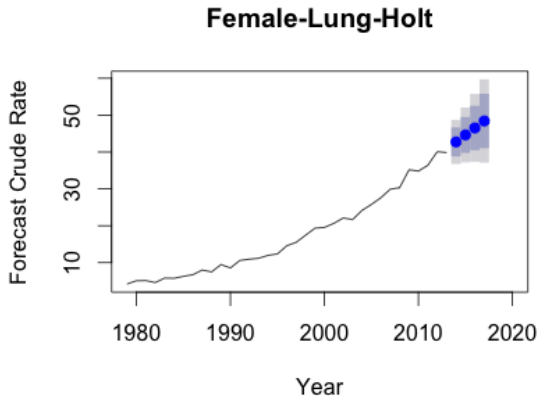


Male - Colorectal	Lower (Built-in)	Upper (Built - in)	Lower (Empirical)	Upper (Empirical)
1 Step Ahead	-13.02847	13.02847	-3.340696	4.185142
2 Step Ahead	-15.91821	15.91821	-3.226496	6.153105
3 Step Ahead	-19.92613	19.92613	-2.79681	7.566996
4 Step Ahead	-24.92542	24.92542	-2.250277	8.245619
Male - Liver	**Lower (Built-in)**	**Upper (Built - in)**	**Lower (Empirical)**	**Upper (Empirical)**
1 Step Ahead	-5.54008	5.54008	-1.775	5.549
2 Step Ahead	-7.83485	7.83485	1.0905	9.9505
3 Step Ahead	-9.5957	9.5957	1.571	14.9285
4 Step Ahead	-11.08016	11.08016	2.009	19.171
Male - Lung	**Lower (Built-in)**	**Upper (Built - in)**	**Lower (Empirical)**	**Upper (Empirical)**
1 Step Ahead	-4.7723	4.7723	-3.466787	2.885311
2 Step Ahead	-5.3165	5.3165	-5.596635	3.448938
3 Step Ahead	-6.01438	6.01439	-4.872683	3.941398
4 Step Ahead	-6.8511	6.8511	-5.585939	5.422609

Female







Female - Breast	Lower (Built-in)	Upper (Built - in)	Lower (Empirical)	Upper (Empirical)
1 Step Ahead	-11.80055	11.80054	0.8905	6.4335
2 Step Ahead	-14.13383	14.13384	2.0135	11.455
3 Step Ahead	-16.62057	16.62057	4.5325	15.895
4 Step Ahead	-19.15198	19.15198	6.457	20.2845
Female - Colorectal	**Lower (Built-in)**	**Upper (Built - in)**	**Lower (Empirical)**	**Upper (Empirical)**
1 Step Ahead	-13.02847	13.02847	-3.340696	4.185142
2 Step Ahead	-15.91821	15.91821	-3.226496	6.153105
3 Step Ahead	-19.92613	19.92613	-2.79681	7.566996
4 Step Ahead	-24.92542	24.92542	-2.250277	8.245619
Female - Lung	**Lower (Built-in)**	**Upper (Built - in)**	**Lower (Empirical)**	**Upper (Empirical)**
1 Step Ahead	-5.200233	5.200233	-1.185097	2.102325
2 Step Ahead	-6.627085	6.627085	-1.863896	2.405845
3 Step Ahead	-8.369088	8.369088	-2.400384	3.265269
4 Step Ahead	-10.39497	10.39497	-3.015806	4.447905