# A Time Series Based Method for Analyzing and Predicting Personalized Medical Data

Qinwin Vivian Hu[1], Xiangji Jimmy Huang[1],
William Melek[2], and C. Joseph Kurian[2]

[1] Information Retrieval and Knowledge Management Research Lab
York University, Toronto, Canada
[2] Alpha Global IT, Toronto, Canada
vhu@cse.yorku.ca, jhuang@yorku.ca, {william,cjk}@alpha-it.com

**Abstract.** In this paper, we propose a time series based method for analyzing and predicting personal medical data. First, we introduce an auto-regressive integrated moving average model which is good for all time series processes. Second, we describe how to identify a personalized time series model based on the patient's history information, followed by estimating the parameters in the model. Furthermore, a case study is presented to show how the proposed method works. In addition, we forecast the laboratory tests for the next twelve months in the future, with giving the corresponding prediction limits. Finally, we draw our contributions as our conclusions.

## 1 Introduction and Motivation

Like many areas in medicine, medical tests conduct on small samples collected from the human being's body and then provide the information a doctor needed to evaluate a human being's health or to understand what is causing an illness. Sometimes, doctors need to order tests to find out more. With the development of the health care theories, techniques and methods, all kinds of clinic laboratory tests are available. Then, how to make good use of these large amount of data and how to predict the laboratory tests in the future are important for the health care systems [7, 12].

In this paper, we are motivated to analyze the personalized time series process for a patient for predicting her/his laboratory tests in the future. The data are from a real research project, which will be introduced in Section 2. We have 79 monthly laboratory test records for each patient. Therefore, for each patient, we build up a time series process to predict the laboratory tests in the next $N^{th}$ month or the next $N^{th}$ year. First, we employ a general auto-regression integrated moving average (ARIMA) model [3, 8, 9]which is good for any time series process. Then, according to the history data, we identify a personalized time series model for each patient by conducting transformations, calculating the sample auto-correlation function (ACF) [6, 10]and the sample partial auto-correlation function (PACF) [6, 10]. Third, we estimate the parameters in the personalized model, based on the modified stationary model for the patient. Later, we

Y.Y. Yao et al. (Eds.): BI 2010, LNAI 6334, pp. 288–298, 2010.

present a case study to show how the proposed method works, with forecasting the laboratory tests in the future and giving the 95% prediction interval.

The remainder of this paper is organized as follows. First, we describe the data set in Section 2. Then Section 3, we introduce a personalized model identification for each patient. An ARIMA model which is the most general model fitting for every time series process is shown and the steps for how to set up a model for a patient according to his/her unique time series are presented, followed by estimating the parameters in the model. After that, we present a case study to present the experimental results, discuss and analyze the influences of our work in Section 4. Finally, we briefly draw the contributions of this paper in Section 5.

## 2 Data Set Description

The datasets in our experiment are obtained from Alpha Global IT [1]. Alpha Corporate Group is an authorization that has been providing Medical Laboratory, Industrial/Pharmaceutical Laboratory, Diagnostic Imaging services and Managed Care Medical Clinic in addition to providing commercial Electronic Medical Record and Practice Management Software. The medical test datasets contains 78 monthly patients' blood testing records. We first extract data for each patient and rank the data according to time order. Then we apply a general time series model and identify a personalized stationary model for predictions.

In order to understand the data set better, we present some sample data in Table 1. There are five attributes employed in this paper, in which SDTE stands for service date, PNUM for patient health card number, PSEX for patient gender, BDTE for patient date of birth and TSEQ for test sequence number. In particular, for the sake of privacy, the information in Table 1 is faked and it only shows the format of a class of datasets.

**Table 1.** Criteria of Theoretical ACF and PACF for Stationary Processes

| SDTE | PNUM | PSEX | BDTE | TSEQ |
|---|---|---|---|---|
| 20020101 | patient number | female | mm/dd/yyyy | $test_1$ |
| ... | ... | ... | ... | ... |
| 20030201 | patient number | female | mm/dd/yyyy | $test_9$ |
| ... | ... | ... | ... | ... |
| 20040101 | patient number | female | mm/dd/yyyy | $test_5$ |
| ... | ... | ... | ... | ... |
| ... ... | | | | |
| 20080601 | patient number | female | mm/dd/yyyy | $test_1$ |
| ... | ... | ... | ... | ... |

## 3 Personalized Model Identification

In this section, we introduce a personalized model identification for each patient. First, we introduce an ARIMA model which is the most general model fitting for every time series process. Then, we describe the steps for how to set up a model for a patient according to his/her unique time series.

### 3.1 The General ARIMA Model

In statistics, and in particular in time series analysis, not all time series are always stationary. A homogeneous non-stationary time series can be reduced to a stationary time series by taking a proper degree of differencing. The auto-regressive model, the moving average model and the auto-regressive moving average model, are useful in describing stationary time series. Then the auto-regressive integrated moving average (ARIMA) model is built as a large class of time series model using differencing, which is useful in describing various homogeneous non-stationary time series.

The general ARIMA model can be presented in Equation 1. 1.

$$ARIMA(p,d,q) : \phi_p(B)(1-B)^d Z_t = \theta_0 + \theta_q(B)a_t \tag{1}$$

where $B$ is a back shift operator [3]; $\phi_p(B)$ is the stationary AR operator [8, 9] with $\phi_p(B) = (1-\phi_1 B - ... - \phi_p B^p)$; $\theta_q(B)$ is the invertible MA operator [11, 13] with $\theta_q(B) = (1-\theta_1 B - ... - \theta_q B^q)$; $\phi_p(B)$ and $\theta_q(B)$ share no common factors; $\theta_0$ is a parameter related to the mean of the process; and $a_t$ is a white noise process [3, 9].

The parameter $\theta_0$ plays very different roles for $d = 0$ and $d > 0$. When $d = 0$, the original process is stationary, and we get $\theta_0 = \mu(1 - \phi_1 - ... - \phi_p)$. When $d > 0$, however, $\theta_0$ is called the deterministic trend term and is often omitted from the model unless it is really needed.

### 3.2 Steps for Model Identification

To illustrate the model identification, we consider the general$ARIMA(p,d,q)$ model introduced in Section 3.1.

Model identification refers to the methodology in identifying the required transformations, the decision to include the deterministic parameter $\theta_0$, and the proper order of $p$ the AR operator and $q$ for the MA operator. Given a time series of a patient, we use the following steps to identify a tentative model for predicting the lab tests in the future.

*Step 1.* Plot the time series data and choose proper transformations.

In a time series analysis, plotting the time series data is always the first step. Through careful examination of the plot, we usually get a good idea about whether the series contains a trend, seasonality, outliers, non-constant means, non-constant variances, and other abnormal and non-stationary phenomena. This understanding often provides a basis for postulating a possible data transformation. Since we prefer examining the plot automatically, there are more than one way to understand the series, such as simulating a distribution for the data.

Differencing and variance-stabilizing transformations are two commonly used transformations, in time series analysis. Because variance-stabilizing transformations such as the power transformation require non-negative values and differencing may create some negative values, we should always apply variance-stabilizing transformations before taking differencing. A series with non-constant variance often needs a logarithmic transformation. More generally, we refer to the transformed data as the original series in the following discussion unless mentioned otherwise.

*Step 2.* Compute and examine the sample ACF and the sample PACF of the original series to further confirm a necessary degree of differencing so that differenced series is stationary. We employ two rules as follows.

First, if the sample ACF decays very slowly (the individual ACF may not be large) and the sample PACF cuts off after lag 1, then it indicates differencing is needed. In general, we try taking the first differencing $(1 - B)Z_t$. One can also use the unit root test proposed by Dickey and Fuller (1979) [4]. In a borderline case, differencing is recommended by Dickey, Bell and Miller (1986) [5].

Second, in order to remove non-stationary, we may need to consider a higher order differencing $(1 - B)^d Z_t$ for $d > 1$. In most cases, $d$ is ether 0, 1, or 2. Note that if $(1 - B)^d Z_t$ is stationary, then $(1 - B)^{d+i} Z_t$ for $i = 1, 2, ...$ are also stationary.

*Step 3.* Compute and examine the sample ACF and the sample PACF of the properly transformed and differenced series to identify the orders of $p$ and $q$ where we recall that $p$ is the highest order in the AR polynomial $(1 - \phi_1 B - ... - \phi_p B^p)$, and $q$ is the highest order in the MA polynomial $(1 - \theta_1 B - ... - \theta_q B^q)$. Usually, we obtain the needed orders of $p$ and $q$ less than or equal to 3. We also present a table in Table 2 to summarize the important criteria for selecting $p$ and $q$.

**Table 2.** Criteria of Theoretical ACF and PACF for Stationary Processes

| Process | ACF | PACF |
|---|---|---|
| $AR(p)$ | Tails off as exponential decay or damped sine wave | Cuts off after lag $p$ |
| $MA(q)$ | Cuts off after lag $q$ | Tails off as exponential decay or damped sine wave |
| $ARMA(p, q)$ | Tails off after lag $(q - p)$ | Tails off after lag $(p - q)$ |

*Step 4.* Test the deterministic trend term $\theta_0$ when $d > 0$.

For a non-stationary model, $\phi B(1 - B)^d Z_t = \theta_0 + \theta(B) a_t$, the parameter $\theta_0$ is usually omitted so that it is capable of representing series with random changes in the level, slope or trend. If there is reason to believe that the differenced series contains a deterministic trend mean, however, we can test for its inclusion by comparing the sample mean $\bar{W}$ of the differenced series $W_t = (1 - B)^d Z_t$ with its approximate standard error $S_{\bar{W}}$. To derive $S_{\bar{w}}$, we note that

$$lim_{n \to \infty} nVar(\bar{W}) = \sum_{j=-\infty}^{\infty} \gamma_j \tag{2}$$

Hence, we get

$$\sigma^2_{\bar{W}} = \frac{\gamma_0}{n} \sum j = -\infty^{\infty} \rho_j = \frac{1}{n} \sum j = -\infty^{\infty} \gamma_j = \frac{1}{n} \gamma(1) \tag{3}$$

where $\gamma(B)$ is the auto-covariance generating function and $\gamma(1)$ is its value at $B = 1$. Thus, the variance and the standard error for $\bar{W}$ is model dependent. For example, for the $ARIMA(1, d, 0)$ model, $(1 - \phi B)W_t = a_t$, we have:

$$\gamma(B) = \frac{\sigma_a^2}{(1 - \phi B)(1 - \phi B^{-1})} \tag{4}$$

so that

$$\begin{aligned} \sigma_{\bar{W}}^2 &= \frac{\sigma_a^2}{n}\frac{1}{(1-\phi)^2} = \frac{\sigma_W^2}{n}\frac{1-\phi^2}{(1-\phi)^2} \\ &= \frac{\sigma_W^2}{n}\frac{1+\phi}{1-\phi} = \frac{\sigma_W^2}{n}\frac{1+\rho_1}{1-\rho_1} \end{aligned} \tag{5}$$

where we note that $\sigma_W^2 = \sigma_a^2/(1 - \phi^2)$. The required standard error is

$$S_{\bar{W}} = \sqrt{\frac{\hat{\gamma}_0}{n}(\frac{1+\hat{\rho}_1}{1-\hat{\rho}_1})} \tag{6}$$

Expressions of $S_{\bar{W}}$ for other models can be derived similarly. At the model identification phase, however, because the underlying model is unknown, most available software use the approximation

$$S_{\bar{W}} = [\frac{\hat{\gamma}_0}{n}(1 + 2\hat{\rho}_1 + 2\hat{\rho}_2 + ... + 2\hat{\rho}_k)]^{1/2} \tag{7}$$

where $\hat{\gamma}_0$ is the sample variance and $\hat{\rho}_1, ..., \hat{\rho}_k$ are the first $k$ significant sample ACFs of $\{W_t\}$. Under the null hypothesis $\rho_k = 0$ for $k \geq 1$, we can reduce Equation 6 to

$$S_{\bar{W}} = \sqrt{\hat{\gamma}_0/n} \tag{8}$$

Alternatively, one can include $\theta_0$ initially and discard it at the final model estimation if the preliminary estimation results is not significant.

### 3.3 Parameter Estimation

After we identify a personalized model in Section 3.2, we have to estimate the parameters in the model. In this section, we apply the method of moments for parameter estimation.

The method of moments consists of substituting sample moments such as the sample mean $\bar{Z}$, sample variance $\hat{\gamma}_0$ and the sample ACF $\hat{\rho}_i$ for their theoretical counterparts and solving the resultant equations to obtain estimates of unknown parameters. For better understanding, we take an auto-regressive process $AR(p)$ as an example. In a similar way, we can deal with a moving average process $MA(q)$ and an auto-regression moving average process $ARMA(p, q)$ at the same way.

In an $AR(p)$ process, we have

$$\dot{Z}_t = \phi_1 \dot{Z_{t-1}} + \phi_2 \dot{Z_{t-2}} + ... + \phi_p \dot{Z_{t-p}} + a_t \tag{9}$$

the mean $u = E(Z_t)$ is estimated by $\bar{Z}$. To estimate $\phi$, we first use that $\rho_k = \phi_1\rho_{k-1} + \phi_2\rho_{k-2} + ... + \phi_p\rho_{k-p}$ for $k \geq 1$ to obtain the following system of Yule-Walker [13] equations:

$$\begin{aligned}
\rho_1 &= \phi_1 + \phi_2\rho_1 + \phi_3\rho_2 + ... + \phi_p\rho_{p-1} \\
\rho_2 &= \phi_1\rho_1 + \phi_2 + \phi_3\rho_1 + ... + \phi_p\rho_{p-2} \\
&... \\
\rho_p &= \phi_1\rho_{p-1} + \phi_2\rho_{p-2} + \phi_3\rho_{p-3} + ... + \phi_p
\end{aligned} \tag{10}$$

Then, replacing $\rho_k$ by $\hat{\rho_k}$, we obtain the moment estimators $\hat{\phi_1}, \hat{\phi_2}, ..., \hat{\phi_p}$ by solving the above linear system of equations. That is,

$$\begin{pmatrix} \hat{\phi_1} \\ \hat{\phi_2} \\ .. \\ \hat{\phi_p} \end{pmatrix} = \begin{pmatrix} 1 & \hat{\rho_1} & \hat{\rho_2} & .. & \hat{\rho_{p-2}} & \hat{\rho_{p-1}} \\ \hat{\rho_1} & 1 & \hat{\rho_1} & .. & \hat{\rho_{p-3}} & \hat{\rho_{p-2}} \\ .. & & & & & .. \\ \hat{\rho_{p-1}} & \hat{\rho_{p-2}} & \hat{\rho_{p-3}} & .. & \hat{\rho_1} & 1 \end{pmatrix}^{-1} \begin{pmatrix} \hat{\rho_1} \\ \hat{\rho_2} \\ .. \\ \hat{\rho_p} \end{pmatrix} \tag{11}$$

These estimators are usually called Yule-Walker estimators [13].

Having obtained $\hat{\phi_1}, \hat{\phi_2}, ..., \hat{\phi_p}$, we use the result

$$\begin{aligned}
\gamma_0 &= E(\dot{Z}_t\dot{Z}_t) = E[\dot{Z}_t(\phi_1\dot{Z_{t-1}} + \phi_2\dot{Z_{t-2}} + .. + \phi_p\dot{Z_{t-p}} + a_t)] \\
&= \phi_1\gamma_1 + \phi_2\gamma_2 + .. + \phi_p\gamma_p + \sigma_a^2
\end{aligned} \tag{12}$$

and obtain the moment estimator for $\sigma_a^2$ as

$$\sigma_a^2 = \hat{\gamma}_0(1 - \hat{\phi_1}\hat{\rho}_1 - \hat{\phi_2}\hat{\rho}_2 - \hat{\phi_p}\hat{\rho}_p) \tag{13}$$

## 4 Case Study

In this section, we present an example to show our proposed personalized time series model.

### 4.1 Model Identification

The number of laboratory tests can be attractive for many health care systems. Figure 1 shows a time series for a female patient who had done her blooding tests from January, 2002 to July, 2008. In total, there are 79 records as the observations of consecutive months. From this figure, it indicates that the series is not stationary in the mean and variance. We compute the sample ACF and sample PACF of the time series $Z$ are shown in Figure 2 and 3 for choosing transformations or differencing. In Figure 2 and 3, we can see that the sample ACF doesn't decays very slowly, and the sample PACF does not cut off after lag 1. Therefore, we do not need to consider a degree of differencing to make the time series stationary. At the same time, to investigate the required transformation for variance stabilization, we apply the power transformation analysis [2] to suggest

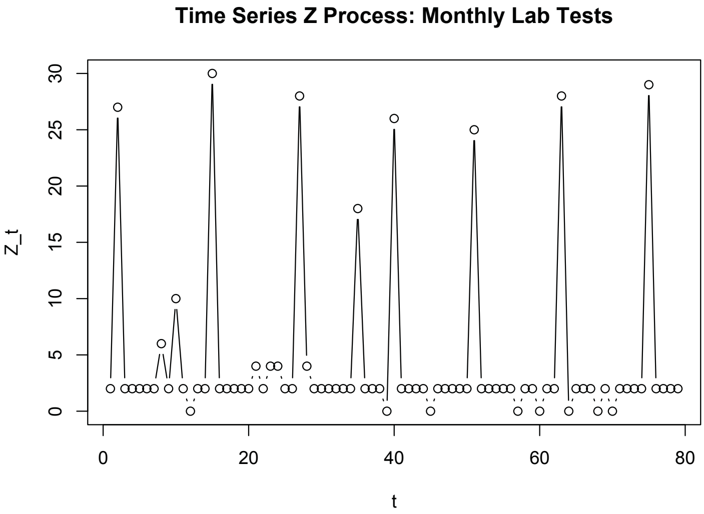

**Fig. 1.** 79 Monthly Blood Testing Records: Z

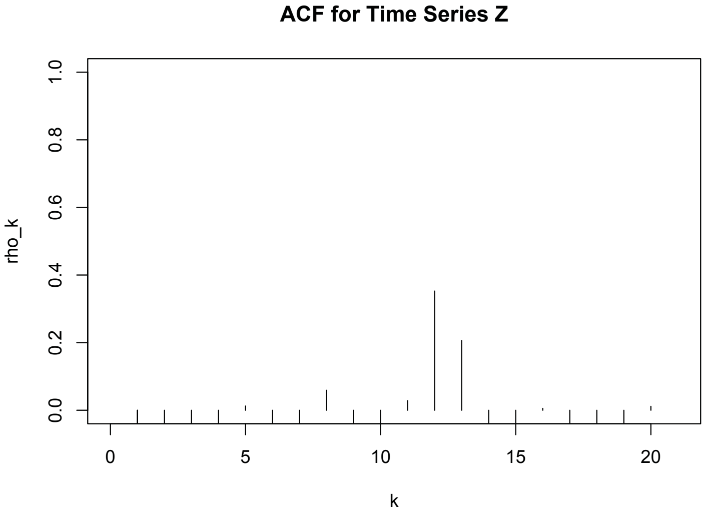

**Fig. 2.** Sample ACF for Time Series Z

an optimal parameter as $\lambda = 0.25$. The power transformation is presented in Equation 14

$$W_t = T(Z_t) = \frac{Z_t^{\lambda} - 1}{\lambda} \tag{14}$$

The transformed time series process $W$ is plotted in Figure 4, in which we can see that $W$ is stationary in the mean but may not be stationary in the

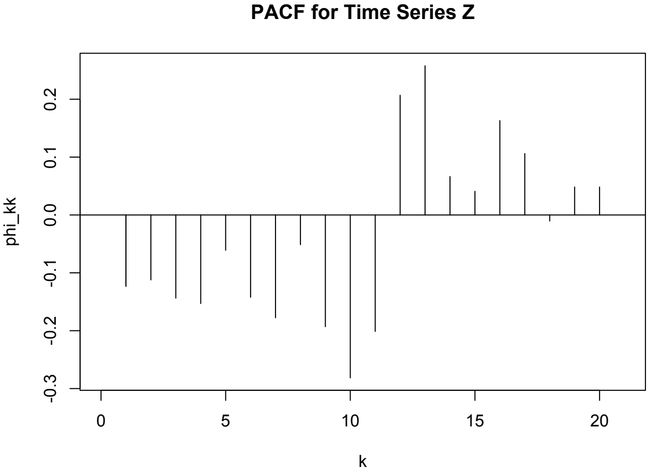

**Fig. 3.** Sample PACF for Time Series Z

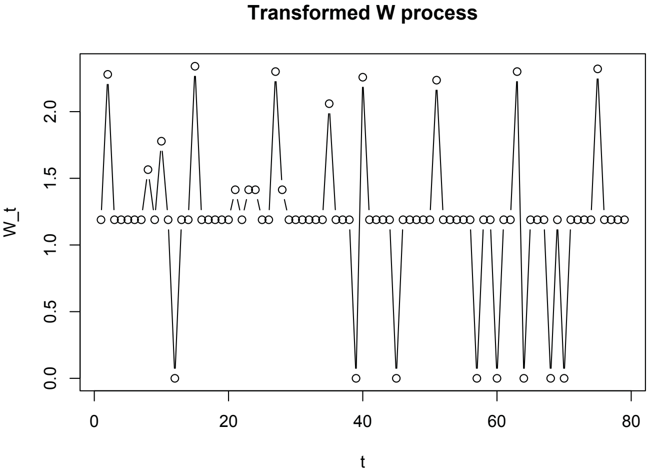

**Fig. 4.** Transformed Time Series: W

variance. Hence, we further compute the sample ACF and sample PACF for the transformed series $W$, which are shown in Figure 5 and 6.

The sample ACF shows a dample sin-consine wave and the sample PACF has relatively large spike at lag 1, 8 and 13, suggesting that a tentative model may be an $AR(1)$ model in Equation 16.

$$(1 - \phi B)(W_t - \mu) = a_t \tag{15}$$

where $W_t = T(Z_t) = \frac{Z_t^\lambda - 1}{\lambda}$ in Equation 14 with $\lambda = 0.25$.

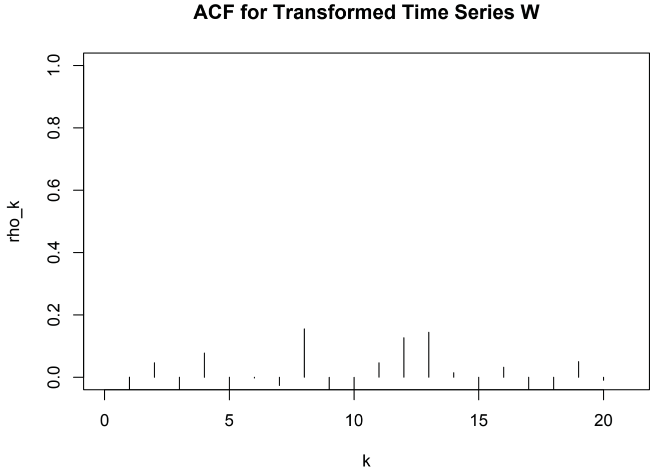

**Fig. 5.** Sample ACF for Transformed Time Series W

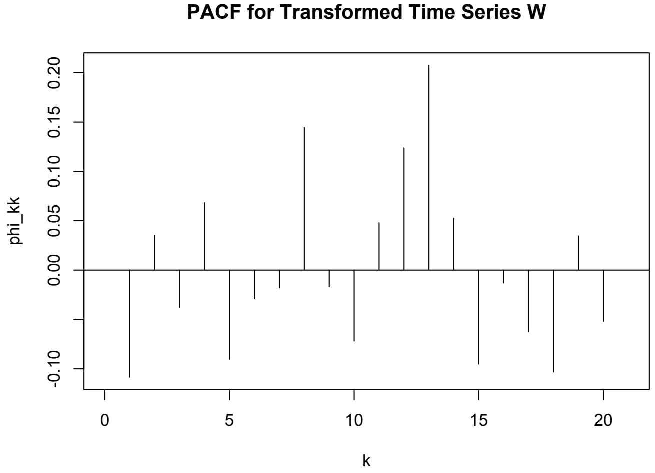

**Fig. 6.** Sample PACF for Transformed Time Series W

### 4.2 Forecasting

We have identified an $AR(1)$ model in Section 4.1 for the transformed series. In this section, we also use this transformed series to forecast the next $N^{th}$ month laboratory tests in the future [3].

For this $AR(1)$ model, we have

$$(1 - \phi B)(W_t - \mu) = a_t \tag{16}$$

where $\phi = -0.1$, $\mu = 1.2$ and $\sigma_a^2 = 0.1$. In this case, we have 79 observations and want to forecast the next year of twelve months with their associated 95% forecast limits.

First of all, we write the $AR(1)$ model as

$$W_t - \mu = \phi(W_{t-1} - \mu) + a_t \tag{17}$$

and the general form of the forecast equation is

$$\begin{aligned} \hat{W_t}(l) &= \mu + \phi(\hat{W_{t-1}}(l-1) - \mu) \\ &= \mu + \phi^l(W_t - \mu) \end{aligned} \tag{18}$$

Thus, the following twelve months' predictions are computed in Equation 19 and the results are shown in Table 3.

$$\begin{aligned} \hat{W_{79}}(1) &= 1.2 - 0.1 * (\frac{2^{0.25} - 1}{0.25}) \\ \hat{W_{79}}(2) &= 1.2 + (-0.1)^2 * (\frac{2^{0.25} - 1}{0.25}) \\ &... \\ \hat{W_{79}}(12) &= 1.2 + (-0.1)^{12} * (\frac{2^{0.25} - 1}{0.25}) \end{aligned} \tag{19}$$

**Table 3.** Forecasting for the Next Twelve Months

| Predicted Month | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Numbers | 2.4 | 2.0 | 2.1 | 2.1 | 2.1 | 2.1 | 2.1 | 2.1 | 2.1 | 2.1 | 2.1 | 2.1 |

**Table 4.** 95% Forecasting Limits for the Next Twelve Months

| Predicted Month | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Intervals | [0.2, 12.1] | [0.1, 10.9] | [0.1, 11.0] | [0.1, 11.0] | [0.1, 11.0] | [0.1, 11.0] |
| Predicted Month | 7 | 8 | 9 | 10 | 11 | 12 |
| Intervals | [0.1, 11.0] | [0.1, 11.0] | [0.1, 11.0] | [0.1, 11.0] | [0.1, 11.0] | [0.1, 11.0] |

Second, in order to obtain the forecast limits, we calculate the weights called $\psi$ from the relationship

$$(1 - \phi B)(1 + \psi_1 B + \psi_2 B^2 + ...) = 1 \tag{20}$$

That is,

$$\psi_j = \phi^j, \forall j \geq 0 \tag{21}$$

Therefore, the 95% forecast limits for the forecasting results in Table 3 are computed in Table 4.

$$\hat{W_{79}}(1) \pm 1.96 * \sqrt{\sigma_a^2} \tag{22}$$

## 5 Conclusions

In this paper, we propose a time series method on how to identify a personalized model based on the patient's laboratory test records. After successfully building up the personalized model, we predict the laboratory tests in the future. In addition, we also give a predictive limits for the forecasting, which is useful for many health care systems. The case study shows that the proposed method provides a good way for personalization analysis.

In the future, we will continue on working for personalization tools, such as making this time series method be a GUI tool. Furthermore, we plan to work on group information for predictions.

## Acknowledgements

This research is supported part by NSERC CRD project. We would also like to thank three anonymous reviewers for their useful comments on this paper.

## References

[1] Alpha Global IT, http://www.alpha-it.com/

[2] Box, G.E.P., Cox, D.R.: An Analysis of Transformations. Journal of the Royal Statistical Society, Series B 26(2), 211–252 (1964)

[3] Box, G.E.P., Jenkins, G.M.: TIme Series Analysis Forecasting and Control, 2nd edn. Holden-Day, San Franscisco (1976)

[4] Dickey, D.A., Fuller, W.A.: Distribution of the Estimators for Autoregressive Time Series With a Unit Root. J. Amer. Statist. Assoc. 74, 427–431 (1979)

[5] Dickey, D.A., Bell, B., Miller, R.: Unit Roots in Time Series Models: Tests and Implications. The American Statistician 40(1), 12–26 (1986)

[6] Dunn, P.F.: Measurement and Data Analysis for Engineering and Science. McGraw–Hill, New York (2005) ISBN 0-07-282538-3

[7] Garg, A., Adhikari, N., McDonald, H., Rosas-Arellano, M., Devereaux, P., Beyene, J., Sam, J., Haynes, R.: Effects of Computerized Clinical Decision Support Systems on Practitioner Performance and Patient Outcomes: A Systematic Review. Jama 293(10), 1223 (2005)

[8] Mills, T.C.: Time Series Techniques for Economists. Cambridge University Press, Cambridge (1990)

[9] Pandit, S.M., Wu, S.-M.: Time Series and System Analysis with Applications. John Wiley & Sons, Inc., Chichester (1983)

[10] Percival, D.B., Walden, A.T.: Spectral Analysis for Physical Applications: Multitaper and Conventional Univariate Techniques, pp. 190–195. Cambridge University Press, Cambridge (1993) ISBN 0-521-43541-2

[11] Slutzky, E.: The Summation of Random Causes as the Source of Cyclic Processes. Econometrica 5, 105–146 (1937); Translated from the earlier paper of the same title in Problems of Economic Conditions

[12] Stead, W.W., Garrett Jr., L.E., Hammond, W.E.: Practicing nephrology with a computerized medical record. Kidney Int. 24(4), 446–454 (1983)

[13] Yule, G.U.: On a method of Investigating Periodicities in Disturbed Series with Special Reference to Wolfer's Sunspot Numbers. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 226, 267–298 (1927)