AGRICULTURAL JOURNAL

MEDWELL PUBLICATIONS

Forecasting Chronic Kidney Disease Stages and Urgency level of Dialysis Using Time Series Algorithm

August Anthony N. Balute, Dennis B. Gonzales, Jennifer T. Carpio and Albert A. Vinluan
Department of Graduate School, University of the East, Manila, 1008 Metro Manila, Philippines

Key words: Acute Renal Failure (ARF), ARIMA model, Time Series algorithm, Chronic Kidney Disease (CKD), Electronic Medical Record (EMR), R programming, scrum methodology

Corresponding Author:
August Anthony N. Balute
Department of Graduate School, University of the East, Manila, 1008 Metro Manila, Philippines

Page No.: 143-148
Volume: 15, Issue 6, 2020
ISSN: 1816-9155
Agricultural Journal

Abstract: As medical data may contain diagnoses and treatments that are subject to error rates, imprecision and uncertainty, medical data mining methods and tools require medical research using data mining methods and artificial intelligence techniques to systematically come up with a suited analysis of the medical database. A Time Series algorithm specifically Auto Regressive Integrated Moving Average (ARIMA) model can be used to detect and analyze frequency and probability of data by assessing essential attributes to predict and forecast trends which in this study is predicting the urgency of dialysis and Chronic Kidney Disease (CKD) stage, to determine the urgency level and prioritization of selected kidney patients. As a valuable tool for predicting future health events demanding services and healthcare needs, preventive measures and intervention strategies will be recommended by doctors easily for decision-making support using a time series algorithm.

INTRODUCTION

The annual mortality rate per 100,000 people from chronic kidney disease in the Philippines has increased by 16.2% since, 1990 which is an average of 0.7% per year. With the increase in number of patients needing dialysis and the resource limited setting of our health care system, it is often encountered that there are a lot of patients needing dialysis and there are only a few dialysis machines available. Since, patients lined up for emergency dialysis are usually seen by different physicians, determining who to prioritize first once a slot is available becomes a challenge. The managers/supervisors at hemodialysis units are also unaware of the overall clinical status of the patient and only rely on the note of the attending physician.With the multitude of factors that needs to be considered in clinical decision making regarding dialysis, the use of machine learning algorithms applied to patient data sets may help clinicians and dialysis units become more efficient and increase cost effectiveness of treatments and quality of care.

Statement of the problem: This study aims to create a forecasting model to determine the CKD stages and urgency level of dialysis using Time Series algorithm.

- What are the clinical parameters that can help in forecasting the CKD stages and urgency level of dialysis?
- How the Time Series algorithm can help forecast the CKD stages and urgency level of dialysis?
- How the Time Series algorithm can help provide the quality of life of a dialysis patient?

Literature review: The challenge in the medical industry nowadays is how to exploit the huge amount of data the field generates. Approaches are required to discover knowledge for decision making. Time series are data types common in the medical domain and require specialized analysis techniques and tools. Auto-Regressive Integrated Moving Average (ARIMA) is one popular linear model in time series forecasting[1]. Time series data in business, economics, environment, medicine and other scientific fields tend to exhibit patters such as trends, seasonal fluctuations, irregular cycles and occasional shifts in level or variability[2]. The analysis of time series for knowledge discovery requires application of special-purpose tools such as the key information of interest to the expert concentrated within a particular time series region, known as events. The concept of data mining to analyze volumes of stored medical data to discover knowledge has a huge potential[3].

ARIMA model was utilized together with Kalman Filter algorithm to forecast the incidence of gonorrhea. The results show that Kalman Filter algorithm based on ARIMA model could perform the prediction of the disease more precisely and accurately, comparing the range of Absolute Error (AE), the Mean Absolute Error (MAE), and the Mean Absolute Percentage Error (MAPE). Furthermore, ARIMA also accurately predicts the time of death, institutionalization and need for full-time care of Alzheimer's disease patients serving as a clinical, research, and public health need[4].

As data mining delivers the technology and procedure to convert ordinary data into evidences of planning and decision making, the huge volume of data in the healthcare industry have been beneficial to generate associations among attributes and extract valuable information. Milicevic conducted a study in Serbia on its physician and nurse supply using time series data based on historic trends from 1961-2008. The study identified variables that were significantly related to physician and nurse employment rates in the public healthcare sector namely, annual total national population, gross domestic product adjusted to 1994 prices, inpatient care discharges, outpatient care visits, students enrolled in the first year of medical studies at public universities and the annual number of graduated physicians. Based on the historic trends in 2015 modeled using ARIMA, the model yielded a stable and significant forecast of physician supply and nurse supply, identifying population and GDP as the most significant predictors. The model also predicts a 7 years mismatch between the supply of graduates and vacancies in the public healthcare sector, concluding the need for a coherent healthcare objective in professional mobility[5].

On the other hand, classification of multivariate time series data is considered challenging but necessary for medical care and research. According to Moskovitch and Shahar in their study on the classification-driven temporal discretization of multivariate time series, a series of raw-data time points can be abstracted into a set of time intervals which can be used for the classification of multivariate time series[6]. ARIMA models are useful to assess trends over time, like evaluating a population based on trends in the use of medical treatments (extracorporeal shock wave lithotripsy, ureteroscopy and percutaneous nephrolithotomy) as well as to assess the re-treatment rate and morbidity from treatment over time.

However, Fisher and Mehta[7] emphasized that algorithms inferring particularly to ecological interactions must overcome three major obstacles: a correlation between the abundances of two species does not imply that those species are interacting, the sum constraint on the relative abundances obtained from metagenomic studies makes it difficult to infer the parameters in time series models, and errors due to experimental uncertainty, or mis-assignment of sequencing reads into operational taxonomic units, bias inferences of species interactions due to a statistical problem called “errors-in-variables”. In the study, an approach in time series was proposed learning interactions from microbial time series, using linear regression with bootstrap aggregation for microbial dynamics. Results show that it could reliably infer the topology of the ecological interactions using time series[7].

The use of Convolutional Neural Networks (CNN) for time series classification has been widely used although a lot of training data needs to be efficient. Data augmentation techniques and learning the network in a semi-supervised way using training time series from different datasets can be used for benchmarking[8]. With this, Artificial Neural Networks (ANNs) can also be suggested to be a promising alternative to traditional linear models in forecasting performance. A combination of ARIMA and ANN models can be an effective way to improve forecasting accuracy achieved by either of the models used separately[9].

With the use of ARIMA, anomalies can be detected by analyzing instantaneous frequency and amplitude of probability. The algorithm learns the system’s normal behavior and does not require the existence of anomalous data for assessing its statistical significance which is an essential attribute in applications that require customization[10].

MATERIALS AND METHODS

The Scrum approach has been developed for managing the software development process. It is an empirical approach applying the ideas of process control theory to software methodology resulting in an approach that reintroduces the ideas of flexibility, adaptability and productivity. The main idea of Scrum is that systems

development involves several environmental variables (e.g., requirements, time frame, resources and technology) that is likely to change during the process. This makes the development process to require flexibility of the systems development process for it to be able to respond to the changes. Most medical systems that have been developed utilize the Scrum software methodology[11]. Scrum process includes three phases: pre-game, development and post-game.

The pre-game phase includes two sub-phases: Planning and Architecture/High Level Design (Fig. 1). Planning includes the definition of the system being developed. A product backlog list will be created containing all the requirements that are currently known. The requirements are prioritized and the effort needed for their implementation is estimated. Planning also includes the definition of the project, tools and other resources. In every iteration the updated product backlog is reviewed to gain commitment for the next iteration. In the architecture phase, the high-level design of the software including the architecture will be planned based on the current items in the Product backlog. In case of an enhancement to existing software, the changes needed for the implementing the backlog items are identified along with the problems that it may cause. A design review for the implementation and decisions are made on the basis of this review. In addition, preliminary plans for the contents of releases are prepared. The development phase (also known as the game phase) is the agile part of the Scrum approach. This phase is treated as a "black box" where the unpredictable is expected. The different environmental and technical variables (such as time frame, quality, requirements, resources, implementation technologies and tools and even development methods) identified in Scrum. In the development phase the software is developed in Sprints. Sprints are iterative cycles where the functionality is developed or enhanced to produce new increments. Each Sprint includes the traditional phases of software development: requirements, analysis, design, evolution and delivery phases. The architecture and the design of the system evolve during the Sprint development.

The post-game phase contains the closure of the release. This phase is entered when an agreement has been made that the environmental variables such as the requirements are completed. In this case, no more items and issues can be found nor can any new ones be invented. The system is now ready for the release and the preparation for this is done during the post-game phase, including the tasks such as the integration, system testing and documentation. In this sutdy, the system architecture of the Degree of Urgency and Progression Predictive Model using Hybrid algorithm is demonstrated and each components of the system were described. The infrastructure design of the system is composed of five components as can be seen in Fig. 2. Figure 3 shows the

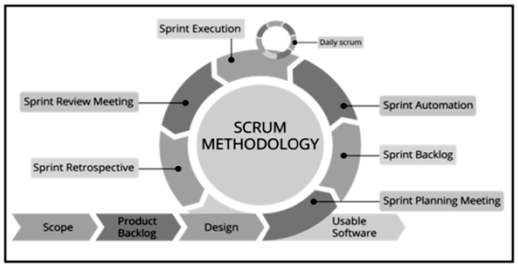

Fig. 1: The SCRUM methodology

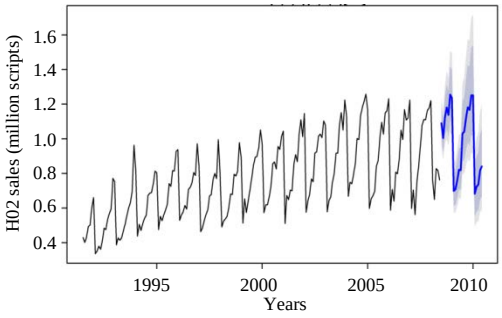

Fig 2: The ARIMA model

main components and the sub-components and a clear explanation of their features and interaction amongst them.

Pre-processing using ARIMA model: Time series forecasting is an important area in data mining research. Feature preprocessing techniques have significant influence on forecasting accuracy and are essential in a forecasting model. Effective feature preprocessing can significantly enhance forecasting accuracy[12].

Auto Regression (AR): AR is a class of linear model where the variable of interest is regressed on its own lagged values. The ARIMA Model Equation is:

$$y_t = \delta + \left\{ \begin{matrix} \phi 1 yt-1 + \phi 2 yt- \\ 2+,...,+\phi 2 yt-p \end{matrix} \right\} + \{\phi 1 \in t-q\} + \in t$$

$$= yt = \delta + \sum_{i=1}^{p} \phi i yt-i + \sum_{j-1}^{q} \phi j \in t-j + \in t$$

Figure 4 illustrates the ARIMA Pre-processing and ARIMA Feature Selection by selecting year, time and month as significant attributes for forecasting. Auto-Regressive (AR) model predicts future behavior based on past behavior. It is used for forecasting when there is some correlation between values in a time series and the values that precede and succeed them.

Figure 5 shows that statistical analysis model (ARIMA) that uses time series data to predict future

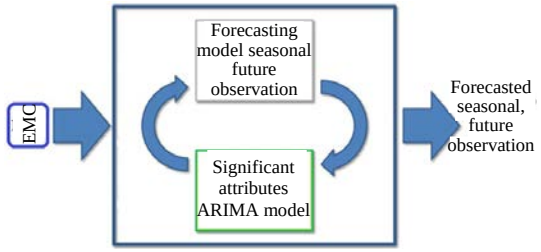

Fig. 3: Simplified conceptual framework

	Point Forecast	Lo 95	Hi 95
Jan 2018	800.4155	192.36908	1408.462
Feb 2018	808.7643	145.02496	1472.504
Mar 2018	966.1473	251.03926	1681.255
Apr 2018	922.0023	158.97607	1685.029
May 2018	868.9989	60.89089	1677.107
Jun 2018	907.2524	56.44796	1758.057
Jul 2018	852.5477	-38.91045	1744.006
Aug 2018	1174.3849	244.04774	2104.722
Sep 2018	1090.7104	123.05508	2058.366
Oct 2018	967.2466	-36.34014	1970.833
Nov 2018	909.2651	-129.01036	1947.540
Dec 2018	868.3443	-203.49779	1940.186

Fig. 4: Data pre-processing using ARIMA Model

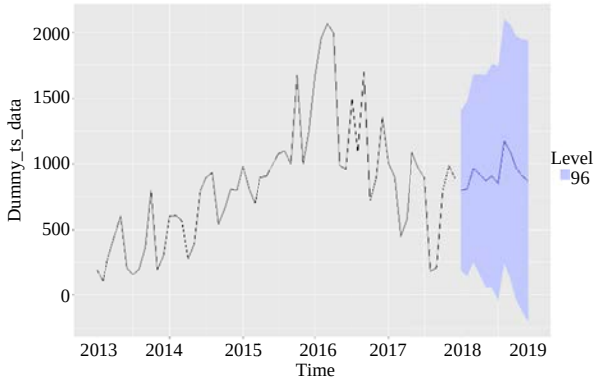

Fig. 5: Data pre-processing using ARIMA Model

trends. It is a form of regression analysis that seeks to predict future movements by examining the differences between values in the series instead of using the actual data values. Lags of the differenced series are referred to as "autoregressive" and lags within forecasted data are referred to as "moving average".

Survey forms were prepared for the acceptability test and were also included as added features of the dashboard. The software engineering quality of the study is ISO/IEC 25010 systems and software quality requirements and evaluation compliant. Questions were grouped into 8 structured set characteristics:

Table 1: Likert scale

Verbal interpretation	Scale values	Range
Strongly Agree (SA)	4	4.50-5.00
Agree (A)	3	3.50-4.49
Disagree (D)	2	1.50-2.49
Strongly Disagree (SD)	1	1.00-1.49

- Functional suitability
- Performance efficiency
- Compatibility
- Usability
- Reliability
- Security
- Maintainability
- Portability (Table 1)

RESULTS AND DISCUSSION

Applying ARIMA Model: A valuable tool for predicting future health events or situations such as demands for health services and healthcare needs is through health forecasting. It facilitates preventive medicine and health care intervention strategies by pre-informing health service providers to take appropriate actions to minimize risks and manage demand. It also requires reliable data, information and appropriate analytical tools for the prediction of specific health conditions or situations. As there is no single approach to health forecasting, various methods have often been adopted to forecast aggregate or specific health conditions. However, there are no defined health forecasting horizons (time frames) to match the choices of health forecasting methods/approaches that are often applied. The key principles of health forecasting have not also been adequately described to guide the process[13].

In medical applications, time series forecasting models have been successfully applied to predict the progress of the disease, estimate the mortality rate and assess the time dependent risk. However, the vast availability of many different techniques, in which each type excels in particular situations, makes the process of choosing an appropriate model more challenging[14].

A time series modeling approach (ARIMA Model) has been used in this study to forecast the possible urgency level and chronic kidney disease stages. Time series algorithm was also used to forecast the level of urgency and CKD stages of patients.

Figure 6 illustrates the patient's observation on urgency timeline for 19 days. The observation was from May 12 of 86% to May 30, 2018 of 70%, with unstable observations throughout the duration of predicting the patient's need for a dialysis.

Figure 7 illustrates the future observation of the patient's urgency rate and level. The system forecasted that the patient will be on 93% of urgency rate in the next 6 days and will still be on emergency care and CKD Stage 5.

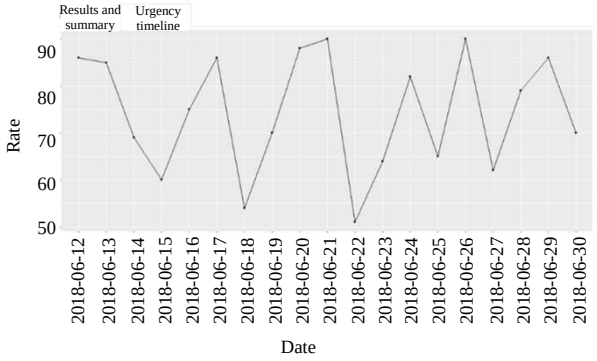

Fig. 6: Seasonal timeline

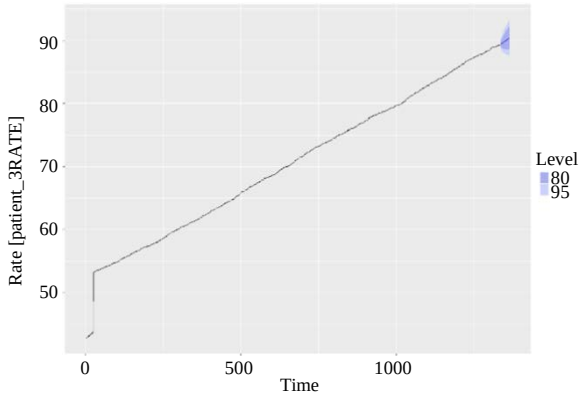

Fig. 7: Urgency level timeline

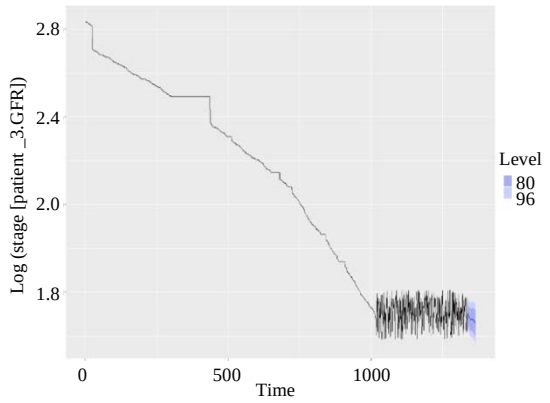

Fig. 8: CKD stage timeline

Figure 8 illustrates the patient's future observation of CKD Stages. The patient's creatinine level is gradually decreasing from 2.9-2.5,2.3, 2.2, 2.0, 1.8, 1.5, 1.4 then it goes unstable increasing and decreasing from 1.6-1.3. The system has forecasted that it will go below at 1.1.

Figure 9 shows that the patient status is under emergency care with the urgency rate of 76% and is currently on CKD stage 5.

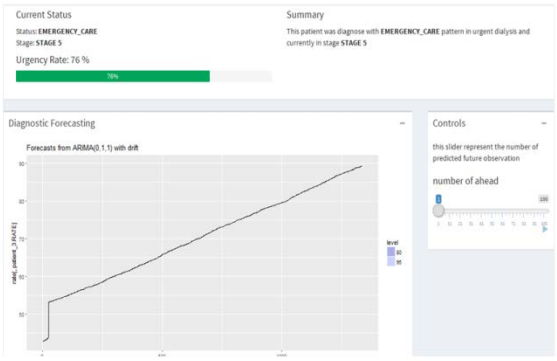

Fig. 9: Recommendation page

CONCLUSION

The objective of this study was to create a prototype to be used in forecasting the CKD stages and urgency level of dialysis patients. The insights and findings provided were able to forecast the CKD stages and urgency of a patient's need for dialysis and the study has also developed a model and dashboard enhancing efficiency and accuracy of results based on patient's medical record in which nephrologists, dialysis centers, hospital administrators and patients can highly benefit from. In line with this, the following recommendations for future research may be considered:

- Consider the level of urgency and CKD stages using clinical variables forecasted by the algorithm
- Conduct a future study that can assist future researchers by giving an idea and enhancement of their forecasting management and data mining topics
- The design of tools to automate some resource-consuming time series analysis tasks, such as preparation

Other recommendation in the field of medicine include: The increase of 0.7% annual mortality rate per year since 1990-2013 exhibits an alarming situation of the Philippines when it comes to treating chronic kidney disease alone. Efforts in improving the health care system of the country may be looked into as problems may be posited gradually with the use of technology, particularly machine-learning systems utilizing the use of data mining tools to better measure the status of patients with other chronic illnesses.

The Hospital management or administrators may consider a venture in artificial intelligence for the healthcare industry which can provide the essential patient electronic information enhanced by the power of analytics and machine learning. This kind of innovative approach can definitely help the Nephrologists and Hospital administrators.

REFERENCES

01. Lafta, R., J. Zhang, X. Tao, Y. Li and W. Abbas *et al*., 2017. A fast Fourier transform-coupled machine learning-based ensemble model for disease risk prediction using a real-life dataset. Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, May 11-14, 2017, Singapore, Singapore, pp: 654-670.
02. Tiao, G.C., 2015. Time Series: ARIMA Methods. In: International Encyclopedia of the Social and Behavioral Sciences, Wright, J. (Ed.). Elsevier, Amsterdam, Netherlands, pp: 15704-15709.
03. Anguera, A., J.M. Barreiro, J.A. Lara and D. Lizcano, 2016. Applying data mining techniques to medical time series: an empirical case study in electroencephalography and stabilometry. Comput. Struct. Biotechnol. J., 14: 185-199.
04. Razlighi, Q.R., E. Stallard, J. Brandt, D. Blacker and M. Albert *et al*., 2014. A new algorithm for predicting time to disease endpoints in Alzheimer's disease patients. J. Alzheimer's Dis., 38: 661-668.
05. Santric-Milicevic, M., V. Vasic and J. Marinkovic, 2013. Physician and nurse supply in Serbia using time-series data. Hum. Resour. Health, Vol. 11, 10.1186/1478-4491-11-27
06. Moskovitch, R. and Y. Shahar, 2015. Classification-driven temporal discretization of multivariate time series. Data Min. Knowl. Discovery, 29: 871-913.
07. Fisher, C.K. and P. Mehta, 2014. Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. PloS One, Vol. 9, No. 7.
08. Guennec, A.L., S. Malinowski and R. Tavenard, 2016. Data augmentation for time series classification using convolutional neural networks. Proceedings of the ECML/PKDD Workshop on Advanced Analytics and Learning on Temporal Data, September 2016, Riva Del Garda, Italy, pp: 1-9.
09. Zhang, G.P., 2003. Time series forecasting using a hybrid arima and neural network model. Neurocomputing, 50: 159-175.
10. Kanarachos, S., S.R.G. Christopoulos, A. Chroneos and M.E. Fitzpatrick, 2017. Detecting anomalies in time series data via a deep learning algorithm combining wavelets, neural networks and Hilbert transform. Expert Syst. Appl., 85: 292-304.
11. Schwaber, K. and M. Beedle, 2001. Agile Software Development with Scrum. 1st Edn., Prentice Hall, New Jersey.
12. Zhu, M., R.Q. Zu, X. Huo, C.J. Bao and Y. Zhao *et al*., 2011. The application of time series analysis in predicting the influenza incidence and early warning. Chin. J. Preventive Med., 45: 1108-1111.
13. Soyiri, I.N. and D.D. Reidpath, 2013. An overview of health forecasting. Environ. Health Preventive Med., 18: 1-9.
14. Bui, C., N. Pham, A. Vo, A. Tran, A. Nguyen and T. Le, 2017. Time series forecasting for healthcare diagnosis and prognostics with the focus on cardiovascular diseases. Proceedings of the International Conference on the Development of Biomedical Engineering in Vietnam, June 27-29, 2017, Springer, Ho Chi Minh, Vietnam, pp: 809-818.