Predicting ICU Mortality:

A Comparison of Stationary and Nonstationary Temporal Models

Mehmet Kayaalp, M.D., M.S.[1], Gregory F. Cooper, M.D., Ph.D.[1], Gilles Clermont, M.D., M.Sc.[2]

[1]Center for Biomedical Informatics, Intelligent Systems Program
[2]Department of Anesthesiology, University of Pittsburgh Medical Center
University of Pittsburgh
Pittsburgh, Pennsylvania
kayaalp@acm.org, gfc@cbmi.upmc.edu, clermontg@anes.upmc.edu

Objective: *This study evaluates the effectiveness of the stationarity assumption in predicting the mortality of intensive care unit (ICU) patients at the ICU discharge.*

Design: *This is a comparative study. A stationary temporal Bayesian network learned from data was compared to a set of (33) nonstationary temporal Bayesian networks learned from data. A process observed as a sequence of events is stationary if its stochastic properties stay the same when the sequence is shifted in a positive or negative direction by a constant time parameter. The temporal Bayesian networks forecast mortalities of patients, where each patient has one record per day. The predictive performance of the stationary model is compared with nonstationary models using the area under the receiver operating characteristics (ROC) curves.*

Results: *The stationary model usually performed best. However, one nonstationary model using large data sets performed significantly better than the stationary model.*

Conclusion: *Results suggest that using a combination of stationary and nonstationary models may predict better than using either alone.*

INTRODUCTION

Temporal modeling is important in numerous clinical domains, including chronic diseases at one extreme and rapidly–progressing acute problems at the other. For these classes of medical problems, we need a robust methodology for providing consistent and reliable temporal decision support to contribute to improved quality of care.

This paper analyzes a key question in temporal process modeling: When should we assume stationarity? Before formally defining stationarity in the next section, we provide an informal definition of stationarity: A process observed as a sequence of events is stationary if its stochastic properties stay the same when the sequence is shifted in a positive or negative direction by a constant time parameter; *i.e.*, its course in a given period is independent of its starting time point.

Sometimes it is possible to detect stationarity in the process by investigating the character of the longitudinal data.[1] This is usually possible when data are collected for a long period, or the domain expert happens to know that the process to be modeled is stationary. Many times, however, the analysis of the data does not yield any conclusive evidence about the stationarity of the process, and the designer has to make a design assumption about stationarity.[2]

In this study, we induce a stationary and 33 nonstationary temporal models from the same medical data, and we compare the differences in predictive performance between the models. Our goal is to determine the merits and drawbacks of the stationary and nonstationary modeling approaches for predicting mortality in the ICU, and to evaluate conditions under which the stationary or nonstationary models are relatively more effective.

BACKGROUND

We used a database of demographic, physiologic and outcome variables collected on 1,449 patients admitted to 40 different ICUs in May 1995. The database contains 11,418 records. Each record contains one day of data on one patient; *i.e.*, the temporal granularity of variables is fixed at one day, except for those variables that are atemporal. The data were originally collected for a prospective study to evaluate a newly established Sequential Organ Failure Assessment (SOFA) score that was intended to assess the incidence and severity of organ dysfunction or failure of ICU patients.[3]

Each record contains the following eight atemporal fields: (**1**) center number (1–40), (**2**) the day in the ICU (1–33; data were collected up to 33 days), (**3**) age (12–95 years of age), (**4**) sex (M/F), (**5**) type of problem motivating admission (1–5; elective surgery, emergency surgery, trauma, medical, and cardiac), (**6**) the origin of the admission (1–5; emergency room, floor, operating room, other acute care hospital, and other origin), (**7**) whether or not it was a readmission to the ICU (Y/N), and (**8**) the status on discharge from the ICU (deceased/survived).

The database contains the following 23 temporal fields: (**1**) oxygenation index, (**2**) mechanical ventilation (Y/N), (**3**) platelet count, (**4**) bilirubin, (**5**) mean arterial pressure,

1067—5027/00/$5.00

doses of (6) dopamine, (7) dobutamine, (8) epinephrine, and (9) norepinephrine, (10) Glasgow Coma Scale, (11) blood urea nitrogen, (12) serum creatinine, (13) urine output, (14) white blood cell count, (15) lowest and highest heart rates, (16) lowest and highest temperature, (17) current state of infection (Y/N), SOFA system scores for the (18) respiratory, (19) cardiovascular, (20) hematological, (21) neurological, and (22) hepatic systems (each between 0–4, where 0 is normal and 4 is pathologically worst), and (23) total SOFA score (linear addition of the former six SOFA system scores).

Patient variables are continuous unless stated otherwise above. We discretized the continuous variables based on medical knowledge and their statistical variances observed in the sample population. In our study, we excluded some other variables contained in the original data set to ensure fairness in forecasting; *e.g.*, the binary (Y/N) variable "do not resuscitate order" can boost model prediction performance significantly, since it may inherently imply a grim prognosis and a non-aggressive therapeutic course.

Among important forecasting problems facing ICU physicians is the probability of patient survival at the discharge from the ICU. We formulated the problem as a stochastic process: Given a sequence of temporal patient data up through day d, what is the probability the patient will die (or its complement, will survive) on day $d+1$.

This is a multivariate stochastic problem. In this study, a multivariate stochastic process is defined as a set of measurable event sequences, where each sequence is comprised of a set of random variables $\mathbf{X}=\{X\}$ associated with time points $t_1,t_2,...\in T$ defined on the temporal space T.

A multivariate stochastic process can be modeled as a temporal Bayesian network. A temporal Bayesian network can be defined in terms of a structure $M=(V,A)$ and a probability space. The structure is comprised of a directed acyclic graph, where nodes $V=\{X(t)\}$ represent temporal random variables, and arcs $A=\{(X_i(t_i),X_j(t_j))\}$ represent pairwise interactions between variables, where $t_i \le t_j$ if $X_i \neq X_j$; otherwise $t_i < t_j$. A temporal Bayesian network is strictly *stationary*, if for every $t_1,t_2,...\in T$

$$P(X_1(t_1),...,X_n(t_n))=P(X_1(t_1+t),...,X_n(t_n+t)) \quad (1)$$

Bayesian networks can be manually constructed by an expert by identifying problem variables and interactions between variables, and by assigning prior probabilities to the event set. In the present research report, we used a machine learning approach to construct Bayesian networks from data automatically. By evaluating the probability distributions in the database, this method can assign a probability score to each possible Bayesian network model encountered during the model search. Among all Bayesian networks considered, one can select a network that best fits the data.[4]

Techniques used for learning atemporal Bayesian networks are applicable to learning temporal Bayesian networks as well. The variable space in temporal Bayesian networks is increased by the factor of time parameters $|T|$, where $t\in T$. Such an increase generally leads to a sparse data set. This problem is known as *the curse of dimensionality*.

One frequently applied remedy to this problem is to assume stationarity in parts of the stochastic process, which leads to partitioning the duration of the process into smaller periods, $[t_1,...,t_n]$, which may be called windows. The subprocess in each window is assumed to be a representative recurrent unit of the entire process; therefore, properties of the stochastic process are assumed to stay the same when the event sequence $(X_1(t_1),...,X_n(t_n))$ is shifted in a positive or negative direction by a constant time parameter t (see Eq.(1)).

Table 1: White Blood Cell Counts (WBCs) of a Patient

Days	WBC	...
1	high	...
2	high	...
3	normal	...
4	normal	...

Consider a patient with four records shown in Table 1, where only one field, WBC, of each record is shown. A nonstationary model would associate each record with an absolute day, on which the measurement was made; thus, the nonstationary model would consist of four days, and use a single data point per temporal variable (see Table 2).

Table 2: Four Days of Records Used as a Single Event in a Nonstationary Model

WBC_1	WBC_2	WBC_3	WBC_4	...
high	high	normal	normal	...

On the other hand, a stationary model with two time-slices would associate measurements with both stationary variables sequentially (see Table 3).

Table 3: Four Days of Records Used as Four Events in a Stationary Model With Two Time-slices

WBC_1	WBC_2	...
unknown	high	...
high	high	...
high	normal	...
normal	normal	...

As seen in Table 2 and Table 3, the nonstationary model treats two WBCs on each day as being unique, whereas the stationary model groups sequential pairs of WBC values.

METHODS

The methods used in this study can be described in three parts: (1) preprocessing the data, (2) learning models from the data, and (3) testing and evaluating the models.

The first data preprocessing step was variable selection. We include only those variables described above. The second data preprocessing step was the discretization of the continuous variables. Each variable distribution was analyzed separately. Variables were discretized manually based on the range of normal values and prior known relationships of variables to mortality; *e.g.*, very low and very high white blood cell counts are both associated with higher mortality, so this variable was discretized in three categories. Missing values were labeled with the category *unknown* and processed along with other categorical values.

The third step of the data preprocessing was determining training and test sets. The 1,449 patient cases were randomly split into two disjoint sets: one with 949 patients for the training, and the other with 500 patients for the test. The test data set was not used in any part of the model learning process.

For inducing the nonstationary models, the training data set was partitioned into 33 subsets, where the length of ICU stay was the same for all patients within each subset. The length of stay of patients in the database varied between 1 and 33 days with the following exception: four of nine patients who stayed longer than 33 days died after the 34th day; nonetheless, we treated them as if they stayed only 33 days, and died on day 34. We treated the other five patients as if they were discharged on day 33.

The criteria used in preprocessing the data for inducing the stationary model were as follows. The outcome variable (*i.e.*, ICU mortality) belongs to the time-slice n for a patient case with $n-1$ records. The preprocessing method should take into account that the patient was alive prior to day n; therefore, in all stationary event sequences, except in the last one if the patient died, the mortality variable should be instantiated as alive. In this experiment, we set the stationarity window with five time-slices, where the fifth time-slice contains only the mortality variable model.

First, the temporal variables (*e.g.*, white blood cell count), atemporal variables (*e.g.*, sex), and the outcome variable were partitioned into different sets. For each patient, a set of training records was created whose fields consisted of (1) values of the atemporal variables, (2) values of the temporal variables corresponding to *every* four consecutive ICU days, and (3) the value of the classification variable on the fifth consecutive day.

The last step of preprocessing involved forming proper test data for the stationary and nonstationary experiments. For the nonstationary experiments, preprocessing of the test data did not differ from that done for the training data. For the stationary experiment, the temporal data during the last five days of ICU stay of each patient were collected along with the atemporal data. For the patients who stayed in the ICU for d days, where $d < 4$, values of the temporal variables between day 1 and day $4-d$ were set to *unknown*.

Data preprocessing was followed by model learning from the training data. For the stationary experiments, there was only one set of data; consequently, one stationary model was constructed based on that data set. For the nonstationary experiments, however, there were 33 distinct data sets; thus, 33 nonstationary models were learned.

Finding a Bayesian network that fits the data is a model selection problem. Because the number of all possible models grows exponentially with the number of variables, the common approach for finding a "good" model is heuristic search, which does not guarantee finding the best model. The model scoring metric used in this study is based on the following Bayesian score: [4,5]

$$P(M \mid D) \propto \prod_{i}^{n} \prod_{j}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})} \qquad (2)$$

Here, D is the database of training cases; n is the number of nodes (variables) in the Bayesian network model M; Γ is the gamma function; q_i is the number of joint states of the parents of node i ($q_i = 1$ if node i has no parents); r_i is the number of states of node i; N_{ijk} is the number of cases in D that node i has value k and the parents of node i have the state denoted by j; α_{ijk} denotes a Dirichlet prior parameter. We assume uniform priors for variables, namely $\alpha_{ijk} = 1$ for all i, j, k. $\alpha_{ij} = \sum_k^{r_i} \alpha_{ijk}$ and $N_{ij} = \sum_k^{r_i} N_{ijk}$. Eq.(2) is not an equality, but rather a proportionality, where a uniform prior $P(M)$ is assumed for all structures. Further assumptions made in this modeling methodology can be found in other reports. [4,5]

The intended use of the model M is forecasting the mortality R of a patient given data D, *i.e.*, $P(R \mid D, M)$. The inference requires only a subset of variables in D that is the set of parent nodes of the mortality variable. Let R be the mortality variable and $\pi(R)$ be parent variables of R, then the probability of interest is $P(R \mid D, \pi(R))$. In other words, the model search task can be simplified to the identification of a set of parents of R. Since the structure of model M consists of R and $\pi(R)$ only, $P(M \mid D)$ in Eq.(2) can also be denoted as $P(\pi(R) \mid D)$. The search algorithm given below assumes that the class of interest is the n^{th} node; *i.e.*, $\pi(R) = \pi(n)$.

1. $M \leftarrow (\{1,...,n\},\{\})$, *i.e.*, $\pi(n) \leftarrow \{\}$
2. $score \leftarrow P(\pi(n) \mid D)$
3. **for** $i:1 \rightarrow n-1$ and $i \notin \pi(n)$ and $|\pi(n)| < n-1$
   a. **push** i **to** $\pi(n)$
   b. **if** $P(\pi(n) \mid D) > score$
      then $score \leftarrow P(\pi(n) \mid D)$, $flag \leftarrow$ up,
         $candidate \leftarrow$ **pop** $\pi(n)$
      else pop $\pi(n)$
4. **if** $flag$ = up
   then push *candidate* **to** $\pi(n)$, $flag \leftarrow$ down, **goto** 3
   else return $\pi(n)$

This search algorithm returns a model that maximizes the Bayesian score for the model structure given the database. As this is a stepwise-forward, greedy algorithm, the global maximum is not guaranteed; *i.e.*, the result is a local maximum.

The resulting models were applied to the test data in both stationary and nonstationary cases. For each test case C, the probability of patient mortality was computed as

$$P(R=d \mid C) = \frac{n(R=d, \pi_c(R)) + 0.22}{n(R=d, \pi_c(R)) + n(R=s, \pi_c(R)) + 1}$$

where d and s stand for deceased and survived, respectively; $n(\cdot)$ denotes the frequency count of the instantiated variables in the training data set. $\pi(R)$, the parents of the mortality variable, are found via the search described above. $\pi_c(R)$ denotes the parents of R, with values determined by the variable values of the test case C during inference. We assumed that the prior probability of ICU mortality can be assessed from an independent data set. In inference, we set the priors for deceased as 0.22, which is the frequency of mortality in the training sample.

The forecasting results were evaluated using an ROC metric.

RESULTS

The stationary model that locally maximizes Eq.(2) was a Bayesian network with two nodes; namely, the *SOFA total score on the last day prior to discharge from the ICU*. The ROC curve of the stationary model is the top curve in Figure 1. The area under the ROC curve is 0.83, where 1.0 indicates the entire area.

Twenty-four of 33 nonstationary models have single predictors of mortality. There are no other predictors of mortality, presumably due to small sample sizes for model induction. For the first nonstationary model, which is the model of patients who stayed only one day in the ICU, the predictive variable was the *dose of administered dopamine*. For the second nonstationary model, the *total SOFA score on the second day* was identified as the predictive variable, as in the stationary case. The

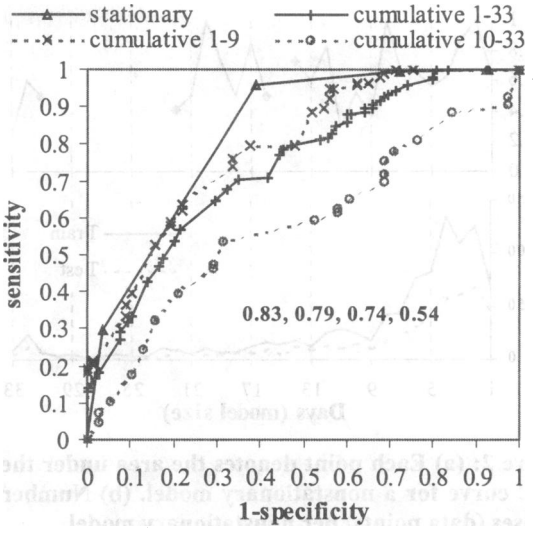

Figure 1: ROC curves for stationary and nonstationary models. Areas under the ROC curves for the stationary model and for all nonstationary models were 0.83 and 0.74, respectively. Combined predictions of nonstationary models 1 to 9 and 10 to 33 have ROC areas of 0.79 and 0.54, respectively.

nonstationary models between three and nine days had the same predictive variable, which was *the mechanical ventilation on the last day prior to discharge.*

According to these results, for patients staying a single day in the ICU, the presence of hypotension (or its complement) is highly prognostic of outcome. Similarly, the prognosis of a patient who stays in the ICU for more than two days depends on the presence or absence of mechanical ventilation.

Over all nonstationary models, there were 11 models with the variable, *mechanical ventilation*, and 6 models with hypotension related variables (2 *dopamine*, 1 *dobutamine*, 1 *norepinephrine*, 1 *SOFA cardiac*, and 1 *mean arterial pressure measure*).

Compared to the test set of the stationary experiment, nonstationary data sets were rather sparse. In certain nonstationary test sets, no patients survived, whereas in some others, no patients died. The ROC metric is relatively uninformative in those situations, and we therefore excluded those sets from analysis. Figure 2(a) shows ROC areas plotted for the nonstationary models.

The fluctuation observed in Figure 2(a) is due to the small numbers of test cases for some nonstationary models. The reliability of ROC scores improves when the number of data points in both test and training sets increases.

In Figure 1, the curve labeled as "*cumulative 1–33*" delineates the ROC points for all nonstationary models, 1 to 33; *i.e.*, predictions of all nonstationary models are

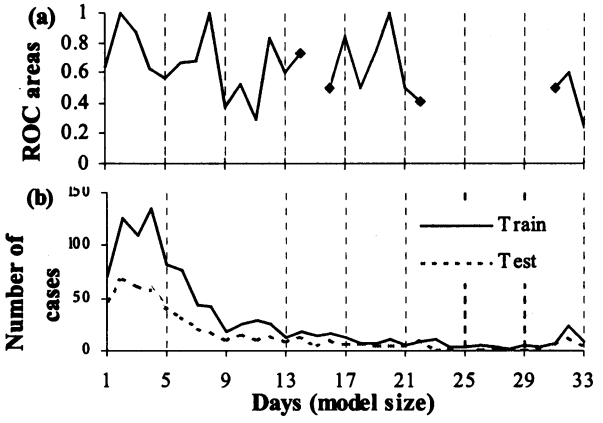

Figure 2: (a) Each point denotes the area under the ROC curve for a nonstationary model. (b) Number of cases (data points) per nonstationary model.

evaluated with a single ROC curve. The curves plotted with dashed lines are cumulative ROC curves for the nonstationary models 1 to 9 (the second curve from the top) and 10 to 33 (the curve at the bottom), where the area under the ROC curve decreases from 0.79 to 0.54. This decay is due to the small number of both training and test cases. When the number of training cases is small, the constructed structure and its parameterization are suboptimal, whereas when the number of test cases is small, the results are statistically not meaningful; therefore, in Figure 2, (a) and (b) are plotted next to each other.

While the model size gets larger, the number of data points decreases; therefore, the predictive performance of nonstationary models decays significantly while the model size grows, as seen in Figure 1 and Figure 2(b).

Computations were executed on a SUN workstation. Each nonstationary model was constructed in a few seconds (on average in nine seconds), whereas the stationary model was constructed in approximately two minutes. The time required for inference was one second per five test cases using the stationary model, whereas it took only three seconds for all 500 test cases using nonstationary models.

CONCLUSIONS

The results of this study are consistent with clinical experience.[6,7] The total SOFA score, reflecting the collective burden of organ system dysfunction, was found to be predictive in the second nonstationary model and in the stationary model. One explanation for this outcome is that the number of nonstationary training cases for model two is high enough to identify the total SOFA score as a highly predictive parent variable. In the nonstationary model on day 2, predictive performance is far better than that of the stationary model; areas under the ROC curves were 1.0 vs. 0.83, respectively. Because of the large number of test cases in both experiments, it is unlikely that this difference in performance is incidental.

This result indicates that nonstationary models may perform as well as or better than stationary models when there are a large number of training cases. The stationarity assumption increases the number of effective data points by reducing the model dimensions; however, due to the limitations of the assumption, parameterization of the Bayesian networks are suboptimal, which negatively influences predictive performance.

We plan to investigate methods that use a hybrid stationary and nonstationary modeling methodology. The goal is to take advantage of any predictors that are approximately stationary, yet also model other predictors that are nonstationary.

Acknowledgements

We thank Drs. Jean-Louis Vincent, Rui Moreno, and the European Society of Intensive Care Medicine for the provision of the SOFA dataset and their support of this study. This work was supported by the National Library of Medicine with the grant "Integrated Advanced Information Management Systems" No. G08-LM06625 and with the grant No. R01-LM06696.

References

1. Riva A, Bellazzi R. Learning temporal probabilistic causal models from longitudinal data. *Artificial Intelligence in Medicine* 1996; 8:217–234.
2. Manuca R, Savit R. Stationarity and nonstationarity in time series analysis. *Physica D* 1996; 99:134–161.
3. Vincent J-L, de Mendonca A, Cantraine F, Moreno R, Blecher S. Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units: Results of a multicenter, prospective study. *Critical Care Medicine* 1998; 26(11):1793–1800.
4. Cooper GF, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning* 1992; 9:309–347.
5. Heckerman D, Geiger D, Chickering DM. Learning Bayesian networks: the combination of knowledge and statistical data. *Machine Learning* 1995; 20(3):197–243.
6. Angus DC, Linde-Zwirble WT, Clermont G. The incidence of organ failure and its impact on mortality and resource use in hospitalized community acquired pneumonia. *Am J Resp Crit Care Med* 1997; 155(4):A929.
7. Linde-Zwirble WT, Clermont G, Coleman MB, Brodak S, Angus DC. Incidence of ARDS in the US, Europe and Japan. *Intensive Care Med* 1996; 22(Suppl 3):272.