Using electronic health records to predict severity of condition for congestive heart failure patients

Costas Sideris
University of California, Los Angeles
3256N, Boelter Hall
Los Angeles, CA 90095
costas@cs.ucla.edu

Behnam Shahbazi
University of California, Los Angeles
3256N, Boelter Hall
Los Angeles, CA 90095
bshahbazi@cs.ucla.edu

Mohammad Pourhomayoun
University of California, Los Angeles
3256N, Boelter Hall
Los Angeles, CA 90095
mpourhoma@cs.ucla.edu

Nabil Alshurafa
University of California, Los Angeles
3256N, Boelter Hall
Los Angeles, CA 90095
nabil@cs.ucla.edu

Majid Sarrafzadeh
University of California, Los Angeles
Boelter Hall
Los Angeles, CA 90095
majid@cs.ucla.edu

Ubicomp '14, September 13 - 17 2014, Seattle, WA, USA
Copyright © 2014 ACM 978-1-4503-3047-3/14/09$15.00.
http://dx.doi.org/10.1145/2638728.2638815

Abstract

We propose a novel way to design an analytics engine based exclusively on electronic health records (EHR). We focus our efforts on Congestive Heart Failure (CHF) patients, although our approach could be extended to other chronic conditions. Our goal is to construct statistical models that predict a CHF patient's length of stay and by extension the severity of his/her condition. We show that it is possible to predict length of hospital stay based on physiological data collected from the first day of hospitalization. Using 10-fold cross validation we achieve accurate predictions with a root mean square error of 3.3 days for hospital stays that are less than 15 days in duration. We also propose a clustering of patients that organizes them to risk groups according to their estimated severity of condition.

Author Keywords

Electronic Health Records, Remote Monitoring Systems, Heart Failure, Intervention, Readmission

ACM Classification Keywords

J.3 [Computer Applications]: Life and Medical Sciences; H.1.2 [Information Systems]: Human information processing

Introduction

Recent advances in wireless sensors, mobile technologies, and cloud computing have made continuous remote monitoring of patients possible [1],[6],[8],[12],[13]. These Remote Health Monitoring Systems (RMS) provide a continuous stream of patient physiological data that allows nurses and doctors to make timely decisions and help patients manage their chronic conditions while minimizing hospital readmission rates.

Conventional RMS rely on threshold based alerts. That is, thresholds based on medical expertise are put in place to alert clinicians when physiological data deviate from the set thresholds. Analytics-based RMS on the other hand employ machine learning algorithms to predict the risk of a medical adverse event. It has been shown [9] that analytics-based RMS work better than threshold-based ones and can help reduce treatment costs.

Traditionally, designing such machine learning algorithms is guided by an initial, usually smaller pilot study that provides sufficient data for design and validation. This approach however suffers from several limitations. First of all, it requires a substantial effort to collect such initial data, spanning months or even years depending on the studied chronic condition. In addition, selecting which physiological data to collect as well as the collection frequency is based on guesswork since there is limited data to support the decision.

In this paper we advocate the use of electronic health records to guide the design of such machine learning algorithms. The benefits of such an approach is that it is based on a vast amount of electronic health records that are available and can be created before the RMS is even deployed. We focus our efforts on Congestive Heart Failure (CHF) patients. We present two methodologies to predict and classify the severity of a patient's condition. These algorithms can be used to identify severity of condition for patients in an RMS and assess the need for intervention, as well as to help schedule clinician time.

Methodology

Data Processing

We used electronic health records (EHR) from the Ronald Reagan UCLA Medical Center between 2005 and 2009. We first identified CHF hospitalizations for adult patients using the same International Classification of Diseases version 9 (ICD-9) codes used in the Centers for Medicare & Medicaid Services (CMS) 30-day readmission measure. A total of 1179 admissions were extracted with primary diagnosis a CHF related ICD-9 code. These records correspond to 913 unique patients out of which 169 had more than one CHF related admission. For each admission record we extracted 7 features: min, max, range, average, median, standard deviation and variance, computed from the patient's heart rate, systolic/diastolic blood pressure, and weight during the first day of hospital stay. Furthermore, 19 categorical features are calculated based on the patient's age, gender, race, ethnicity, ZIP code, type of external care, insurance coverage, perceived severity of condition, perceived mortality and the first 10 comorbidities reported.

Attribute Selection

We rank these features based on their correlation with a patient's length of stay in the hospital. Using correlation-based feature subset selection [5] we find that the most prominent categorical and numerical features are age, gender, ethnicity, minimum systolic pressure and heart rate range for the first day (See Table 1).

1.	age
2.	gender
3.	ethnicity
4.	min. systolic pressure first day
5.	range of heart rate first day
6.	std systolic pressure first day
7.	std diastolic pressure first day
8.	perceived severity
9.	perceived mortality
10.	insurance
11.	type of outside care
12.	2nd co-morbidity
13.	4th co-morbidity
14.	5th co-morbidity
15.	8th co-morbidity

Table 1: Attributes as ranked by the correlation-based feature subset selection

Predicting Length of Stay

A good indicator of a patient's severity of condition is the Length of Stay (LOS) in the hospital. As multiple previous studies have indicated [3],[10],[11] contextual information such as age, sex and comorbidities can affect the LOS of a patient along with the actual condition. To predict the LOS of a patient, we use the most correlated categorical features as well as the most correlated vitals that correspond to the first day. Using generalized linear regression we combine the effects of a patient's vitals with the effects of his contextual information to predict LOS. Table 2 presents the average root mean square error (RMSE) using 10-fold cross validation and including only patients that stayed less than 15 days, 30 days, and 60 days respectively. It can be seen that as we try to predict longer stays the RMSE increases.This could be due to the lack of data for these longer stays since the vast majority of patients stay less than 15 days. It could also be due to the fact that for longer stays we need intermediate information from a patient's hospital stay.

LOS Considered	RMSE (days)	No.Records
$\leq$ 15 days	3.2788	948
$\leq$ 30 days	4.9527	1049
$\leq$ 60 days	8.2770	1125

Table 2: Average root mean square error of generalized linear regression with 10-fold cross validation .

Grouping Patients by Severity of Condition

In the previous section we showed that it is possible to predict a patient's length of stay based on contextual information and first day vitals. This prediction model can be useful when evaluating a patient on a RMS. It provides nursing staff with an estimate of how long that patient would stay in the hospital if he/she were to be admitted that day. To simplify the decision process for the support staff, we design a clustering algorithm that groups patients in terms of their individual risk based on the top two vital features (minimum systolic pressure, heart rate range for the first day). In this study, we define LOS plus 3 times ICU stay as a "risk factor":

$$riskfactor = LOS + 3 * ICU \qquad (1)$$

We cluster the admission records based on hierarchical clustering [7]. Through experimentation we found that using Chebyshev's distance function [2] and complete linkage for the cost function [4] results in the most stable clusters. Figure 1 demonstrates the results of the hierarchical clustering using 12 clusters. We rank and color the clusters based on the average risk from low to high. Red corresponds to the highest and blue/green to the lowest risk.

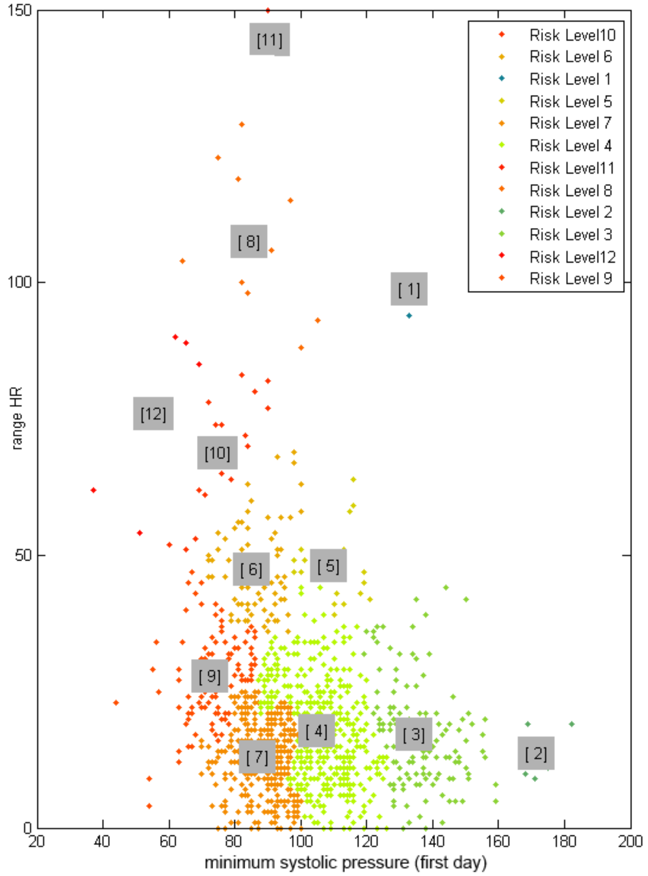

Figure 1: Hierarchical Clustering of patients according to their minimum systolic pressure, heart rate range during the first day of hospitalization. Colors represent average risk per group, with blue/green and red representing the lowest and highest risk respectively.

As it can be seen in Figure 1, there are some issues with our risk factor as the rankings do not follow medical expertise. For example, people with really high systolic blood pressure appear to have shorter duration of stay at the hospital than people with normal levels. The main reason for these discrepancies is that there is a strong age bias on how long a patient stays in the hospital as previous studies have shown [3].
Figure 2 shows that CHF patients between the ages of 55 and 65 years old for example end up staying in the hospital longer on average than older or younger patients.

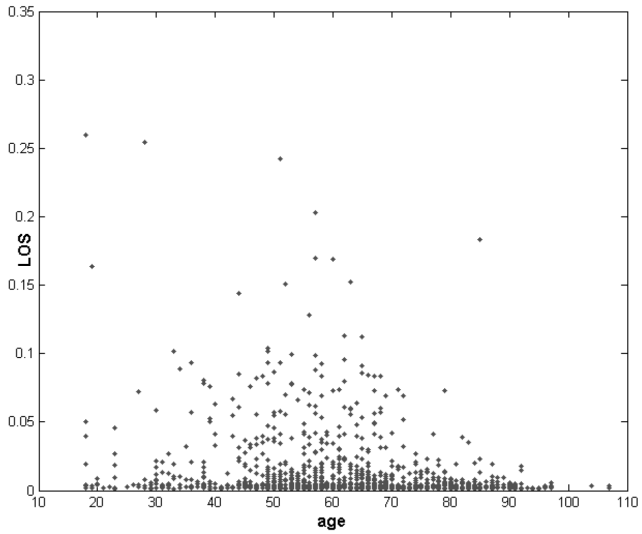

Figure 2: Age versus Length of Stay for CHF patients. It can be seen that there is an age bias on how long a patient stays in the hospital.

To amend these issues and obtain a more objective risk grouping regardless of age/sex, we normalize the risk factor of each patient by multiplying it with a value proportional to the average risk factor of his age and sex group. With that modification, and with the same clustering methodology, we obtain a risk ranking that is closely aligned with medical expertise. Figure 3 demonstrates the results of the hierarchical clustering with the modified risk function.

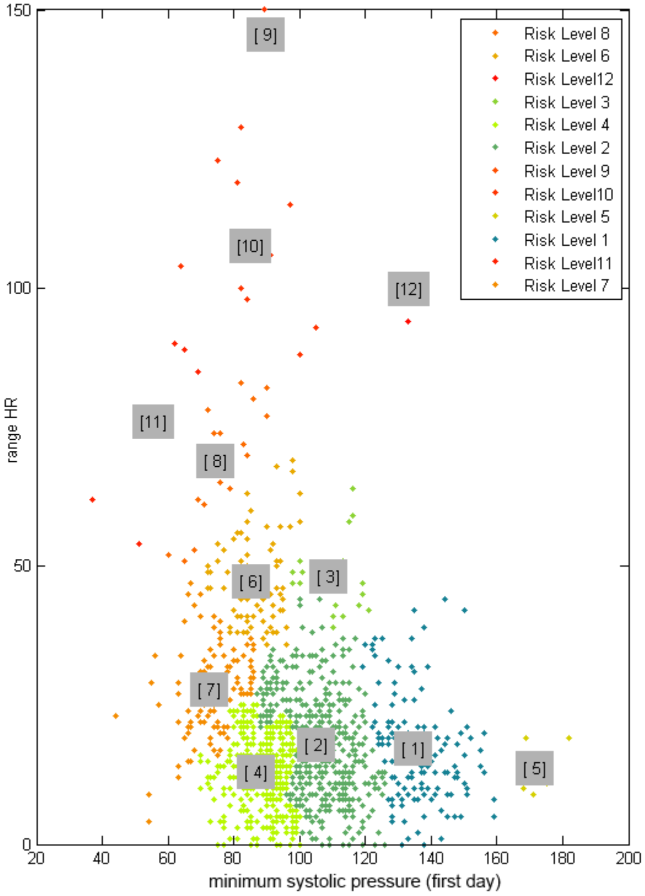

Figure 3: Hierarchical Clustering of patients with the risk factor normalized with regards to age/sex.

The above clustering suggests that it is possible to rank a patient's severity of condition based on systolic pressure and heart rate measurements. This information can be useful in the context of a RMS for clinicians. Threshold alerts such as "systolic pressure above 130 millimeters of mercury" end up overwhelming the clinicians as they tend to accumulate very fast with a growing number of patients. Our clustering methodology provides a much more detailed status for each patient as it allows ranking severity and making more informed decisions. In addition, such a scheme could be employed to suggest short term and long term health improvements for the patients.

Conclusions

Using statistical models constructed from electronic health records, we have shown that it is possible to predict the length of stay of a CHF patient with high accuracy. We have also presented a methodology for modeling patient groups in terms of severity of condition based solely on hospital records.

These models can be used to analyze daily information collected from an RMS and allow a nurse/doctor to better design their intervention. The risk factor information could also be used to provide personalized advice to the patient on how to improve their condition in the short and long term as to avoid readmission.

In the future we plan to further validate our approach by analyzing data collected by a RMS for congestive heart failure patients and comparing it with our predicted patient risk. We also plan to extend our approach to different chronic conditions such as diabetes and liver disease.

References

[1] Bar-Or, A., Healey, J., Kontothanassis, L., and Van Thong, J. Biostream: A system architecture for real-time processing of physiological signals. In *Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE*, vol. 2, IEEE (2004), 3101–3104.

[2] Cha, S.-H. Comprehensive survey on distance/similarity measures between probability density functions. *City 1*, 2 (2007), 1.

[3] Charlson, M. E., Pompei, P., Ales, K. L., and MacKenzie, C. R. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of chronic diseases 40*, 5 (1987), 373–383.

[4] Defays, D. An efficient algorithm for a complete link method. *The Computer Journal 20*, 4 (1977), 364–366.

[5] Hall, M. A., and Smith, L. A. Practical feature subset selection for machine learning.

[6] Jovanov, E., Raskovic, D., Price, J., Krishnamurthy, A., Chapman, J., and Moore, A. Patient monitoring using personal area networks of wireless intelligent sensors. *Biomedical Sciences Instrumentation 37* (2001), 373–378.

[7] Kaufman, L., and Rousseeuw, P. J. *Finding groups in data: an introduction to cluster analysis*, vol. 344. John Wiley & Sons, 2009.

[8] Lan, M., Samy, L., Alshurafa, N., Suh, M.-K., Ghasemzadeh, H., Macabasco-O'Connell, A., and Sarrafzadeh, M. Wanda: An end-to-end remote health monitoring and analytics system for heart failure patients. In *Proceedings of the Conference on Wireless Health*, WH '12, ACM (New York, NY, USA, 2012), 9:1–9:8.

[9] Lee, S. I., Ghasemzadeh, H., Mortazavi, B., Lan, M., Alshurafa, N., Ong, M., and Sarrafzadeh, M. Remote patient monitoring: What impact can data analytics have on cost?

[10] Librero, J., Peiró, S., and Ordiñana, R. Chronic comorbidity and outcomes of hospital care: length of stay, mortality, and readmission at 30 and 365 days. *Journal of clinical epidemiology 52*, 3 (1999), 171–179.

[11] Rochon, P. A., Katz, J. N., Morrow, L. A., McGlinchey-Berroth, R., Ahlquist, M. M., Sarkarati, M., and Minaker, K. L. Comorbid illness is associated with survival and length of hospital stay in patients with chronic disability: a prospective comparison of three comorbidity indices. *Medical care 34*, 11 (1996), 1093–1101.

[12] Suh, M.-k., Chen, C.-A., Woodbridge, J., Tu, M. K., Kim, J. I., Nahapetian, A., Evangelista, L. S., and Sarrafzadeh, M. A remote patient monitoring system for congestive heart failure. *Journal of medical systems 35*, 5 (2011), 1165–1179.

[13] Suh, M.-k., Moin, T., Woodbridge, J., Lan, M., Ghasemzadeh, H., Bui, A., Ahmadi, S., and Sarrafzadeh, M. Dynamic self-adaptive remote health monitoring system for diabetics. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, IEEE (2012), 2223–2226.