

Heart Failure Hospitalization Prediction in Remote Patient Management Systems

M. Pechenizkiy, E. Vasilyeva, I. Žliobaitė
Department of Computer Science
P.O. Box 513, 5600 MB, TU Eindhoven
the Netherlands

{m.pechenizkiy,e.vasilyeva,i.zliobaite}@tue.nl

A. Tesanovic, G. Manev
Philips Research Laboratories
HTC 5, 5656 AE Eindhoven
the Netherlands

aleksandra.tesanovic@philips.com

Abstract

Healthcare systems are shifting from patient care in hospitals to monitored care at home. It is expected to improve the quality of care without exploding the costs. Remote patient management (RPM) systems offer a great potential in monitoring patients with chronic diseases, like heart failure or diabetes. Patient modeling in RPM systems opens opportunities in two broad directions: personalizing information services, and alerting medical personnel about the changing conditions of a patient. In this study we focus on heart failure hospitalization (HFH) prediction, which is a particular problem of patient modeling for alerting. We formulate a short term HFH prediction problem and show how to address it with a data mining approach. We emphasize challenges related to the heterogeneity, different types and periodicity of the data available in RPM systems. We present an experimental study on HFH prediction using, which results lay a foundation for further studies and implementation of alerting and personalization services in RPM systems.

1 Introduction

Chronic diseases are the leading cause of death and healthcare costs in the developed countries. Healthcare systems are shifting from patient care in hospitals to monitored care at home [7]. Adequate patient monitoring, instruction, education and motivation can be done outside of the hospital using Remote Patient Management (RPM) systems. RPM systems are expected to assist in normalization of patient condition and preventing re-hospitalization. Figure 1 shows an example of an RPM system.

Recently, a possible architecture of the next generation personalized RPM systems was introduced [5]. The study presented a general process of knowledge discovery from RPM data. It identifies potentially useful features and pat-

terns, which are used for patient modeling and constructing the adaptation rules.

In this study we formulate and address the problem of hospitalization prediction, which is a part of patient modeling. In broader terms hospitalization means severe worsening of a patient condition. Informally we distinguish five horizons of prediction: next step, short, medium, long term and life-long prediction. Life-long prediction is out of the scope of patient modeling. Next step horizon means a couple of hours, short term means a couple of weeks, medium means months and long term means years. Different prediction horizons relate to different possible actions.

We focus on a *short term* hospitalization prediction, which is the most relevant in terms of extraordinary medical actions to prevent the upcoming worsening (Section 2). We study a case of repeated heart failure. Repeated means that the patients have already had a heart failure and are being monitored. Previously, decision rules that should trigger an alarm in case of possible Heart Failure Hospitalization (HFH) have been designed manually based on the domain expertise. We employ a data mining approach (Section 3) for patient modeling, which utilizes information across different data sources. Our study, preliminary result of which were discussed in [4], shows that it is possible to learn predictive models, that outperform the trigger rules, authored by the experts, in terms of their accuracy (Section 4).

2 Background and problem definition

This section presents RPM setting and the problem of HFH prediction in its context.

2.1 Remote Patient Management systems

Existing commercial RPM systems typically provide an end-to-end infrastructure that connects patients at home with medical professionals at their institutions (Figure 1).

In general, RPM systems, e.g. *Card Guard* iTV (www.cardguard.com), *Philips Motiva* (www.healthcare.philips.com), support two workspaces. The first is for the doctors to monitor conditions of the patients and adjust therapy. The second is for the patients to post the symptoms and exchange information with the responding medical professional. Some systems allow the delivery of personalized documents to the patients such as information about the disease, healthy lifestyle recommendations or suggestions on a diet.

Home monitored patients generate various types of data recorded by different means (Table 1). Data collected during RPM process contains typically both objective (vital signs) and subjective (questionnaires) measurements about the condition of a patient. The vital sign measurements are collected using sensors and transferred to the monitoring and management server via application hosting device. The signs to be monitored depend on the chronic disease in question. Weight and blood pressure is typically monitored for heart failure (HF) patients, glucose and weight – for diabetes patients. Subjective measurements, collected from the patients via questionnaires, include symptoms and quality of life (QoL) scores. The questionnaires can be presented to patients directly via an application hosting device or a feedback device, such as a TV.

Based on the indicated deviations from the normal values, a medical professional can adjust the treatment plan including medications, nutrition and physical activity.

2.2 Predicting HF hospitalization

Heart failure is one of the most severe cardiovascular diseases. It causes high mortality and implies high treatment costs. Early and accurate detection of HF situations makes it possible for RPM systems to intervene with appropriate education, instructions, and medications. Timely intervention is expected to improve the condition of a patient as well as to reduce the future treatment costs.

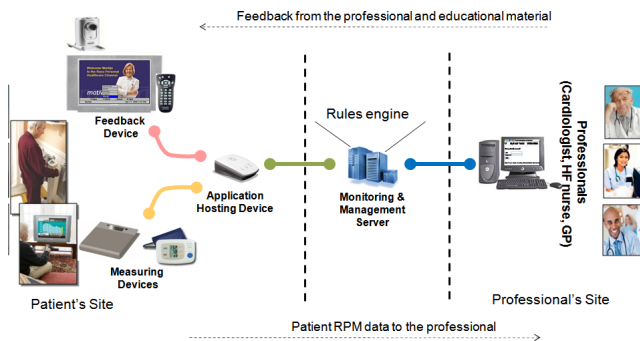


Figure 1. Architecture of an RPM system.

Table 1. Types of home patient data.

Data classes		Collected via	(Typical) Frequency
Medical history	Causes	Face to face meeting at a medical professional's institution	Once, when diagnosis for chronic condition is made
	Co-morbidities		
	Prior hospitaliz-s		
	Implanted devices		
Baseline data	Vitals	Face to face meeting at a medical professional's institution	Every few month, during regular follow-up
	Hight		
	Other diagnosis		
	Lab results		
Vital signs	Weight	An RPM system at a patient's home	Daily
	Blood pressure		
	Pulse		
Question-naires	Symptoms	Several alternatives: - An RPM system at a patient's home, but also can be collected: - Via a telephone contact by a medical professional - Via face to face meeting during regular checkups at medical professional institution	Varies depending on the protocol of care and can be collected: - Daily (RPM) - Weekly (RPM) - Montly (telephone) - Few months (face to face meetings)
	Depression		
	Anxiety		
	Overall health		
	Overall QoL		
	Stress		
	Sleep patterns		
	Fatigue		
Lonliness			
Bio-markers		Face to face meeting at a medical professional's institution	(Few) months
Medications	Disease related drugs	- Via a telephone contact by a medical professional	Few weeks to few months
	Non-disease related drugs	- Via face to face meeting during regular checkups	

Patients with chronic HF have phases of clinical stability interrupted by episodes of worsening. In such case, a previously stable chronic HF patient shows worsening symptoms that the body cannot compensate any more. Worsening of HF may lead or not lead to hospitalization of the patient. Although, both are important, we focus on the problem of heart failure hospitalization (HFH) prediction as the case of prior importance. HF patients might also be hospitalized due to non HF reasons. In this study we consider a hospitalization as HFH, if the first diagnosis was 'heart failure' or the primary admission reason was worsening HF.

We consider a *short term* prediction of worsening. In this study we define short term as two weeks. Short period is particularly relevant for HFH prediction in terms of possible follow up actions, like an extraordinary appointment with a doctor. If the period is too long, like several months, then the prediction output is not relevant in terms of actions.

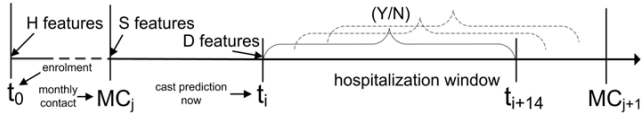


Figure 2. Predicting within the next 14 days.

HFH is expected for HF patients anyway. If the period is too short, like a day or so, then it is too late to take an action other than to hospitalize the patient anyway.

We formulate the problem of HFH prediction in the following way. For a given patient the task is to predict if HF hospitalization will occur within the next 14 days. Assume today is day t_i . The predictive features are formed using the patient data collected up and including day t_i . Positive prediction means that HFH is likely to occur in the period from day t_{i+1} to t_{i+14} , which we call *the hospitalization window*.

The prediction is casted every day. Every day new measurements become available and the predictive features are updated. Every day the hospitalization window moves one day further. Based on the daily prediction output an alarm is raised if deemed necessary. The timeline of the data availability is illustrated in Figure 2. The data is used by medical experts or an automated classifier for the prediction and the following decision making.

The patient data has different periodicity. Medical history data (H features) is recorded at the time of enrolment (t_0). It includes information related to previous hospital admissions, existence of valve diseases, evidence of coronary diseases, arrhythmias, devices implanted and more. A record may contain dozens of fields, typically it is recorded only once. Quality of life symptoms (S features) are recorded approximately every month, during a phone contact (MC_j). The patients are also asked to report additional data such as disease and non-disease medication or medication change; a number of visits or contacts in the last month at home, by phone, at the office, at the clinic. The vital signs (D features), such as weight or blood pressure, are measured on a daily basis using sensors.

3 Data mining approach to HFH prediction

HFH prediction can be addressed as a time-series prediction or a classification task. We focus on the classification task formulation.

In order to learn a classifier we need labeled training data with both positive and negative instances, each represented by a set of the features, which are expected to be predictive. The process of forming a training set is illustrated in Figure 3. Given a large number of H , S and D features as well

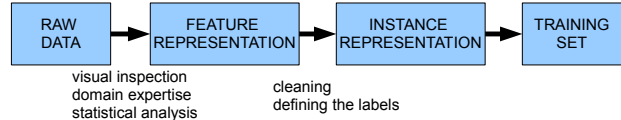


Figure 3. Training set formation process.

as different periodicity they are measured, a feature set construction is the major challenge. In this section we discuss how to identify potentially useful features and how to form positive and negative instances.

3.1 Feature extraction and construction

HFH may be explained by a number of factors [1, 3, 6]. Demographics and baseline measurements are considered to be important for a long term prediction of the health status. Daily, weekly or monthly measurements are expected to have a short-term prognosis value. From the medical domain it is known that different symptoms and signs may cause and possibly predict HFH. However, recent studies focused mainly on daily measurements such as weight dynamics for predicting HFH.

In this study we assemble the primary set of predictive features by visually exploring the data¹. We aim to get a better understanding of what features may potentially describe patient current state and their short and long-term dynamics. To gather this information we employ event-pattern analysis and exploration of time-series data.

For *the event-pattern analysis* we use dot charts. Dot charts give an insight into *frequency* of and *precedence* of events starting from the beginning of the clinical study. They also assist in finding the outliers or errors, for example, a monthly contact via phone or a measurement that took place while patient was in the hospital.

Dot charts also help to notice *potentially interesting patterns* that in turn help to identify and construct potentially useful features, for instance, weight dynamics. Moreover, it can be clearly observed that a number of patients are measuring themselves during the working days, but not during weekends. That suggests that the use and impact of RPM system is dependent on a lifestyle of the patients. Another strong relation was observed between the frequency of measuring and a contact with medical professionals. For instance, if a patient does not measure herself for some time, a clinical visit or a monthly contact triggers the patient to restart measuring. This suggests that communication increases patient motivation to use the system.

Exploration of time series data shows how the values of daily measurements or symptoms change over time. It also

¹We keep the description of the feature set construction at a high level for proprietary reasons. This is in line with our goal to give a broad perspective and suggest potentially relevant problem formulations.

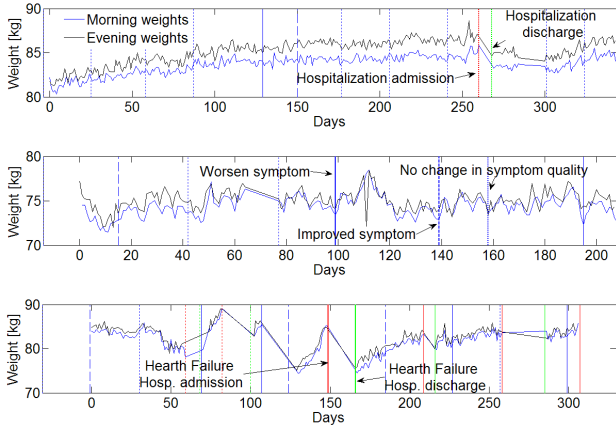


Figure 4. Weight dynamics of three patients.

indicates how these measurements correlate with certain events, such as hospitalization based on particular symptoms. An example of weight timeseries for three patients is shown in Figure 4. Here relations between a rapid weight increase, hospitalization and change of a prominent symptom for HF (ankle swelling) can be observed.

Databases are typically noisy and our case is not an exception. We apply explorative statistical data analysis, outlier detection, data cleaning approaches and handling the missing values to filter *relevant data*.

3.2 Positive and negative instances

Having a collection of relevant data and knowledge about potentially relevant features we can form training instances. This is not trivial due to different periodicity at which the measurements are obtained.

To form the positive training instances for a patient we first find which day HFH has actually occurred (t_h). Then we take a period of 14 days backwards [t_{h-14}, t_h] to compute the features related to daily measurements (D features). Figure 5 illustrates the timeline. It should be noticed that *data* for computing these features may go back further than this two week window. A feature itself might code dynamics of the measurement or exceeding some predefined threshold. An example of such feature could be, how many times the weight exceeded 100 kg in the last three days. The value of the feature for day t_{h-14} will be computed using the data from [t_{h-16}, t_{h-14}]. The medical history (H) and symptom (S) features are computed based on the last available data. Further discussion related to the feature space construction is given in the next section.

A similar approach is used to construct the negative instances. The trick is how to choose a reference time t_h . We set t_h to be an average between the time of two consecutive

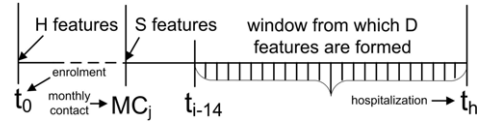


Figure 5. Forming one training instance.

monthly contacts, given that no HFH happened in between. This way we expect to smoothen the effect of how fresh the symptom information is.

It should be noted that for each patient present in the training database several negative instances and potentially more than one positive instance can be constructed. Exact numbers depend on the duration of the observed period for the corresponding patient and on how many times she has been hospitalized during it.

4 Experimental study

We performed a quantitative evaluation of our approach on an extract from the TEN-HMS database [2] containing information about 426 patients with cardiovascular diseases, of whom 143 patients (112 alive and 31 dead) were HF patients home telemonitored during the period of two years and had records at least for 50 days period. 43 patients had at least one HFH.

4.1 Experimental set up

Our experiments had two major goals: to assess the performance of different classification methods and to explore the predictive power of different types of features. We evaluated the results against nine rules, which were established based on domain expertise and are in use in current RPM systems. The rules themselves cannot be disclosed.

The experimental set up consisted of two major steps. First we tested a number of classifiers: support vector machines (SVM), decision trees (J48), and rule-based learners (JRip) [8]. We experimented with different parameter settings on the training data. Then, the selected best classifiers were compared against the triggering rules on the test data.

For each combination of parameters we experimented with different feature subsets: only symptom features (S), symptom and daily measurement features ($S + D$), symptom and medical history features ($S+H$), and their union ($S+D+H$). Additionally, we tried some of these subsets and finally an exhaustive search (FS) for the best feature subset ($S+D+H+FS$). We fixed the best parameters for each classification technique on the training data using ten fold cross-validation. In each category we left only those combinations, which were statistically significantly better than

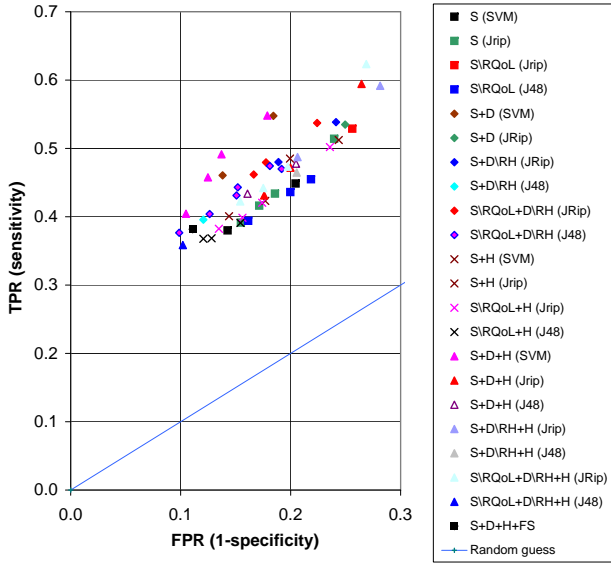


Figure 6. HFH prediction accuracies for different parameterizations and feature sets on the training set (cross-validated).

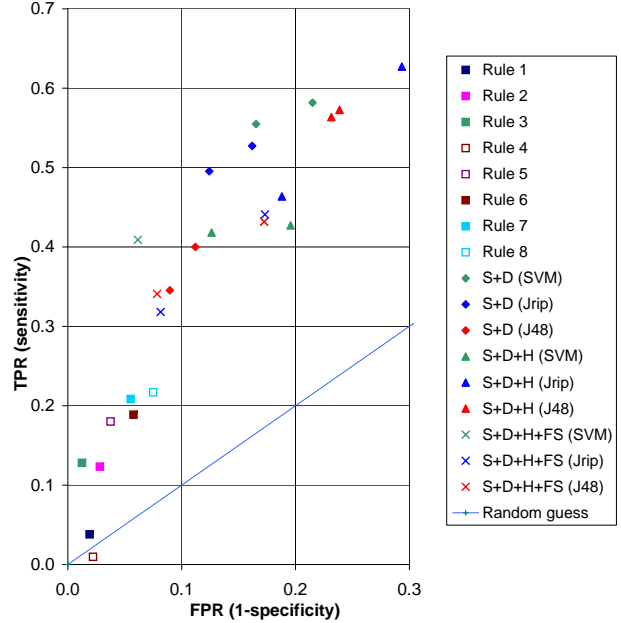


Figure 7. HFH prediction accuracies of the classifiers vs. the expert rules on the test set.

others according to paired t-test with respect to Youden index (YI). The test regards true positive rate (TPR) and false positive rate (FPR) as equally important, $YI = TPR - FPR$.

Finally, we tested the performance of the selected models on the test dataset. To form the test set we held out 29 patients. For each test patient we formed test instances for all monitoring days, the scenario was presented in Figure 2. At day t_i we aimed to predict whether HFH will occur within the next 14 days $[t_{i+1}, t_{i+14}]$. In total there were 220 positive test instances and 9605 negative instances, obtained from 29 patients, which formed the test set.

4.2 Results

We present validation and test results. In Figure 6 the performance of different parameterizations and feature subsets on the training set is presented (cross-validated). The results are plotted against the true positive and false positive rates. Note, that the performance of an ideal classifier would correspond to $TPR = 1$ and $FPR = 0$, the top left corner. The diagonal represents random predictions.

The results suggest that combining the daily measurement (D) and symptom (S) features improves the performance of classification techniques, as compared to only S , the points are closer to the top left corner. Adding the medical history features (H) in many cases improves even further. SVM shows the good results in terms of FPR, while JRip classifier is better in terms of TPR. J48 has shown the

best FPR, but overall it was slightly behind the other two classifiers in terms of Youden index, which combines both FPR and TPR.

The results on the test dataset are plotted in Figure 7 and complemented with Table 2. When interpreting the results, it needs to be taken into account, that the number of correct classifications can be higher than the actual number of hospitalizations. It is due to the problem formulation and the experimental set up. The label is considered to be true if HFH occurred within 14 days. Thus it is also true for the next prediction, but now within 13 days and so on. In this setup one hospitalization generates up to 14 positive instances. Therefore, besides TPR we also report the hospitalization prediction rate (HR) that is how many HFHs have been predicted out of the total number of actual HFHs.

The key result of the experimental study is the following: classification approaches perform much better than individual predefined expert-rules (*Rule1 – Rule8*). The results are consistent in terms of Youden index and hospitalization rate. The relative performance of different classifiers is comparable to their performance on the training data.

4.3 Challenges with symptom features

The results suggest that symptom features S play a key role in the performance of classifiers. We finalize the case study with discussing challenges related to S features.

Classifiers showed rather high false positive rate, which

Table 2. Prediction accuracies on the test set.

Classification model	TPR	FPR	YIndex	HRate	
Rule 1	0.038	0.019	0.019	0.316	
Rule 2	0.123	0.028	0.095	0.211	
Rule 3	0.128	0.012	0.115	0.211	
Rule 4	0.010	0.022	-0.012	0.053	
Rule 5	0.180	0.038	0.143	0.263	
Rule 6	0.189	0.058	0.131	0.316	
Rule 7	0.209	0.055	0.153	0.421	
Rule 8	0.217	0.075	0.142	0.474	
S+D	max TPR	0.582	0.215	0.367	0.526
(SVM)	min FPR	0.555	0.124	0.371	0.474
S+D	max TPR	0.527	0.162	0.365	0.684
(JRip)	min FPR	0.495	0.124	0.371	0.579
S+D	max TPR	0.400	0.112	0.288	0.526
(J48)	min FPR	0.345	0.090	0.256	0.579
S+D+H	max TPR	0.427	0.196	0.232	0.474
(SVM)	min FPR	0.418	0.126	0.292	0.727
S+D+H	max TPR	0.627	0.293	0.334	0.737
(JRip)	min FPR	0.464	0.188	0.276	0.579
S+D+H	max TPR	0.573	0.239	0.334	0.632
(J48)	min FPR	0.564	0.231	0.332	0.632
S+D+H+FS	max TPR	0.432	0.172	0.259	0.632
(SVM)	min FPR	0.409	0.062	0.348	0.579
S+D+H+FS	max TPR	0.441	0.173	0.268	0.632
(JRip)	min FPR	0.318	0.082	0.237	0.368
S+D+H+FS	max TPR	0.432	0.172	0.259	0.632
(J48)	min FPR	0.341	0.078	0.263	0.474

can be attributed to the impact of S features. Indeed, S features by their own allow to predict HFH in many cases. From the domain perspective, symptoms are the early warning signals of worsening, but typically over a longer horizon. S features normally allow to say that there is a high chance of the hospitalization within a month, but not within a short 14 days period. Thus, if classification is based primarily on S features the model can generate a large number of false alarms in a row. To reduce FPR, additional handling mechanisms need to be introduced, which is a subject of further investigation.

Periodicity of S features requires separate attention. The direct measurement of S features may become outdated, due to relatively long intervals between monthly contacts. As a result, a particular symptom might have changed but not yet be recorded. In addition, there might be completely or partially missing values due to the organizational or technical reasons. In such cases, predictive modeling of the symptom features might improve the performance of HFH prediction. We experimented with predicting two most prominent symptoms, breathlessness and swelling of ankles. The results were promising and opened a future research direction.

5 Conclusion

We presented a generic approach for modeling patient state for personalized information and alerting of worsen-

ing. Within the scope of modeling for alerting, we formulated a problem of a short term hospitalization prediction. we presented a data mining approach to predict heart failure hospitalization, with a particular focus of training set construction. An experimental study with the data from a real clinical trial demonstrated the benefits of our approach as compared to expert based prediction. It also opened prospects for further research.

The immediate follow up steps of the work include improving HFH prediction via handling different periodicity of the data. Another step would be to switch from crisp to probabilistic prediction within the prediction horizon, outputting hospitalization probability for each day. In addition, we plan to make use of the educational data, motivational messages and other feedback provided to the patient by an RPM system or medical personnel, to obtain reliable and up-to-date information about the symptoms, and to make our prediction approach context-aware.

Acknowledgments. This research is partly supported by EU HeartCycle and KWR MIP projects.

References

- [1] S. I. Chaudhry, Y. Wang, J. Concato, T. M. Gill, and H. M. Krumholz. Patterns of weight change preceding hospitalization for heart failure. *Circulation*, 116:1549–1554, 2007.
- [2] J. Cleland, A. A. Louis, A. Rigby, U. Janssens, and A. Balk. Noninvasive home telemonitoring for patients with heart failure at high risk of recurrent admission and death. *J. of American College of Cardiology*, 45(10):1654–1664, 2005.
- [3] M. Packer, W. Abraham, and M. Mehra et. al. Utility of impedance cardiography for the identification of short-term risk of clinical decompensation in stable patients with chronic heart failure. *J. of the American College of Cardiology*, 46(11):2245–2252, 2006.
- [4] M. Pechenizkiy, A. Tesanovic, G. Manev, E. Vasilyeva, E. Knutov, S. Verwer, and P. De Bra. Patient condition modeling in remote patient management: Hospitalization prediction. In *Adj. Proc. of 18th Int. Conf. on User Modeling, Adaptation, and Personalization: Posters and Demonstrations*, pages 34–36, 2010.
- [5] A. Tesanovic, G. Manev, M. Pechenizkiy, and E. Vasilyeva. ehealth personalization in the next generation rpm systems. In *Proc. 22nd IEEE Int. Symp. on Computer-Based Medical Systems*, pages 1–8. IEEE Press, 2009.
- [6] R. T. Tsuyuki, R. S. McKelvie, M. O. Arnold, A. Avezum, A. C. P. Barretto, A. C. C. Carvalho, D. L. Isaac, A. D. Kitching, L. S. Piegas, K. K. Teo, and S. Yusuf. Acute precipitants of congestive heart failure exacerbations. *Archives of Internal Medicine*, 161(11):2337–2342, 2001.
- [7] H. Wang. Disease management industry and high-tech adoption. *An Industry report from parks associates, Parks Associates*, 2008.
- [8] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.