Linear Time-Invariant System Based Assessment Model for Coronary Heart Disease

Zening Qu*, Yongqiang Lyu*, Yida Tang†, Wenyao Wang†, Zihan Wang*, Jiaming Hong* and Nazim Agoulmine ‡
*Research Institute of Information Technology, Tsinghua University zeningqu@acm.org, luyq@tsinghua.edu.cn
†Fuwai Hospital tangyida@gmail.com, w.wenyao@hotmail.com
‡IBISC Laboratory, University of Evry nazim.agoulmine@ufrst.univ-evry.fr

Abstract—**This study proposes a linear time-invariant (LTI) system based assessment approach for coronary heart disease (CHD) risk. Unlike traditional risk regression models, the new approach considers accumulated effects of CHD factors with time and can thus perform time-based simulation and real time assessment for the progression of coronary heart disease. There are several LTI-based models achieved in this study via black box system identification process on a 1,549-men cohort. These models have good fitting on the sample data and can adequately reproduce results of the current dominant risk models. Our findings verified that the LTI-based modeling approach works for time-based CHD assessment —- such models can be used for time-based simulation and real time evaluation if sufficient sample data is observed.**

I. Introduction

Coronary heart disease (CHD) is the narrowing or blockage of the coronary arteries, usually caused by atherosclerosis; it is a major cause of mortality in many countries [1]–[3]. Real time CHD risk assessment can improve individuals' awareness of their own coronary artery health condition and take actions in a timely manner. The estimated risk can also help clinicians identifying high-risk patients and comparing effects of different therapies. This is becoming increasingly beneficial given recent advances in e-health and ubiquitous computing: with sensors and microprocessors fabricated into living environments, individuals could be informed of their well-beings before thorough medical examination is performed. A prerequisite of this goal, however, is the availability of an accurate CHD risk assessment model that can be personalized and work with individuals in a real time manner.

Huge effort has been made in assessing CHD risk and identifying risk factors. Traditionally, the fundamental approach is to develop regression models from cohort whose CHD rates was recorded for years, even decades. The most representative work of this approach is the Framingham risk score [4], which estimates the likelihood for an individual to develop CHD in 10 years after the score is calculated. Framingham model and its variations are the most widely used CHD risk models in hospitals today and they have played an important role in informing patients about their coronary health conditions. However, the Framingham score is relatively static, and is thus not a good option for health monitoring in a nonclinical environment, which requires prompt reaction to real time dynamics.

In this study, we propose a linear time-invariant (LTI) system based modeling approach to realize real time CHD risk assessment. We introduce *black box system identification procedure* (see section II) as the fundamental modeling method. Picture a coronary artery system that is developing CHD as a black box, system identification process can generally be employed to reveal mathematically the content of the black box. It is important to note that these models will eventually shed light on the CHD pathology, which makes them valuable triangulations besides modern anatomy and statistics. Moreover, this new approach is able to address the dynamics of CHD progression because it takes in to account the temporal effects of CHD risk factors, which traditional methods ignore. Hence, the result models are very dynamic and cope well with the real time requirement mentioned earlier. Last but not least, system identification models can be personalized to an individual's detailed medical history, which is a favorable feature especially for long-term users.

Besides linear models such as LTI, nonlinear system models are also equally eligible candidates for depicting CHD progression (at least from system identification's perspective). However, nonlinear systems are much more complicated and resource-consuming than linear ones, which makes them less favored choice for ubiquitous computing devices such as mobile phones and standalone sensors. Given that linear systems have already been widely adopted to describe complex systems in other fields such as automated control [5], [6], we prioritize linear models as our primary exploration goal.

To verify the LTI system modeling approach, we performed two experiments and identified a set of models that adequately reproduced results of the currently dominant risk models. The results showed that LTI-based modeling approach works in CHD assessment and the dynamic feature of the models makes it possible to assess CHD progression in a real time fashion.

The rest of this paper is organized as follows. Section II introduces workflow and basic ingredients of system identification. Section III describes our experimental effort in finding the set of models that best describe a coronary artery system's well-being. We discuss limitations of our experiments in section IV and conclude in section V.

II. System Identification

System identification refers to the process of going from observed data to a mathematical model, which is quite fundamental in science and engineering [6]. The model (difference or differential equations) is then a description of the system and can be used for simulation and prediction [5]. In our specific case, we hope to find a model that best describes the

978-1-4673-5801-9/13/$26.00 ©2013 IEEE

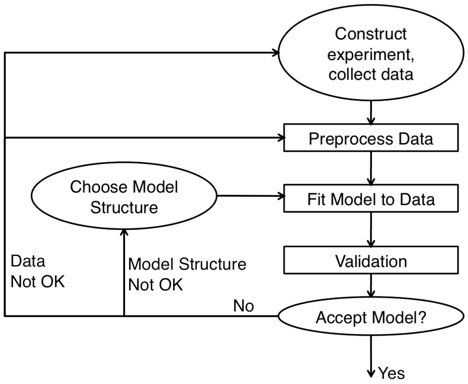


Fig. 1: System identification workflow [6].

human coronary artery system. If that happens we can use the established model to predict (estimate) the likelihood of future CHD occurrences.

System identification is generally an iterative procedure (see figure 1). It is characterized by four ingredients: 1) the observed data, 2) a set of candidate models, 3) a criterion of fit and 4) model validation principles. The observed data usually consists of input and output signals of the target system, sampled/collected over experiment time. Candidate models are in their nature difference or differential equations, with parameters to be estimated. The identification procedure starts with preprocessing the observed data, preparing them for parameter estimation. A model is then chosen from the candidate set and fitted to the data. This is done by running parameter estimation algorithms until the criterion of fit is satisfied or the allocated time or resource is exhausted. The output of this sub-procedure is a model whose parameters are determined. This model then has to go through a validation procedure for its quality to be examined. If, according to the validation principles, the model is good, it can be accepted. Otherwise we should go back to the model selection step, or the data preprocess step, or even the experiment design step and start over.

With many complex systems, such as the human coronary artery system, our knowledge of the target system is not sufficient for us to know the model structure a prior. Fortunately there are a variety of *black box models* that we can try and choose from. Because black box models (e.g. ARX, ARMAX and artificial neural networks) have successfully described a large set of phenomenon in the physical world, when a new relationship needs to be described, it is often rewarding to try the black box models first. On the other hand, if property of the target system is already partially known, we can replace the black box models with *gray box models*, which are capable of combining insights into the model structure. Note that neither black box models nor gray box models should be treated as a "true" description of the system —- they are just fitted to input and output data and are useful for generating more output data. As Ljung puts it, "The acceptance of models depends on the 'usefulness' rather than the 'truth'." [5].

It is not surprising that the set of black box models is a large one. Instead of elaborating each model structure here, we highlight the most general forms of linear and nonlinear systems and introduce models as families.

A linear system can be characterized by Eq. (1) [5], [6]

$$y(t) = G(q, \theta)u(t) + H(q, \theta)e(t) \tag{1}$$

$y(t)$ and $u(t)$ are continuous-time output and input signals of the linear system. $e(t)$ denotes white noise. $G(q, \theta)$ and $H(q, \theta)$ are transfer functions where q is a shift operator and θ is a vector of parameters to be estimated. Parameterizing $G(q, \theta)$ and $H(q, \theta)$ gives a family of linear black box models, among which the ARX, ARMAX, OE and BJ models are the most commonly used.

Another class of linear models are the state space models:

$$\dot{x}(t) = A(\theta)x(t) + B(\theta)u(t) \tag{2}$$

$$y(t) = C(\theta)x(t) + v(t) \tag{3}$$

Here, $x(t)$ is a state vector and $A(\theta)$, $B(\theta)$ and $C(\theta)$ are state space matrices. This model structure can be converted to that of Eq. (1) but is often not done so because the state vector $x(t)$ often carries insight to the system under study.

For nonlinear systems we have a more general model form [6]:

$$\hat{y}(t|\theta) = g(\theta, Z^{t-1}) \tag{4}$$

Here, $\hat{y}(t|\theta)$ is the output predictor. g is a mapping from past observations Z^{t-1} to the next output. g can be parameterized to generate a family of nonlinear models whose popular members include wavelets, neural networks and fuzzy models.

III. Black Box Identification of the Coronary Artery System

We define the coronary artery system as a Multi-Input Single-Output (MISO) system. The single output $y[n]$ is an individual's real time CHD risk. The multiple inputs $u_i[n], i = 1, 2, 3...$ are the individual's various baseline variables that could affect his real time CHD risk. Note that here $y[n]$ and $u_i[n]$ discrete time-series signals, because they are samples of continuously changing data points. Given this, any model who successfully maps $u_i[n]$ to $y[n]$ is a good CHD risk model —- and this model is personalized (u_i and y only denotes one's personal data) and real time (with built-in time indicator n).

In this section we elaborate our two black box experiments by defining inputs (section III-A) and outputs (III-B) and comparing model predictions with our validation data (section III-C). The experiments were based on a 1,549-patient cohort consists of Chinese men aged 41 – 74. Usage of this data was permitted by informed consent from all participants.

A. Input Data

As described above, input signals are known factors that could affect CHD development. Such signals can be CHD risk factors, biomarkers, gene-markers and so on [7]. In our case, we want to choose a set of factors that is large enough to successfully explain output data, and small enough so that factors are relatively independent. For our two experiments, we chose the following factors as input signals: blood pressure, cigarette smoking, total cholesterol (TC), LDL-C, HDL-C, C-reactive protein (CRP), plasma glucose, and triglyceride.

The input signals should cover a reasonably long lifespan of a patient in order to unveil the chronic development of CHD. Ideally, an $u_i[n]$ signal would contain a data point each year and cover a participant's entire adulthood. Unfortunately, no such cohort is immediately available to the authors. As a workaround, we leveraged our cohort data and used age-group means to replace decades of follow-up data. We divided the 1,549 male patients aged 41 – 74 into 34 age groups. For each age group, we computed mean value for each input signal. In this way we obtained dynamic input signals that cover a 34-year lifespan (see figure 2).

The input signals reveal trends across different age groups and they are no substitute of follow-up data. For example, follow-up inputs of a real person would carry temporary information in successive sample points, which is a huge help for the system identification procedure to uncover the chronic pathology of CHD. Temporal changes in age group means, on the other hand, are not so helpful. Despite this fact, we can still use the age group mean input signals to try out the system identification procedure. We can expect to learn more about the pathology significant follow-up data becomes available.

B. Output Data

The output signal $y[n]$ is also a discrete time series signal. Choosing a metric for $y[n]$ is hard because $y[n]$ indicates CHD risk and must be measurable. The choice of $y[n]$ also directly decides the quality of our result model. The later should be expected to perform at *most* as good as $y[n]$. From our experience, the best candidate metric for $y[n]$ is the SYNTAX score [8], which is directly computed from coronary angiographs, the standard reference for CHD diagnose. Unfortunately, because coronary angiography is invasive and thus usually performed on people who already have cardiovascular diseases, it is hard to obtain repeated SYNTAX scores of the same patient, not to mention repeated scores of a healthy person.

Due to the above reason, in our experiments we chose two metrics as model output: the Framingham score [4] and the left ventricle ejection fraction (LVEF). The first metric is a 10 year CHD risk estimation produced by Cox regression functions. The second metric is a measurement of the pumping ability of the heart. We use these metrics as a substitution of the SYNTAX score, assuming that they reflect the health condition of a person's coronary artery system, though in far less accuracy. We plotted the two outputs in figure 3 and found the two metrics to be consistent: the Framingham risk score tend to increase with age (see figure 3a), while the LVEF tend to decrease (see figure 3b). This means as the patient population ages, their CHD risk tend to be higher and their heart pumping capability tend to decay.

1) Details about calculating the Framingham score: Because the patients under discussion are all Chinese males, we used the CMCS recalibrated Framingham function for men [9] to compute the risk score for each age group. The time series Framingham scores depicted in figure 3a was computed using the time series inputs smoking, fasting blood glucose, blood pressure, total cholesterol and HDL-C (figure 2a through 2f). Stratification of blood pressure, total cholesterol and HDL-C was done using the same criteria as that of [9]. Regarding the binary parameters of the Framingham model, our assumption was that imaginary person characterized by age group means does smoke but does not have diabetes, as this was the dominant phenomenon in our cohort. (Among the 1,549 male patients, the smoker v.s. nonsmoker ratio is 1099 v.s. 450, and the diabetes v.s. non-diabetes ratio is 401 v.s. 1148.)

2) Details about the LVEF: In our 1,549-male cohort, 520 males have LVEF record. We processed LVEF in the same way we did to the inputs and the age group mean is plotted in figure 3b. Whether the output signal was the Framingham score or LVEF, the same identification procedure was applied to the corresponding data set. To save space, we do not visualize the input time series of this smaller cohort, but it is not hard to imagine that the variation among age group means becomes larger as there are fewer sample points.

C. Validation

We performed black box identification for both Framingham score and LVEF metrics. Our candidate model set includes: transfer function models, ARX, ARMAX, OE, BJ, state space models, process models, correlation models, nonlinear ARX and the Hammerstein-Wiener model.

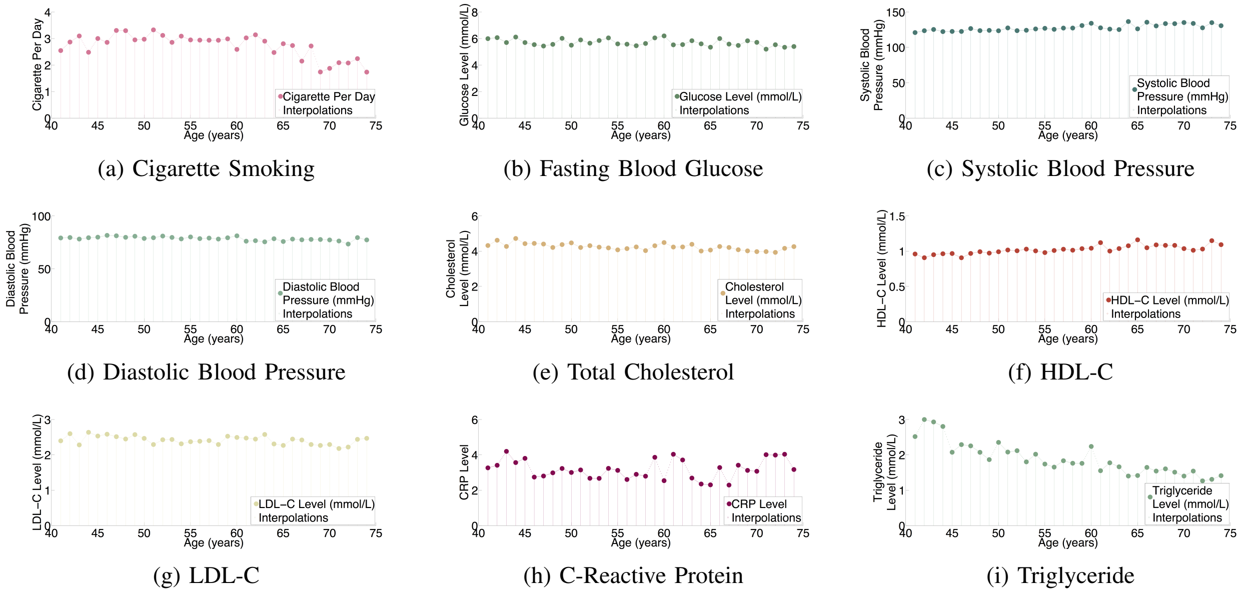
For each input-output combination, we performed black box identification on each of the above model structures. The procedure involves both order selection and parameter estimation. The result models were validated by comparing their prediction outputs with our validation data (which is part of the observed input data). We use a 0 – 100 scale, borrowed from Matlab system identification toolbox, to intuitively evaluate the quality of the result models. The higher the score is, the better a model fits our validation data. Using this score, we were able to rank all result models and pick out models that represent best mappings from the input data to our two output metrics. For our first experiment where Framingham 10 year CHD score was used, we found the four best mappings to be: state space model, Hammerstein-Wienner model, ARX and ARMAX (see figure 4). These models scored 81.08, 80.52, 78.30 and 71.20 in our 0 – 100 score system. In our second experiment where the LVEF value becomes the output signal, the two models that scored over 50 were state space model (scored 91.92) and ARX (scored 72.83) (see figure 5).

From our two experiments it seems that both linear and nonlinear models can adequately map input risk factors to output metrics, which is not uncommon in the field of system identification. As mentioned earlier, a model is not a "true" description but a good mapping of observed data. Note that in both experiments, linear black box models exhibit good quality, even if the output signals are of quite different nature. Such observation is in agreement with the common consent that linear models are quite expressive and covers a large class of phenomena in the physical world.

IV. Discussion

Our current experiments were not designed to build CHD risk model from our cohort. Instead, the purpose was to introduce system identification as an new approach by doing some quick demonstrations. Therefore, the two experiments we have done so far have several limitations.

First, the Framingham score and LVEF are only metrics we chose to quantize the well-being of a coronary system, not



(a) Cigarette Smoking

(b) Fasting Blood Glucose

(c) Systolic Blood Pressure

(d) Diastolic Blood Pressure

(e) Total Cholesterol

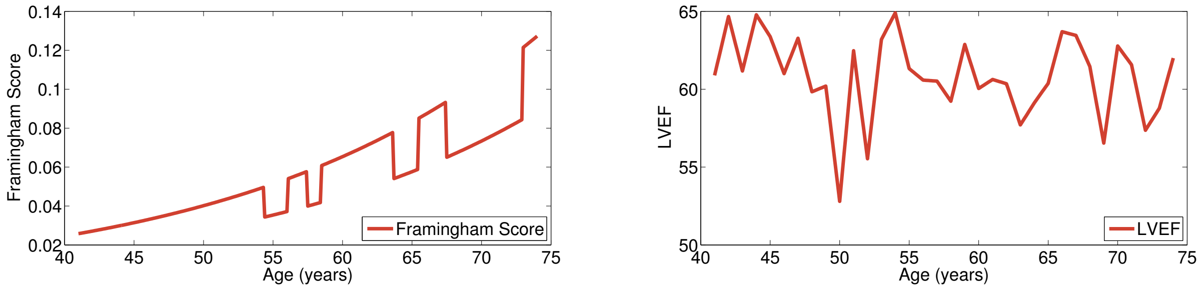
(f) HDL-C

(g) LDL-C

(h) C-Reactive Protein

(i) Triglyceride

Fig. 2: Discrete dynamic input signals computed from age group means. Note that linear interpolation is applied to the signals in order to facilitate identification. Also note that we have already excluded younger (below 41) and elder (above 74) age groups because there were not enough data samples to compute meaningful age group means.



(a) Framingham 10 Year CHD Risk Score

(b) Left Ventricular Ejection Fraction

Fig. 3: Output signals computed using age-group mean inputs.

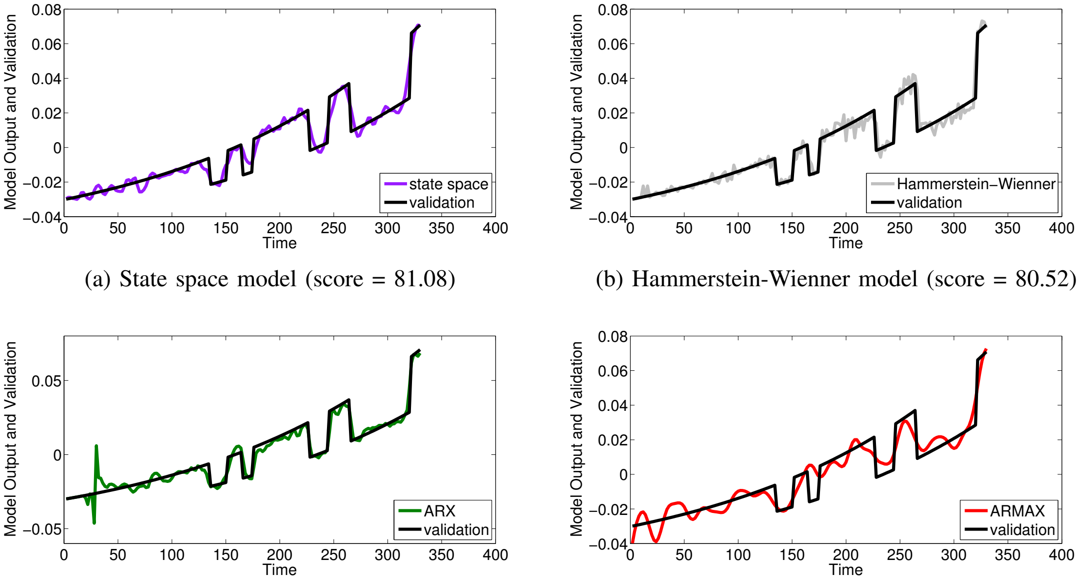
direct quantitative descriptions of the real artery system. The models we identified from these metrics should be expected to be at *most* as accurate as these metrics, if the two metrics were accurate at all. However, the accuracy of the system identification models can readily be improved once higher-quality output data (such as the SYNTAX score mentioned above) becomes available. The exactly same set of techniques can be used to produce models that will eventually outperform traditional models. One might question, however, when new data do become available, would the models be able to express the new relationships? Our answer is yes. In our current experiments, the fact that black box models successfully map risk factors to both Framingham score and LVEF demonstrates the expressiveness of these models. They should also fit the new data as long as it resembles our current two metrics.

The second limitation with our experiments is that we are using a static cohort to substitute what the identification procedure expects: years of follow-up of an individual. The primary drawback of this approach is that we lose the temporal relationship between successive data points in our time series signals. To shed more light on the chronic affects among risk factors, it is necessary to obtain follow-up cohort is model inputs.

The two limitations we have mentioned here also reflect a pragmatic challenge: cohort data that satisfies system identification's specific requirements (dynamic follow-ups that last for years) is hard to obtain.

V. Conclusion

We proposed system identification as a promising alternative to traditional regression models in assessing CHD risk. The major advantages of system identification models are that they can be individualized and used in real time. Through our two preliminary experiments, we have shown that the currently available models are very expressive and fit well with the dominant CHD risk metrics used today. With the support of



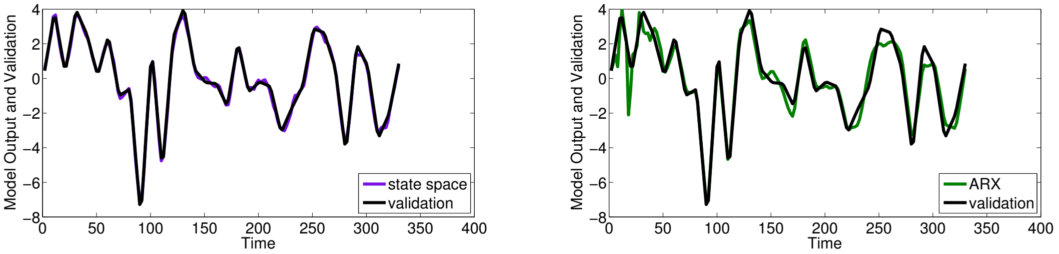
(a) State space model (score = 81.08)

(b) Hammerstein-Wienner model (score = 80.52)

(c) ARX Model (score = 78.3)

(d) ARMAX Model (score = 71.2)

Fig. 4: Four best models for the Framingham case.



(a) State space model (score = 91.92)

(b) ARX model (score = 72.83)

Fig. 5: Four best models for the LVEF case.

more descriptive data, system identification models have great potentials in outperforming models available today.

Acknowledgment

This work is funded by NSFC 61201357. The authors also thank the anonymous reviewers for their feedbacks.

References

[1] A. S. Go, D. Mozaffarian, V. L. Roger, E. J. Benjamin, J. D. Berry, W. B. Borden, D. M. Bravata, S. Dai, E. S. Ford, C. S. Fox, S. Franco, H. J. Fullerton, C. Gillespie, S. M. Hailpern, J. A. Heit, V. J. Howard, M. D. Huffman, B. M. Kissela, S. J. Kittner, D. T. Lackland, J. H. Lichtman, L. D. Lisabeth, D. Magid, G. M. Marcus, A. Marelli, D. B. Matchar, D. K. McGuire, E. R. Mohler, C. S. Moy, M. E. Mussolino, G. Nichol, N. P. Paynter, P. J. Schreiner, P. D. Sorlie, J. Stein, T. N. Turan, S. S. Virani, N. D. Wong, D. Woo, and M. B. Turner, "Heart disease and stroke statistics2013 update: A report from the american heart association," *Circulation*, vol. 127, no. 1, pp. e6–e245, 2013.

[2] X.-H. Zhang, Z. L. Lu, and L. Liu, "Coronary heart disease in china," *Heart*, vol. 94, no. 9, pp. 1126–1131, 2008. [Online]. Available: http://heart.bmj.com/content/94/9/1126.abstract

[3] S. Allender, P. Scarborough, V. Peto, M. Rayner, J. Leal, R. Luengo-Fernandez, and A. Gray, "European cardiovascular disease statistics," *European Heart Network*, vol. 3, pp. 11–35, 2008.

[4] P. W. Wilson, R. B. D'Agostino, D. Levy, A. M. Belanger, H. Silbershatz, and W. B. Kannel, "Prediction of coronary heart disease using risk factor categories," *Circulation*, vol. 97, no. 18, pp. 1837–1847, 1998.

[5] L. Ljung, *System Identification – Theory for the User, Second Edition*. Prentice Hall PTR, 1999.

[6] ——, *System Identification*. John Wiley & Sons, Inc., 1999.

[7] T. H. S. Dent, "Predicting the risk of coronary heart disease i. the use of conventional risk markers," *Atherosclerosis*, no. 2, pp. 345–51, 2010.

[8] G. Sianos, M.-A. Morel, A. P. Kappetein, M.-C. Morice, A. Colombo, K. Dawkins, M. van den Brand, N. Van Dyck, M. E. Russell, F. W. Mohr *et al.*, "The syntax score: an angiographic tool grading the complexity of coronary artery disease," *EuroIntervention*, vol. 1, no. 2, pp. 219–227, 2005.

[9] J. Liu, H. Yuling, R. B. D'Agostino, Z. Wu, W. Wang, J. Sun, P. W. F. Wilson, , W. B. Kannel, and D. Zhao, "Predictive value for the chinese population of the framingham chd risk assessment tool compared with the chinese multi-provincial cohort study," *JAMA*, vol. 291, no. 21, pp. 2591–2599, 2004. [Online]. Available: + http://dx.doi.org/10.1001/jama.291.21.2591