# Predicting Age Using Biomarkers and Physiological Measurements

Huy Seng
Biomedical Computation
Stanford University
Email: huyseng@stanford.edu

December 11, 2009

## 1 Introduction

The general consensus is that if you want to know someone's age, you would just ask that person. However, some unscrupulous people might lie about their age and even falsify birth records for personal gain. This problem is somewhat pervasive in professional sports today. During the 2008 Summer Olympics in Beijing, there was a huge controversy over the ages of the female Chinese gymnasts. Professional baseball prospects in other countries falsify their birth certificates to appear younger than they really are in order to seem more attractive to MLB scouts. There is a lack of simple tests or techniques that could definitively verify someones age. The aging process is still not well understood. To date, there are no accurate predictive models of biological aging in adult human populations.

## 2 Background

Currently, there is no widely accepted predictive model for human biological aging. Some researchers have used biomarker data to build such a model with moderate success[1][2][3]. Fliss et al. used age, gender, and blood biomarker information to build a model for aging using statistical techniques on the National Health and Nutrition Examination Survey (NHANES) data [1]. Chen et al. used Least Angle Regression to reduce the number of features in their model of aging in pediatric patients [2].

## 3 Methods

### 3.1 Data

Data from NHANES was used to test and validate the aging models. The National Health and Nutrition Examination Survey (NHANES) is a series of interviews and physical examinations that help evaluate the health and nutritional status of both children and adults in the United States. A part of the Centers for Disease Control and Prevention (CDC), NHANES surveys a sample of 5000 people that tries to accurately represent the nations demographic. The survey process first includes an interview with demographic, socioeconomic, dietary, and health-related questions. That information is then supplemented by a physical examination and laboratory tests to attain medical, dental, and physiological measurements. NHANES data from 2001-2002 was used. People aged 10-17 were classified as young, and people aged 18-25 were classified as old. Any data with missing elements was thrown out to avoid data imputation.

### 3.2 Model Selection

Weka (Waikato Environment for Knowledge Analysis), a machine learning software that features an abundance of algorithms, was used for the preliminary data analysis. Weka's user-friendly GUI and relative ease of use was optimal for experimenting and fast prototyping. Weka was used to visualize the NHANES data set first. Then, many of the machine

learning algorithms available in Weka were used on the NHANES 2001-2002 data set and each algorithm was evaluated using 10-fold cross-validation. Some of the classifiers that were experimented with were naive Bayes, support vector machines, logistic regression, AdaBoost, and random forests. The classifier with the highest predictive accuracy was chosen. In this case, accuracy is defined as the number of correctly classified instances in the data set. This method was used to select the classifiers for both the male and female models.

### 3.3 Feature Selection

There were initially 45 biomarkers in the feature set. It is inconvenient to use so many biomarkers to predict age. Thus, feature selection was used to reduce the feature set to a more manageable size. Information gain was the metric used to find the best features.

## 4 Results

### 4.1 Model Selection

For the male models bagging performed the best with 88.06% accuracy and an AUC of 0.948, but logistic regression was ultimately chosen since it performed better on the smaller feature set. Logistic regression was also chosen for the female model since it had the highest accuracy and AUC. The accuracy was 81.29%, and the AUC was 0.906. [Tables 1- 2]

### 4.2 Feature Selection

The top ten features are displayed in tables 3 and 4 for males and females respectively. The bolded features are the one chosen for the final model. The accuracy was calculated with the top ten features. Then, the lowest ranking feature was dropped, and the accuracy was calculated again. This was repeated until only one feature was left. The feature set with the highest accuracy was chosen. The male model performed with 88.06% accuracy with only 3 biomarkers, while the female model performed with 81.13% accuracy with 9 biomarkers. The male model actually performed better with only 3 biomarkers than with all 45 features. [Tables 5-6]

## 5 Discussion

The male and female logistic regression models both performed surprisingly well. Remarkably, the male model was able to classify age with only 3 biomarkers with 88% accuracy. The top biomarker for each model was alkaline phosphatase, which is present in children in higher levels. Height, weight, and body mass index also showed up in the top 10 features according to information gain for males. This makes sense since height, weight, and bmi tend to correlate with age, especially in children. The male model probably performed better than the female model due to the cutoff point for the two age groups. Males generally end adolescence and enter young adulthood around 18, while females usually complete puberty before 18. The female model might perform better if the age cutoff were changed to 16. These models are both pretty accurate and sensible. They provide us with some insight into the biological changes that take place from adolescence to young adulthood.

## References


[1] Fliss, A., M. Ragolsky, and E. Rubin, Reverse Translational Bioinformatics: A Bioinformatics Assay Of Age, Gender And Clinical Biomarkers, in 2008 Summit on Translational Bioinformatics, A. Butte, Editor. 2008: San Francisco, CA.

[2] Chen, D., A. Morgan, and A. Butte, Validating pathophysiological models of aging using clinical electronic medical records. J Biomed Inform. 2009 Nov 30..

[3] Nakamura, E. and K. Miyao. A method for identifying biomarkers of aging and constructing an index of biological age in humans. Journal of Gerontology, 2007. 62A(10):p. 1096.

| Classifier | Accuracy | AUC | Precision | Recall |
|---|---|---|---|---|
| Bagging | 88.06% | 0.948 | 0.883 | 0.881 |
| SMO | 88.06% | 0.89 | 0.893 | 0.881 |
| Logistic Regression | 87.43% | 0.953 | 0.877 | 0.874 |
| AdaBoost | 87.08% | 0.939 | 0.874 | 0.871 |
| Multilayer Perceptron | 86.19% | 0.94 | 0.862 | 0.862 |
| Random Forest | 84.94% | 0.918 | 0.849 | 0.849 |
| Naive Bayes | 83.69% | 0.907 | 0.843 | 0.837 |
| Bayes Net | 82.98% | 0.92 | 0.845 | 0.83 |

Table 1: 10-fold cross-validation results of different classifiers on the male data set.

| Classifier | Accuracy | AUC | Precision | Recall |
|---|---|---|---|---|
| Logistic Regression | 81.29% | 0.906 | 0.813 | 0.813 |
| SMO | 80.31% | 0.796 | 0.803 | 0.803 |
| Bagging | 79.57% | 0.878 | 0.796 | 0.796 |
| Multilayer Perceptron | 78.67% | 0.875 | 0.787 | 0.787 |
| Random Forest | 77.28% | 0.865 | 0.774 | 0.773 |
| AdaBoost | 76.78% | 0.86 | 0.767 | 0.768 |
| Bayes Net | 75.31% | 0.854 | 0.753 | 0.753 |
| Naive Bayes | 73.17% | 0.834 | 0.734 | 0.732 |

Table 2: 10-fold cross-validation results of different classifiers on the female data set.

| Rank | Biomarker | Score |
|---|---|---|
| **1** | **Alkaline Phosphatase** | **0.60513** |
| **2** | **Creatinine** | **0.25126** |
| **3** | **Hemoglobin** | **0.18479** |
| 4 | Hematocrit | 0.18285 |
| 5 | LDH | 0.13646 |
| 6 | Weight | 0.13327 |
| 7 | Height | 0.13144 |
| 8 | Body Mass Index | 0.07499 |
| 9 | Mean Cell Volume | 0.07268 |
| 10 | Red Cell Count | 0.06181 |

Table 3: Top 10 biomarkers according to information gain for male model.

| Rank | Biomarker | Score |
|---|---|---|
| **1** | **Alkaline Phosphatase** | **0.28177** |
| **2** | **Total Calcium** | **0.072** |
| **3** | **Segmented Neutrophils** | **0.06982** |
| **4** | **Weight** | **0.06788** |
| **5** | **Cholesterol** | **0.0661** |
| **6** | **Lymphocyte Percent** | **0.0598** |
| **7** | **Body Mass Index** | **0.0515** |
| **8** | **Triglycerides** | **0.04978** |
| **9** | **GGT** | **0.03996** |
| 10 | LDH | 0.03511 |

Table 4: Top 10 biomarkers according to information gain for female model.

| Features | Accuracy |
|---|---|
| All 45 | 87.43% |
| Top 10 | 87.70% |
| Top 9 | 87.79% |
| Top 8 | 87.43% |
| Top 7 | 87.52% |
| Top 6 | 87.70% |
| Top 5 | 87.79% |
| Top 4 | 87.97% |
| **Top 3** | **88.06%** |
| Top 2 | 87.43% |
| Top 1 | 87.43% |

Table 5: 10-fold cross-validation accuracies for feature sets consisting of the top biomarkers by information gain in the male model.

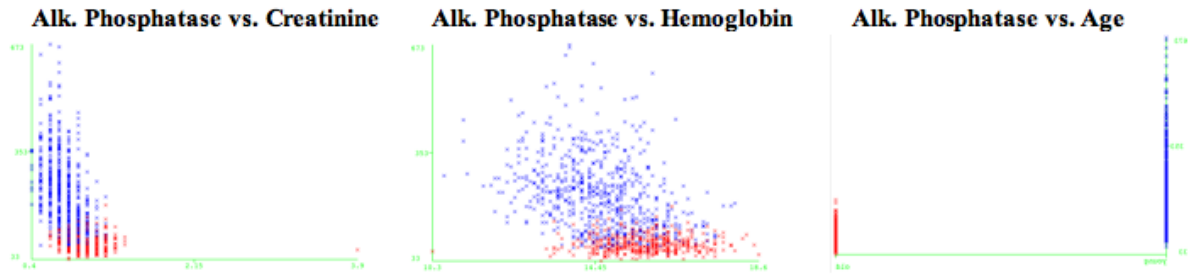


Figure 1: These graphs show weak correlations between the top biomarker and the rest of the biomarkers in the male model

| Features | Accuracy |
|---|---|
| All 45 | 81.29% |
| Top 10 | 81.13% |
| **Top 9** | **81.13%** |
| Top 8 | 80.64% |
| Top 7 | 79.57% |
| Top 6 | 79.08% |
| Top 5 | 79.25% |
| Top 4 | 77.77% |
| Top 3 | 77.11% |
| Top 2 | 75.72% |
| Top 1 | 75.47% |

Table 6: 10-fold cross-validation accuracies for feature sets consisting of the top biomarkers by information gain in the female model.

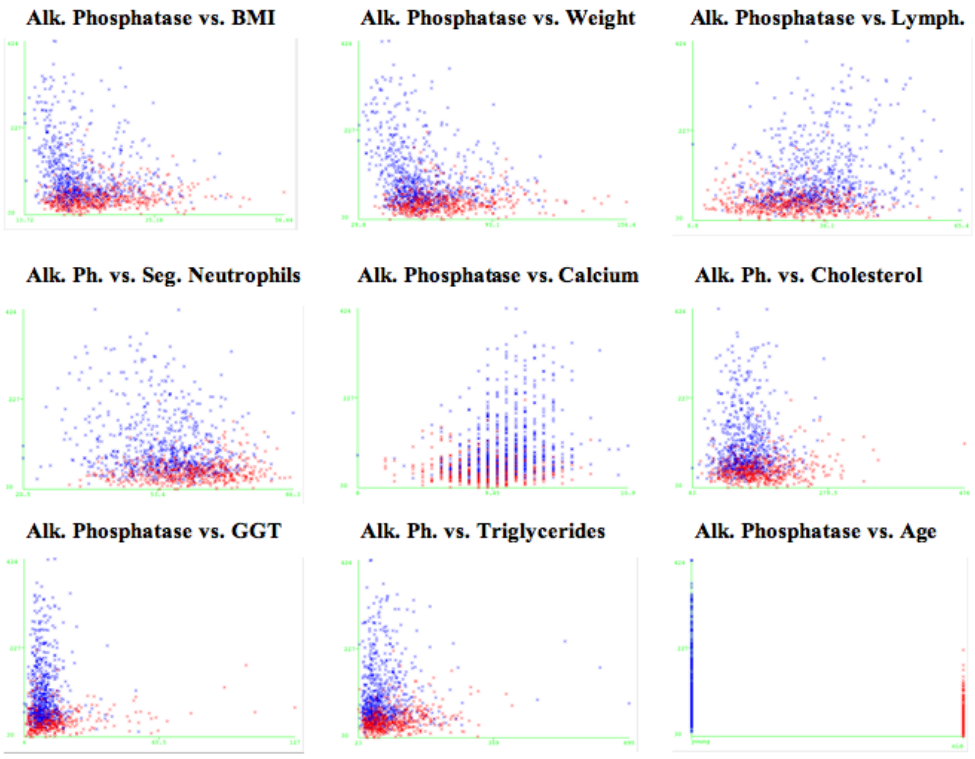


Figure 2: These graphs show weak correlations between the top biomarker and the rest of the biomarkers in the female model