What We Found on Our Way to Building a Classifier: A Critical Analysis of the AHA Screening Questionnaire

Quazi Abidur Rahman (MSc) QUAZI@CS.QUEENSU.CA
Computational Biology and Machine Learning Lab, School of Computing, Queen's University, Kingston, ON
Sivajothi Kanagalingam (MD), Aurelio Pinheiro (MD), Theodore Abraham (MD)
Heart and Vascular Institute, Johns Hopkins University, Baltimore, MD
Hagit Shatkay (PhD) SHATKAY@CIS.UDEL.EDU
Computational Biology and Machine Learning Lab, School of Computing, Queen's University, Kingston, ON
Dept. of Computer and Information Sciences & Center for Bioinformatics and Computational Biology, University Of Delaware, Newark, DE

Abstract

The American Heart Association has recommended a 12-element questionnaire for pre-participation screening of athletes (before they take on athletic activities), in order to reduce and hopefully prevent sudden cardiac death in young athletes. This screening procedure is widely used throughout the United States, but its efficacy for discriminating *Normal* from *Non-normal* heart condition is not clear. As part of a larger study on cardiovascular disorders in young athletes, we set out aiming to pursue a classification task: namely, training a machine-learning-based classifier to automatically categorize athletes into risk-levels based on their respective answers to the AHA questionnaire. We also conducted information-based analysis of each question to identify the ones that may best predict the athletes' heart condition and thus enhance the classification performance. However, surprisingly, rather than obtaining a classifier, the classification results, the information contents of the questions, as well as further probabilistic analysis, all indicate that the AHA-recommended 12 elements screening procedure does not effectively distinguish between *Normal* and *Non-normal* heart as identified by cardiologists using Electro- and Echo-cardiogram examinations. Our results suggest that ECG and (possibly Echo) rather than the questionnaire should be considered for screening young athletes.

1. Introduction

Inherited cardiovascular disease is the main cause of sudden cardiac death (SCD) in young athletes. In the United States the incidence has been reported as 1:50,000 – 1:100,000 per year (Corrado et al., 2005; Maron, 2003; Piggozi & Rizzo, 2008). A larger study in the Veneto region in Italy reported an incidence rate of SCD of 2.1 per 100,000 athletes annually as a result of cardiovascular disease (Corrado et al., 2005). While the incidence of SCD is lower in comparison to other causes of death, it is disconcerting in that these deaths occur in young and otherwise perceived-to-be healthy individuals, most often without any prior cardiac symptoms. Moreover, as most of these deaths occur in athletes of high-school age (Corrado et al., 2005, Wever-Pinzon et al., 2009) they are a cause for much concern in the media, the public and the medical community.

Initial screening through electrocardiogram (ECG) and echocardiogram (Echo) is a first step for identifying morphological anomalies that can lead to cardiac abnormalities, and in extreme cases to sudden death. However, due to considerations involving speed, ease of administration and cost, these standard procedures, while often used in Europe (Corrado et al., 1998) are not used for large-scale screening of young athletes in the United States. As an alternative preventive measure, the American Heart Association (AHA) has recommended a screening procedure (Maron et al., 1996), intended as a cost-effective, practical initial measure for pre-participation screening of athletes. In the United States, the use of this screening procedure has steadily increased over the years since 1997 (Glover & Maron, 2007).

The current, revised, AHA pre-participation screening recommendations were published in 2007, and include 12-element screening guidelines (Maron, 2007) (see Table 1). Under these guidelines, each athlete answers several questions concerning personal and family history and undergoes a physical examination (we refer to the *combination* of *questions and physical exam* as *the questionnaire*). If any of the questions is answered in the affirmative or if the physical examination suggests an abnormality, the athlete is then referred for a more extensive cardiologic evaluation through ECG and Echo, in which responses that are *Non-normal* (i.e., deviate from the *Normal* measures established for athletes, but not conclusively abnormal) can be identified; athletes with *Non-normal* results are referred for further, more extensive and more accurate testing to verify whether any serious heart condition is present. A preliminary study by our group (Kanagalingam, 2010) (presented as an abstract at the AHA symposium), has broadly suggested low predictive power of the AHA screening procedure, without considering its explicit elements and their predictive value.

As a component within a large-scale research of adverse heart conditions, which examines multiple biological and

clinical factors and extensively studies the efficacy of the questionnaire and its possible contribution to predicting cardiac irregularities, we set out to pursue what appeared to be a straightforward task: namely, training a machine-learning-based classifier, based on the answers to the questionnaire from several hundred athletes, in order to automatically predict from these answers the athletes' heart condition. The "heart condition" for the purpose of this study was either *Normal* or *Non-normal*, as determined by a cardiologist based on ECG and Echo readings. We expected to be able to effectively train such a classifier from the questionnaire data, due to the hypothesis driving the AHA guidelines as discussed above: namely, that the answers to the pre-screening questionnaire can indeed be correlated with the diagnosis obtained from the more extensive and time-consuming standard initial cardiovascular tests, (Echo and ECG), administered by a physician. Intending to follow the common machine-learning procedures for learning a classifier from data (e.g., Mitchell, 1997) we also aimed to select the most informative features, that is, identify the items in the AHA-based pre-screening procedure, whose answers are the most predictive of the cardiologist's adjudication.

Machine learning methods have been widely used for disease prediction, risk assessment and patient classification. For instance, in the field of cardiology, arrhythmia classification was performed using support vector machines (Melgani & Bazi, 2008; Osowski et al., 2004), linear discriminant analysis (Chazal & Reilly, 2006) and artificial neural networks (Yu & Chou, 2008). As another example, naïve Bayes classifiers have been used for diagnosis and risk assessment of Long-QT syndrome in children from clinical data (Qu et al., 2010). In the area of cancer diagnosis and prediction, methods such as support vector machines (Akay, 2009), logistic regression (Chhatwal et al., 2009) and random forests (Statnikov & Wang, 2008) have been applied. We thus anticipated that by using filled-in questionnaires from a relatively large population of young athletes, we could train a classifier to distinguish between athletes with potential cardiovascular abnormalities (as determined by ECG and Echo tests) from normal ones.

For the work described here, data was collected from 470 athletes, participating at state-level athletic meets, who have been screened through the procedure using the AHA guidelines, consisting of questions and a basic physical examination. These same athletes have also undergone more extensive tests through ECG and Echocardiograms. The latter two tests were reviewed by an experienced electrophysiology cardiologist, who used these two tests to adjudicate each athlete as *Normal* or *Non-normal.* The cardiologist's adjudication, which is based solely on ECG and Echocardiograms, serves here as the "gold-standard" to which the AHA guidelines results are compared.

Notably, the screening through the AHA procedure is intended as a means to avoid the more costly and cumbersome Echo and ECG tests, Thus the underlying assumption in administering the AHA procedure is that athletes who require further screening (those whose ECG or Echo would thus not be completely *Normal*) would indeed be identified in the screening and referred for further examination (ECG, Echo - and if needed even more extensive testing), while athletes who do not need further screening would have their questions and basic physical show completely *normal* answers.

We also note that *ideally* an evaluation of the AHA questionnaire effectiveness may directly attempt to correlate *actual sudden death* events with specific questionnaire answers. However, such a study is impractical, fortunately, due to the relatively low incidence of actual sudden cardiac death (SCD). That said, we also point out that the questionnaire does not intend to "magically" predict SCD. Rather, the reasoning behind it is that it should suggest (or rule-out) the presence of certain morphological anomalies that can lead to sudden death. Such anomalies are best assessed by a cardiologist through the analysis of the ECG and the Echo tests. Based on this insight, the expectation was that the answers to the questionnaire should be predictive of the Echo/ECG results. As such, our goal was to train a machine-learning-based classifier that will take as input the results obtained from the screening based on the 12-element AHA guidelines for each athlete and predict the cardiologist's Echo/ECG-based adjudication.

In this study we rigorously apply classification techniques and investigate the information-content of each item in the questionnaire. We also conduct probabilistic analysis of the positive and negative answers and their correlation with ECG/Echo test results. However, the classification results and the information contents of the different items, as well as the results from the probabilistic analysis, exposed significant shortcomings in the pre-screening procedure itself. Thus, what started as a classification task, ended up as an in-depth informatics-driven analysis, indicating and revealing important issues with the AHA screening procedure, whose use is advocated as the primary screening tool for athletes.

While the article begins by discussing what appears to be a negative result, its main contribution and the significance of the presented research lies in employing the same statistical, information-based methods that are typically used for developing diagnostic/predictive machine-learning tools, to effectively expose important shortcomings in the current screening procedure. It also points out that other, more discerning, procedures may be required for effective pre-participation screening of athletes (at least until a questionnaire is devised with better predictive capability). Hence, our results suggest that ECG and (possibly Echo) should be considered for screening athletes in the United States. We note that ECG is being used for screening of athletes in Europe, especially in Italy (Corrado et al., 1998) and has been recommended by the consensus statement of European Society of Cardiology (Corrado et al., 2005).

Table 1. The AHA 12-element Screening Guidelines (Maron, 2007).

Guideline #	Question Type	Question Contents as described in the AHA guideline
1	Personal History	Exertional chest pain/discomfort?
2		Unexplained syncope/near-syncope?
3		Excessive exertional and unexplained dyspnea/fatigue, associated with exercise?
4		Prior recognition of a heart murmur?
5		Elevated systemic blood pressure ?
6	Family History	Premature death (sudden and unexpected, or otherwise) before age 50 years due to heart disease, in at least one relative?
7		Disability from heart disease in a close relative younger than 50 years of age?
8		Specific knowledge of certain cardiac conditions in family members: hypertrophic or dilated cardiomyopathy, long-QT syndrome or other ion channelopathies, Marfan syndrome, or clinically important arrhythmias?
9	Physical Exam	Heart murmur
10		Femoral pulses to exclude aortic coarctation
11		Physical stigmata of Marfan syndrome
12		Brachial artery blood pressure (sitting position)

Throughout the rest of the paper we describe the AHA-based questionnaire data, the analysis applied, and the operative conclusions, suggesting that the questionnaire is not an effective tool for assessing risk in young athletes, and that alternative procedures need to be considered.

2. Data

The study included 470 participants, all of whom are young athletes participating at state level athletic meets. They were all asked to fill a questionnaire consisting of 12 *Yes/No* questions as shown in Table 2 (*Q1*-*Q12*), corresponding to AHA elements 1-8 shown in Table 1. They have also undergone a physical exam corresponding to the AHA elements 9-12 in Table 1. The results of the physical (which can either be *normal* or *abnormal*), are listed as Question 13 (*Q13*) in Table 2. Notably, the AHA 12-elements are intended to be clear to physicians but not necessarily to laymen. Therefore, the questionnaire filled by the athletes, as shown in Table 2, uses simpler questions that correspond to each element's intention. In several cases more than one question is needed to cover an element, and some questions address more than a single element. The element number(s) covered by each question is shown in the rightmost column of Table 2.

In addition to answering questions *Q1*-*Q12* and undergoing the basic physical (*Q13*), the participants have separately undergone ECG and Echo tests. The latter two tests were evaluated by an expert cardiologist to draw a more conclusive adjudication regarding each individual's heart condition, based on measurable, observable cardiac parameters as opposed to questions. The two possible conclusions were: *Normal* and *Non-normal,* where *Non-normal* heart condition means that further extensive cardiological evaluation of the athlete is required. While the cardiologist was not officially blinded to the results from the questionnaire, his adjudication was based solely on the ECG and Echo tests, and did not include any analysis or consideration of the questionnaire results. Of the 470 participants, 348 were categorized by the cardiologist as *Normal,* while 122 were categorized as *Non-normal*.

Table 2. The list of questions used in the questionnaire presented to the athletes in this study, along with the AHA guideline number to which each question corresponds. Question 13 summarizes the results of the Physical part of the AHA 12-elements guidelines.

Quest. #	Question content as presented to athlete	AHA Guideline #
Q1	Dizziness/Passed Out during/after exercise?	2
Q2	Chest Pains or shortness of breath?	1
Q3	Become tired quicker than peers during exercise?	3
Q4	Heart murmur/disease?	4
Q5	Skipped heartbeats or racing heartbeats?	1 (discomfort), 4
Q6	Heart disease development or related death in family?	6
Q7	Does anyone in the family have fainting episodes or seizures?	6,7
Q8	Chest discomfort when active?	1
Q9	Have you been told you have high blood pressure?	5
Q10	Have you experiences seizures or exercise related asthma?	1,2
Q11	Anyone in family experienced heart surgery or have a pacemaker or defibrillator under the age of 50 years?	7
Q12	Anyone in family diagnosed with Cardiomyopathy, aneurysm, Marfan's, IHSS?	8
Q13	Physical examination results ab*normal*?	9-12

As not all participants answered all the questions, when analyzing individual questions for information content and conditional probabilities (Sections 3.2 and 3.3), we consider, per-question, only the number of answers that the question has actually received. In Section 3.1, we describe how the missing values are handled by the classifiers. The second row in Table 3 shows how many answers were received for each of the questions, while the third and fourth rows indicate how many of the answers were positive and how many of them were negative, respectively.

3. Methods and Tools

Our analysis of the AHA questionnaire data started by applying classifiers to the data, and was followed by an information-content analysis of each question. We also performed probabilistic analysis of the answers to each

Table 3. Number of answers received for each question along with the number of positive and negative answers.

	Q1	*Q2*	*Q3*	*Q4*	*Q5*	*Q6*	*Q7*	*Q8*	*Q9*	*Q10*	*Q11*	*Q12*	*Q13*
# of answers	469	466	466	436	431	380	423	466	440	468	459	367	451
# of positive answers	94	121	51	33	22	40	45	55	26	65	6	12	40
# of negative answers	375	345	415	403	409	340	378	411	414	403	453	355	411

question, and investigated the possibility of identifying subsets of questions that may together show a stronger association with abnormal outcomes than individual questions do. These methods and related tools are presented in the following subsections.

3.1 The Classifiers

As a baseline for examining the feasibility of predicting the heart condition of young athletes using the AHA questions and physical examination as attributes, we applied three standard classification methods: naïve Bayes (e.g., Mitchell, 1997), random forests (Breiman, 2001) and support vector machine (Cortes & Vapnik, 2005).

We used the standard classification packages in WEKA (Hall et al., 2009) for all three classifiers. Random forests was implemented with 100 trees. SVM used Gaussian radial basis function as kernel[1], where the soft margin parameter C and the kernel parameter γ were selected after trying several combinations of the parameters and choosing the best one in terms of overall accuracy. To train/test and evaluate the performance of the classifiers, we used the standard 10-fold cross-validation procedure.

As not all participants answered all the questions, some values are missing in the questionnaires, as shown in Table 3. For classification purposes, we denote each missing value as *Not Known (NK)*. Hence, each athlete's response to the questionnaire is represented as a 13-dimensional vector $(a_1, a_2, a_3, \dots, a_{13})$, where $a_i \in \{No, Yes, NK\}$, denoting a negative, a positive or a *Not Known* answer, respectively, to question Q_i. The task of the classifier is to assign each such instance (athlete) into one of the two possible classes: *Normal* or *Non-normal*. For the purpose of this study, the true class for each of the 470 athletes is as assigned by the cardiologist based on the results of the ECG and Echo tests (348 have *Normal* conclusion and 122 have *Non-normal* conclusion). As the dataset is biased toward the *Normal* class, to correct for the imbalance, we used the standard procedure of sub-sampling from the over-represented class to create a balanced dataset for training/testing. Under the sub-sampling method, instances are chosen at random from the majority class to make the size of the two classes equal (Clement et al., 2009). By randomly selecting 122 instances from the *Normal* class and taking the whole subset of 122 *Non-normal* instances we obtain a balanced dataset. We have repeated the sub-sampling procedure 5 times to ensure stability of the results. The classifiers have been trained and tested on both the original and the balanced dataset.

To measure the performance of the classifiers, we have used several standard measures, namely, the *Accuracy*, that is, the proportion of correctly classified instances, as well as the widely used measures of *Recall (Sensitivity) Precision (counterpart of Specificity),* and *F-measure*. *Accuracy*, *Precision* and *Recall* are defined below, where *true positives*, denote *Non-normal* cases that are correctly classified as *Non-normal*:

$$Accuracy = \frac{\#\ of\ correctly\ classified\ instances}{Total\ number\ of\ instances};$$

$$Precision = \frac{\#\ of\ true\ positives}{\#\ of\ true\ positives + \#\ of\ false\ positives};$$

$$Recall = \frac{\#\ of\ true\ positives}{\#\ of\ true\ positives + \#\ of\ false\ negatives}.$$

The *F-measure* is the harmonic mean of the *Precision* and the *Recall*. The definition of the *F-measure* is:

$$F - measure = 2 \cdot \frac{Precision.Recall}{Precision + Recall}.$$

An alternative approach to handle imbalanced data is to use cost-sensitive classification, for instance by explicitly penalizing misclassification of *Non-normal* items into the *Normal* class. We have conducted experiments using this approach, but do not discuss them here due to space limits.

3.2 Information Content Analysis

As discussed in more detail in Section 4, using all the questions as attributes results in poor classification performance. Hence we focused on investigating each question individually to assess its predictive capability. To measure each question's predictive capability, we use the well-known Information Gain criterion (e.g., Mitchell, 1997). This criterion is typically used in algorithms for learning tree-based classifiers from data. However, we note that poor performance of tree-based classifiers does not necessarily imply low information gain for the attributes. After the first attribute has been selected as the root of a classification tree, as we proceed further down the tree, less data is available to calculate the information gain of an attribute. This may lead to selection of an attribute that appears to be highly informative, but eventually contributes toward poor classification performance. For this reason, separately calculating each attribute's information gain using the whole dataset is important.

[1] We have also tried linear kernel, but Gaussian radial basis kernel performed marginally better than the linear kernel.

The information gain, calculated for each question, measures how much information is gained about the conclusion (*Normal* or *Non-normal*) when the answer to that question is obtained. It thus indicates how predictive the answer to a question is in classifying participants as having a *Normal* or a *Non-normal* heart-condition. It is calculated as the difference between the unconditional entropy associated with the conclusion and the conditional entropy of the conclusion given the answer to a question. These measures are formally defined as follows: Let C be the set of conclusions (class labels) and A_Q be the answer to question Q. The maximum likelihood estimate for the probability of the conclusion being *Normal (Nor)*, $Pr(C = Nor)$, is calculated as:

$$Pr(C = Nor) \approx \frac{\#\ of\ participants\ with\ Normal\ conclusion}{Total\ \#\ of\ participants},$$

while the probability of *Non-normal (NNor)* conclusion is simply calculated as

$$Pr(C = NNor) = 1 - Pr(C = Nor).$$

Similarly, we define the conditional probability of the conclusion to be *Normal* (or *Non-normal*) given the answer (*Yes* or *No*) to question Q. We define this probability, for a question Q, as: $Pr(C = W|A_Q = X)$ where W is either *Nor* or *NNor* and X is either *Yes* or *No*. The conditional probabilities are estimated from the observed proportions; e.g., the probability of the conclusion being *Non-normal* given that the answer for question Q is positive, $Pr(C = NNor|A_Q = Yes)$ is estimated as:

$$Pr(C = NNor|A_Q = Yes) \approx$$

$$\approx \frac{\#\ of\ participants\ with\ Non\text{-}normal\ conclusion\ and\ positive\ answer\ to\ Q}{Total\ \#\ of\ participants\ who\ have\ answered\ positively\ to\ Q}.$$

The entropy of the conclusion, $H(C)$, is defined as:

$$H(C) = -[Pr(C = Nor)\ log_2\ Pr(C = Nor) + \Pr(C = NNor)\ log_2\ Pr(C = NNor)].$$

Let the conditional entropy of the conclusion, given a positive or a negative answer be $H(C|A_Q = Yes)$ and $H(C|A_Q = No)$, respectively. The conditional entropy of the conclusions set C given the answer to a question Q is calculated as:

$$H(C|A_q) = [Pr(A_Q = Yes) * H(C|A_Q = Yes] + Pr(A_Q = No) * H(C|A_Q = No)]$$

The information gain, $IG(C, A_Q)$, is formally defined as:

$$IG(C, A_Q) = H(C) - H(C|A_q).$$

3.3 Probabilistic Analysis

As all questions lead to a very low information gain (see Section 4), we investigated for each question whether a positive answer to it has a significantly higher probability of indicating *Non-normal* conclusion, compared to a negative answer. Any such question is expected to at least indicate a likely *Non-normal* conclusion (even if it does not reliably identify *Normal* conclusions). We note that correctly identifying *Non-normal* conclusion is more important than correctly predicting *Normal* conclusion, because failure to identify an athlete with a *Non-normal* conclusion can be potentially life-threatening whereas misidentifying a *Normal* conclusion as *Non-normal* will only incur extra cost to conduct further tests. To perform this investigation, we have compared the probabilities $Pr(C = NNor|A_Q = Yes)$ with $Pr(C = NNor|A_Q = No)$ and used the Z-test (Walpole, 2002) to check whether the difference between the two resulting Bernoulli distributions is statistically significant. The procedure is as follows:

Given a question Q, let $T_{A_Q=Yes}$ be the total number of participants answering *Yes* while $T_{A_Q=No}$ denotes the total number of participants answering *No* to the question. The Z-statistic for the probabilities $Pr(C = NNor|A_Q = Yes)$ and $Pr(C = NNor|\ A_Q = No)$ is calculated as:

$$Z = \frac{Pr(C = NNor|A_Q = Yes) - Pr(C = NNor|A_Q = No)}{\sqrt{p(1-p)\left(\frac{1}{T_{A_Q=Yes}} + \frac{1}{T_{A_Q=No}}\right)}},$$

where

$$p = \frac{T_{A_Q=No} * Pr(C=NNor|\ A_Q=Yes) + T_{A_Q=Yes} * Pr(C=NNor|\ A_Q=No)}{T_{A_Q=Yes} + T_{A_Q=No}}.$$

For a two-sided test, if the value of the Z-statistic is greater than 1.96 or smaller than -1.96, the difference between the two probabilities is considered statistically significant with 95% confidence (p-value<=0.05).

3.4 Combination of Questions

We also considered the possibility that there are combinations of two or more questions that when answered together in the affirmative have a non-negligible association with the *Non-normal* conclusion. We investigated this association by identifying all such combinations of questions and counting how many athletes with *Non-normal* conclusions gave positive answers to the questions in each combination.

4. Results

As mentioned in the introduction (Section 1), as a baseline, we attempted to classify the dataset using traditional machine learning methods: *naïve Bayes, random forests, and support vector machine*. The goal was to assign the athletes into the correct adjudicated class (i.e., predict the ECG/Echo conclusion), based on their respective answers to the questions shown in Table 2. All three classifiers performed poorly for the *Non-normal* class, as evaluated using 10-fold cross validation. The classification Accuracy,

Table 4. Classification results from the WEKA implementation of naïve Bayes, random forests (RF) and support vector machine (SVM), using all the questions as attributes on the original (biased) dataset.

Classifier	Accuracy for *Normal class*	Accuracy for *Non-normal class*	Overall Accuracy	Precision	Recall	F-measure
Naïve Bayes	0.968	0.098	0.742	0.522	0.098	0.166
RF	0.905	0.115	0.70	0.298	0.115	0.166
SVM	0.968	0.098	0.742	0.522	0.098	0.166

Precision, Recall and F-measure for the three methods when applied to the original (biased) dataset are shown in Table 4. For the *Normal* class, the naïve Bayes, the random forest and the support vector machine classifiers correctly classified 96.8%, 90.5% and 96.8% instances, respectively, but their performance for the *Non-normal* class is extremely poor. As noted before, the performance over the *Non-normal* class is very important because misclassifying an athlete with abnormal heart condition as *Normal* is unacceptable in a pre-screening process.

We note that the poor performance of the classification for *Non-normal* class may be attributed to the bias in the dataset, which can lead the classifier to assign most of the instances to the majority class. To correct for this, we have used sub-sampling for balancing the set; Table 5 shows the classification results for the balanced datasets, averaged over 5 random sub-samples.

Correcting for the imbalance in the dataset indeed improved significantly the classification results for instances of the *Non-normal* class (in particular, Recall has significantly increased), but still, about 50% of the *Non-normal* cases are misclassified as *Normal* by naïve Bayes and 36% are misclassified as *Normal* by random forests. similarly the SVM classifier misclassifies 45% of the *Non-normal* cases as *Normal.* Moreover, the vast majority of the *Normal* cases (more than 50%, for all three classifiers) have been classified as *Non-normal*. We note that such a low level of performance is close to the classification level expected at random.

As discussed in Section 3.2, to pursue the information-content based analysis of each question, we calculated the information gain per question. The information gain associated with the questions *Q1-Q12* ranges between *0.001-0.003* and for *Q13* it is 0.008. Clearly, the information gain for all of the questions is very low, the highest being only 0.008 for question *Q13*, which is the result of the AHA-recommended physical exam. As a point of comparison, in a hypothetical case in which 70% of the *Yes* answers to question *Q13* would corresponded to a *Non-normal* conclusion, the information gain would have been 0.106, which is significantly higher than any of the gains associated with the questions. This, very low information content of each question explains the poor classification results, especially the close-to-random classification performance over the balanced dataset.

Table 5. Classification results from the WEKA implementation of naïve Bayes, random forests (RF) and support vector machine (SVM) using all the questions as attributes on the balanced dataset.

Classifier	Accuracy for *Normal* class	Accuracy for *Non-normal* class	Overall Accuracy	Precision	Recall	F-measure
Naïve Bayes	0.443	0.508	0.475	0.477	0.508	0.492
RF	0.467	0.639	0.553	0.545	0.639	0.589
SVM	0.459	0.549	0.504	0.504	0.549	0.525

To further analyze whether positive answers to the questions have higher probability of corresponding to *Non-normal* conclusion than negative answers, we have compared the probabilities $Pr(C = NNor|A_Q = Yes)$ and $Pr(C = NNor|A_Q = No)$. The histogram in Figure 1 shows for each question the conditional probability of the conclusion being *Non-normal* given that the answer to the question is *Yes*, side-by-side with the conditional probability of a *Non-normal* conclusion, when the answer to the same question is *No*.

We observe that for seven of the *questions (Q3, Q4, Q5, Q9, Q11, Q12 and Q13)*, the conditional probability $Pr(C = NNor|A_Q = Yes)$ is indeed somewhat higher than the conditional probability $Pr(C = NNor|A_Q = No)$. However, for six of the questions, *Q1*, *Q2*, *Q6*, *Q7*, *Q8*, and *Q10*, the probability of a *Non-normal* adjudication is

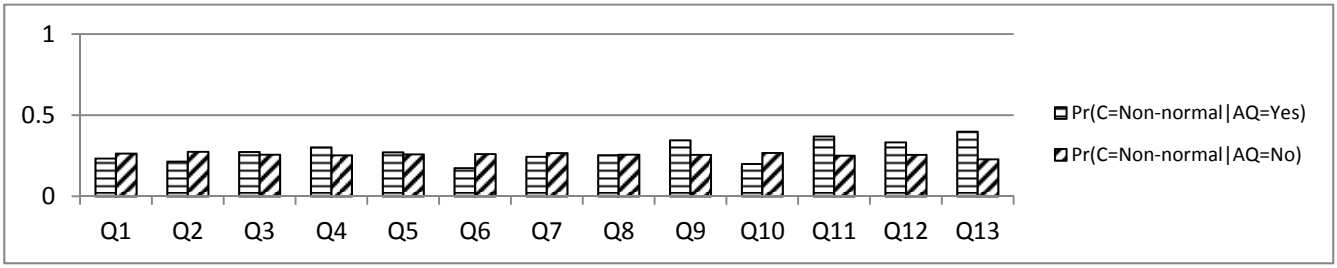

Figure 1. Conditional probability of adjudications being *Non-normal* when the answer to each question is *Yes* vs. *No*. Only seven of the questions (*Q3, Q4, Q5, Q9, Q11, Q12* and *Q13*) have a higher probability of identifying *Non-normal* heart condition when the answer is positive. Among these questions, the difference in the probability is statistically significant only for *Q13*. For six of the remaining questions (*Q1, Q2, Q6, Q7, Q8, Q10*) the probability of being *Non-normal* is *higher* when the answer is negative than when the answer is positive.

actually *higher* when the answer is negative than when the answer is positive. We used the Z-test to verify whether these differences are statistically significant, and found that *only for Q13 (the physical exam)*, the difference is statistically significant with a p-value of 0.016. Thus the only item in the questionnaire that is found to be marginally predictive of a *Non-normal* conclusion when the answer is positive than when the answer is negative, is the physical examination (*Q13*). An abnormal physical examination result corresponds to a higher probability of a *Non-normal* Echo/ECG results than a *normal* physical. However, even in this case the number of false negatives (i.e. the number of *Non-normal*s that are left undetected) is 94 out of a total of 110 *Non-normal*s, which is very high.

As none of the questions except for *Q13* is individually predictive of *Non-normal* conclusions, we investigated the possibility that positive answers to a *combination of two or more questions* may indicate abnormality in athletes' heart condition. Table 6 shows the number of participants who answered in the affirmative to two or more question, along with the number of participants among them who were identified as *Non-normal* by the cardiologist (based on ECG and Echo tests).

Table 6 shows that only a single athlete with a *Non-normal* adjudication has answered 7 or more questions in the affirmative; this is clearly an insufficient sample to draw any conclusions from. Among the 11 athletes who answered a combination of 5 questions in the affirmative, three were identified as *Non-normal*. However, each of these three athletes answered a *different* subset of 5 questions in the affirmative. The same is true for the three participants with *Non-normal* conclusion who answered a combination of 6 questions in the affirmative. Thus there is no subset consisting of 5 or 6 questions that is associated with more than a single abnormal case.

We have thus investigated the correspondence between *Non-normal* conclusions and the affirmative answers to smaller subsets of 2, 3 or 4 questions, looking at such subsets of questions for which at least one of the athletes who answered in the affirmative was identified as *Non-normal*. No combination of 4 questions was answered in the affirmative by more than one athlete. As for combinations of *three* questions, 5 athletes answered in the affirmative to the combination: *Q1*, *Q2* and *Q8*. Among these athletes, *two* were identified as *Non-normal*; there is no other combination of three questions for which more than one participant with *Non-normal* heart condition answered in the affirmative. Similarly, there is no 2-question combination for which more than two athletes with *Non-normal* conclusion answered in the affirmative. Hence there does not exist any combination of questions for which a significant number of athletes with non-normal heart condition answered in the affirmative. As such there is no combination of two or more questions for which a positive answer is strongly indicative of abnormality in the athletes' heart conditions.

Table 6. Number of participants who answered more than one question in the affirmative, and the corresponding number of *Non-normal* identified among them by ECG/Echo tests.

# of questions answered together in the *affirmative*	8	7	6	5	4	3	2
Total # of participants	1	3	7	11	26	46	72
# of *Non-normal* identified	0	1	3	3	9	12	13

All of the results described above demonstrate that relying on normal findings from the physical examination (*Q13*), and on negative answers to questions *Q1*-*Q12* in the AHA questionnaire as a way to assess whether athletes can safely participate in competitive activities leads to a high rate of false negatives. That is, athletes with potential heart abnormalities (identified by a cardiologist through ECG and Echo tests) are very likely to be pre-screened as *Normal*, and not be referred for further examination by a specialist. This is clearly an undesirable scenario in a pre-screening process. Additionally, affirmative answers to one or more questions in the questionnaires are not effective predictors of *Non-normal* conclusions.

5. Conclusion

We set out to build a classifier that could predict potential abnormalities in young athletes' heart-condition, using data from close to 500 athletes who were examined using the AHA-based 12-element screening procedure. The ground truth used for potential abnormality was determined by an experienced cardiologist based on Electro- and Echo-cardiogram tests, which are *not included* in the AHA screening procedure.

The poor performance of several well-studied machine-learning classifiers, (and particularly the close-to-random classification performance measured on the balanced dataset), when using all the elements in the questionnaire as attributes, leads us to conduct an in-depth study of the data and the questions. We aimed to determine each element's ability (or there lack-of) to identify abnormality. Underlying the study was the expectation that the classifiers performance may be improved by using the most informative subset of questions as attributes.

However, surprisingly, our results show that in terms of information content, none of the elements included in the questionnaire contributes significant information about the findings obtained through traditional ECG and Echo-based tests. As such, improvement in the classification results was not attainable using any subset of the questions as attributes. Further analysis of the respective conditional probabilities through statistical tests, indicates that an *abnormal* physical examination (*Q13*) is the only item within the questionnaire that is even associated with a statistically-significantly higher probability of a *Non-normal* ECG/Echo than a *normal* physical examination. But even this item still gives rise to many false negatives.

Thus, the results of this study strongly suggest that the 12-element procedure advocated by the American Heart Association for pre-participation screening of young athletes is not effectively correlated with or predictive of the outcome obtained by a standard, more extensive examination by a cardiologist using ECG and Echo tests.

Pragmatically speaking, the conclusion from this study implies that ECG (and possibly Echo) should be considered for screening athletes in the Unites States. Future research following the machine-learning and informatics-driven approach as used in this study will examine whether *using one or more of the cardiovascular tests* such as electrocardiogram or echocardiogram together with any combination of all or some of the AHA-based questions may improve the efficacy of pre-participation screening.

Acknowledgments

This work was partially supported by HS's NSERC Discovery Award #298292-2009, NSERC DAS #380478-2009, CFI New Opportunities Award 10437, and Ontario's Early Researcher Award #ER07-04-085, and by TA's grant HL 098046 from the National Institutes of Health.

References

Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems with Applications*, *36*(2), 3240-3247. Elsevier Ltd.

Breiman, L. (2001). Random forests. *Machine learning*.

Chazal, P. D., & Reilly, R. B. (2006). patient-adapting heartbeat classifier using ECG morphology and heartbeat interval features. *IEEE Transactions on Biomedical Engineering*, *53*(12), 2535-2543.

Chhatwal, J., et. al . (2009). A logistic regression model based on the national mammography database format to aid breast cancer diagnosis. *AJR. American journal of roentgenology*, *192*(4), 1117-27.

Corrado, D., et al., (1998). Screening for hypertrophic cardiomyopathy in young athletes. *The New England journal of medicine*, *339*(6), 364-9.

Corrado, D. et al., (2005). Cardiovascular pre-participation screening of young competitive athletes for prevention of sudden death: proposal for a common European protocol. Consensus Statement of the Study Group of Sport Cardiology of the Working Group of Cardiac Rehabilitation and. European heart journal, 26(5), 516-24.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 297, 273-297.

Glover, D. W., & Maron, B. J. (2007). Evolution in the process of screening United States high school student-athletes for cardiovascular disease. *The American journal of* cardiology, 100(11), 1709-12.

Hall, M. et al., (2009). The WEKA data mining software. ACM SIGKDD Explorations Newsletter, 11(1).

Kanagalingam, S., et al., (2010). Abstract 19765: Efficacy of the American Heart Association Questionnaire in Identifying Electrocardiographic and Echocardiographic Abnormalities in Young Athletes During Community-based Screening. *Circulation*.

Klement, W., Wilk, et al., (2009). Dealing with Severely Imbalanced Data. *ICEC 2009 Workshop, PAKDD*.

Maron, B. J. (2003). Sudden death in young athletes. *The New England journal of medicine*, *349*(11), 1064-75.

Maron, B. J. (2007). Hypertrophic cardiomyopathy and other causes of sudden cardiac death in young competitive athletes, with considerations for preparticipation screening and criteria for disqualification. *Cardiology clinics*, *25*(3), 399-414, vi.

Maron, B. J., et al. (1996). Cardiovascular Preparticipation Screening of Competitive Athletes: A Statement for Health Professionals From the Sudden Death Committee (Clinical Cardiology) and Congenital Cardiac Defects Committee (Cardiovascular Disease in the Young), AHA *Circulation*, *94*(4), 850-856.

Melgani, F., & Bazi, Y. (2008). Classification of electrocardiogram signals with support vector machines and particle swarm optimization. *IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society*, *12*(5), 667-77.

Mitchell, T. M. (1997). *Machine Learning* (p. 432). McGraw-Hill.

Osowski, S., Hoai, L. T., & Markiewicz, T. (2004). Support vector machine-based expert system for reliable heartbeat recognition. *IEEE transactions on bio-medical engineering*, *51*(4), 582-9.

Pigozzi, F., & Rizzo, M. (2008). Sudden death in competitive athletes. *Clinics in sports medicine*, *27*(1), 153-81, ix.

Qu, L., et al. (2010). A Naïve Bayes Classifier for Differential Diagnosis of Long QT Syndrome in Children. *Int. Conf. on Bioinformatics and Biomedicine* (pp. 433-437).

Statnikov, A., & Wang, L. (2008). A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC bioinformatics*, *9*(319).

Walpole, R. E. et al. (2002). *Probability and statistics for engineers and scientists* (Vol. 6). Prentice Hall Upper Saddle River, NJ:

Wever-Pinzon, O. E., et al. (2009). Sudden cardiac death in young competitive athletes due to genetic cardiac abnormalities. *Anadolu Kardiyol Derg*, *9*(2), 17–23.

Yu, S., & Chou, K. (2008). Integration of independent component analysis and neural networks for ECG beat classification. *Expert Systems with Applications*, *34*(4), 2841-2846.