Comparing Symbolic Representations of Cardiac Activity to Identify Patient Populations with Similar Risk Profiles

Z Syed[1], BM Scirica[2], CM Stultz[1], JV Guttag[1]

[1]Massachusetts Institute of Technology, Cambridge, MA, USA
[2]TIMI Study Group, Brigham and Women's Hospital, Boston, MA, USA

Abstract

This paper proposes electrocardiographic mismatch (ECGM) to quantify differences in the long-term ECG signals for two patients. ECGM compares the symbolic distributions of ECG signals and measures how different patients are electrocardiographically. Using ECGM, we propose a hierarchical clustering scheme that can identify patients in a population with anomalous ECG characteristics. When applied to a population of 686 patients suffering non-ST-elevation ACS, our approach was able to identify patients at an increased risk of death and myocardial infarction (HR 2.8, p=0.003) over a 90 day follow-up period.

1. Introduction

In this paper, we propose a comparative approach to identify patients at increased risk of adverse cardiovascular outcomes. Our approach is based upon the hypothesis that patients at increased risk of death following ACS comprise a minority that is electrocardiographically dissimilar from the much larger set of low risk patients, i.e., high risk patients can be recognized as population outliers.

Evidence suggests that high risk patients constitute a small minority. For example, cardiac mortality over a 90 day period following ACS was reported to be 1.79% for the SYMPHONY trial involving 14970 patients [1] and 1.71% for the DISPERSE2 trial with 990 patients [2]. The rate of myocardial infarction (MI) over the same period for the two trials was 5.11% for the SYMPHONY trial and 3.54% for the DISPERSE2 trial.

In contrast to using specific features, we focus instead on finding cases that are atypical in morphology and dynamics. We propose a new metric, called the *electrocardiographic mismatch* (ECGM), which quantifies the extent to which the long-term ECG recordings from two patients differ. The pairwise differences are used to partition patients into groups with similar ECG characteristics and potentially common risk profiles.

Our hypothesis is that those patients whose long-term electrocardiograms did not match the dominant group in the population, are at increased risk of adverse cardiovascular events. These cases have a high electrocardiographic mismatch relative to the majority of the patients in the population, and form one or more subgroups that are suspected to be at an increased risk of adverse events in the future.

Our approach is orthogonal to the use of specialized high risk features along two important dimensions. Firstly, it does not require the presence of significant prior knowledge. We only assume that ECG signals from patients who are at high risk differ from those of the rest of the population. There are no specific assumptions about the nature of these differences. Secondly, the ability to partition patients into groups with similar ECG characteristics and potentially common risk profiles allows for a more fine-grained understanding of a how a patient's future health may evolve over time. Matching patients to past cases with similar ECG signals could lead to more accurate assignments of risk scores for particular events such as death and MI.

2. Electrocardiographic mismatch

The electrocardiographic mismatch (ECGM) between two patients, p and q, is calculated in two steps.

As a first step, the ECG signal for each patient is symbolized using the techniques similar to those described in [3]. Symbolization involves segmenting the original ECG signal into heart beats, and then separating the beats into different groups based on their morphology.

To segment the ECG signal into beats, we use two open-source QRS detection algorithms with different noise sensitivities. The first of these makes use of digital filtering and integration [4] and has been shown to achieve a sensitivity of 99.69%, while the second is based on a length transform after filtering [5] and has a sensitivity of 99.65%. Both techniques have a positive predictivity of 99.77%. QRS complexes were marked at

locations where these algorithms agreed.

The process of partitioning segmented beats into different groups is carried out using the dynamic time-warping (DTW) algorithm described in [3], which is able to quantify differences in morphology between two beats. Using DTW, beats that are similar in morphology are considered to belong to the same morphology class, while those that have a high morphology difference according to DTW are considered to belong to a different group. Details of DTW and the process of partitioning beats into groups with distinct morphologies are presented below in more detail.

Given any two beats, x_1 and x_2, of length l_1 and l_2 respectively, DTW produces the optimal alignment of the two sequences by first constructing an l_1-by-l_2 distance matrix d such that:

$$d[i,j] = \left(x_1[i] - x_2[j]\right)^2 \quad (1)$$

Each entry (i,j) in this matrix represents the square of the difference between samples $x_1[i]$ and $x_2[j]$. A particular alignment then corresponds to a path, φ, through the distance matrix of the form:

$$\varphi(k) = \left(\varphi_1(k), \varphi_2(k)\right), \ 1 \le k \le K \quad (2)$$

where φ_1 and φ_2 represent row and column indices into the distance matrix, and K is the alignment length.

The optimal alignment produced by DTW minimizes the overall cost:

$$C(x_1, x_2) = \min_{\varphi} C_{\varphi}(x_1, x_2) \quad (3)$$

where C_{φ} is the total cost of the alignment path φ and is defined as:

$$C_{\varphi}(x_1, x_2) = \sum_{k=1}^{K} d\left[x_1[\varphi_1(k)], x_2[\varphi_2(k)]\right] \quad (4)$$

The search for the optimal path is carried out in an efficient manner using dynamic programming. The final energy difference between the two beats x_1 and x_2, is given by the cost of their optimal alignment, and depends on both the amplitude differences between the two signals, as well as the length K of the alignment (which increases if the two beats differ in their timing characteristics).

DTW quantifies changes in morphology resulting from amplitude and timing differences between two beats. Using this information, beats with distinct morphologies can be placed in different groups, with each group being assigned a unique label or symbol. This is done by means of a Max-Min iterative clustering algorithm that proceeds by choosing an observation at random as the first centroid c_1 and setting the set S of centroids to $\{c_1\}$. During the i-th iteration, c_i is chosen such that it maximizes the minimum distance between c_i and observations in S:

$$c_i = \arg\max_{x \notin S} \min_{y \in S} C(x, y) \quad (5)$$

where $C(x,y)$ is the DTW difference between x and y as defined in (3). The set S is incremented at the end of each iteration such that $S = S \cup c_i$.

The number of clusters discovered by Max-Min clustering is chosen by iterating until the maximized minimum dissimilarity measure in (5) falls below a specified threshold θ. At this point, the set S comprises the centroids for the clustering process, and the final assignment of beats to clusters proceeds by matching each beat to its nearest centroid. Each set of beats assigned to a centroid constitute a unique cluster. The final number of clusters, k, obtained using this process depends on the separability of the underlying data.

The overall effect of DTW-based partitioning of beats is to transform the original raw ECG signal into a sequence of symbols, i.e., into a sequence of labels corresponding to the different beat morphology classes that occur in succession. A more detailed discussion of the symbolization process is provided in [3].

Denoting the set of symbols for patient p as S_p and the set of probabilities with which these symbols occur in the electrocardiogram as P_p (for patient q an analogous representation is adopted), we calculate the ECGM between these patients as:

$$ECGM_{p,q} = \sum_{a \in S_p} \sum_{b \in S_q} C(a,b) P_p[a] P_q[b] \quad (6)$$

In (6), $C(a,b)$ corresponds to the dynamic time-warping cost of aligning the centroids of symbol classes a and b.

Intuitively, the electrocardiographic mismatch between patients p and q corresponds to an estimate of the expected dynamic time-warping cost of aligning any two randomly chosen beats from these patients. The ECGM calculation in (6) achieves this by weighting the cost between every pair of symbols between the patients by the probabilities with which these symbols occur.

An important feature of ECGM is that it is explicitly designed to avoid the need to set up a correspondence between the symbols of patients p and q for comparative purposes. In contrast to cluster matching techniques that compare data for two patients by first making an assignment from symbols in one patient to the other, ECGM does not require any cross-patient registration of symbols and performs weighted comparisons between all symbols for p and q.

3. Hierarchical clustering

For every pair of patients in a population, the electrocardiographic mismatch between them is computed using the techniques described in Section 2. The resulting divergence matrix, D, relating the pairwise electrocardiographic mismatches between all the patients

is used to partition the population into groups with similar cardiac characteristics. This process is carried out by means of hierarchical clustering [6].

Hierarchical clustering starts out by assigning each patient to a separate cluster. It then proceeds to combine two clusters at every iteration, choosing clusters that obey some concept of being the "closest" pair. We use a definition of closest that corresponds to merging two clusters A and B for which the mean electrocardiographic mismatch between the elements of the clusters is minimized, i.e., we choose clusters A and B such that they minimize the merge distance, f, which is given by :

$$f = \frac{1}{|A|.|B|} \sum_{x \in A} \sum_{y \in B} ECGM_{x,y} \qquad (7)$$

Where $|A|$ and $|B|$ correspond to the number of elements in each cluster.

Intuitively, this approach picks two clusters to merge that are closest in the sense that the average distance between elements in the two clusters is minimized. This definition of closest is similar to the unweighted pair group method with arithmetic mean (UPGMA) or average linkage criterion [6].

Broadly speaking, there are two approaches to decide when to terminate the iterative clustering process. The simplest approach is to terminate at the iteration when the clustering process has produced a pre-determined number of clusters. However, in this case we have no prior assumptions about the appropriate number of clusters. We therefore use a more complex approach in which the number of clusters is determined by the dataset.

The merge distance defined in (7) is monotonically nondecreasing with iteration number. Small increases in the merge distance suggest that the clustering process is merging clusters that are close. Conversely, large merge distances correspond to clusters being merged that are dissimilar. We therefore use the merge distance to indicate when the clustering process is beginning to show diminishing returns, i.e., merging clusters that are increasingly far apart. Continuing beyond this point may lead to the new clusters created containing heterogenous elements. We therefore terminate the clustering process when the merge distance for the next three iterations would show a quadratic concave up increase.

4. Evaluation and results

The population used for this work comprised patients in the DISPERSE2 trial [7], who were admitted to a hospital with non-ST-elevation (NSTE) acute coronary syndromes. Three lead continuous ECG (cECG) monitoring was performed for a median duration of 4 days at a sampling rate of 128 Hz. The endpoints of cardiac death and MI were adjudicated by a blinded Clinical Events Committee for a median follow-up period of 60 days. The maximum follow-up was 90 days. Data from 686 patients was available after removal of noise-corrupted signals. During the follow-up period there were 14 cardiac deaths and 28 MIs.

To evaluate the ability of ECGM to identify patients at increased risk of future cardiovascular events, we first separated the patients into a dominant normal sub-population (i.e., the low risk ECGM group) and a group of abnormal patients (i.e., the high risk ECGM group). This was done by terminating hierarchical clustering automatically as described in Section 3 and labeling all patients outside the largest cluster as being abnormal and potentially high risk. In the subsequent discussion, we denote this new risk variable as the ECG Non-Dominance (ECGND). Patients placed in the non-dominant group by ECGM clustering were assigned an ECGND value of 1, while those in the dominant group had a value of 0.

Kaplan-Meier survival analysis was used to study the event rates for death and MI. Hazard ratios (HR) and 95% confidence interval (CI) were estimated by using a Cox proportional hazards regression model to study event rates in patients within the dominant and non-dominant groups. The HR for the dominant and non-dominant ECGM groups was compared to other clinical risk variables; age, gender, smoking history, hypertension, diabetes mellitus, hyperlipidemia, coronary heart disease (CHD), prior MI, prior angina and ST depression on holter. The risk variables were also examined using multivariate analysis.

The results of univariate and multivariate analysis are given in Tables 1 and 2. In the case of ECGND, patients who were electrocardiographically mismatched with the dominant group of the population showed an increased risk of adverse cardiovascular events. Patients outside the dominant cluster had a much higher rate of death and MI during follow-up i.e., the occurrence of either of these adverse outcomes, the cumulative incidence in the high risk group was 9.17% as opposed to 3.50% in the low risk group ($p<0.01$).

5. Discussion

In this paper, we explore the hypothesis that patients at increased risk of adverse cardiovascular outcomes following acute coronary syndrome (ACS) may be detected as a minority that is electrocardiographically dissimilar to other patients in the population. To test this, we developed an approach to automatically quantify the difference between a pair of ECG recordings. We also described how this information can be used in a hierarchical clustering framework to partition patients into similar groups, with matching long-term electrocardiograms. Using such an approach, we searched for clusters of patients that are population outliers. Our study of 686 patients showed that patients who are

electrocardiographically mismatched from the majority patient population are at a considerably increased risk of adverse cardiovascular events such as death and MI over a 90 day period following NSTEACS.

Acknowledgement

This work was supported by the Center for Integration of Medicine and Innovative Technology (CIMIT), the Harvard-MIT Division of Health Sciences and Technology (HST) and the Industrial Technology Research Institute (ITRI). The DISPERSE2 trial was supported by AstraZeneca (AZ).

References

[1] Newby LK, Bhapkar MV, White HD, Topol EJ, Dougherty FC, Harrington RA, Smith MC, Asarch LF, Califf RM. Predictors of 90-day outcome in patients stabilized after acute coronary syndromes. European Heart Journal. 2003; 24:172-181.

[2] Cannon C, Husted S, Harrington R, Scirica B, Emanuelsson H, Peters G, Storey R. Safety, tolerability, and initial efficacy of AZD6140, the first reversible oral adenosine diphosphate receptor antagonist, compared with clopidogrel, in patients with non-ST-segment elevation acute coronary syndrome: primary results of the DISPERSE-2 trial. J Am Coll Cardiol. 2007;50:1844-51.

[3] Syed Z, Guttag J, Stultz C. Clustering and symbolic analysis of cardiovascular signals: discovery and visualization of medically relevant patterns in long-term data with limited prior knowledge. EURASIP Journal on Applied Signal Processing, 2007.

[4] Hamilton PS, Tompkins WJ. Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmia database. IEEE Trans Biomed Eng. 1986; 33:1157-1165.

[5] Zong W, Moody GB, Jiang D. A robust open-source algorithm to detect onset and duration of QRS complexes. Comput Cardiol. 2003; 30:737-740.

[6] R, Hart P. Pattern Classification. Wiley-Interscience. 2000;2nd ed.

[7] Cannon C, Husted S, Harrington R, Scirica B, Emanuelsson H, Peters G, Storey R. Safety, tolerability, and initial efficacy of AZD6140, the first reversible oral adenosine diphosphate receptor antagonist, compared with clopidogrel, in patients with non-ST-segment elevation acute coronary syndrome: primary results of the DISPERSE-2 trial. J Am Coll Cardiol. 2007;50:1844-51.

Variable	Hazard Ratio	P Value
Age	1.05	< 0.01
Gender	0.77	0.468
Smoker	1.15	0.670
Hypertension	1.99	0.103
Diabetes	1.84	0.077
Hyperlipidemia	0.74	0.362
CHD	0.85	0.635
Prior MI	1.46	0.281
Prior angina	1.03	0.932
ST depression	1.04	0.910
ECGND	2.58	< 0.01

Table 1: Univariate association of risk variables with death and MI over 90 day follow-up period.

Variable	Hazard Ratio	P Value
Age	1.05	< 0.01
Gender	0.63	0.242
Smoker	1.52	0.285
Hypertension	1.81	0.185
Diabetes	1.41	0.350
Hyperlipidemia	0.76	0.429
CHD	1.14	0.720
Prior MI	1.33	0.451
Prior angina	0.87	0.707
ST depression	0.74	0.387
ECGND	2.43	< 0.01

Table 2: Multivariate association of risk variables with death and MI over 90 day follow-up period.

Address for correspondence

Zeeshan Syed
32 Vassar Street, 32G-916, Cambridge, MA 02139, USA
zhs@csail.mit.edu