The PhysioNet / Computers in Cardiology Challenge 2008:
T-Wave Alternans

GB Moody

Massachusetts Institute of Technology, Cambridge, MA, USA

Abstract

The 9th annual PhysioNet/Computers in Cardiology challenge invited participants to measure T-wave alternans (TWA) in a set of 100 two-minute electrocardiograms that included subjects with a variety of risk factors for sudden cardiac death (including ventricular tachyarrhythmias, transient myocardial ischemia, and acute myocardial infarctions), healthy controls, and synthetic ECGs with calibrated amounts of artificial TWA. The participants' TWA estimates were used to develop a ranking of the 100 test cases in order of TWA content, and the Kendall rank correlation coefficient between this reference ranking and each individual participant's ranking of the 100 cases was calculated as a score (between -1 and 1; actual scores were between 0.11 and 0.92). The challenge yielded insights into the strengths and weaknesses of classic and novel TWA analyses, open-source implementations of a variety of methods, and a set of freely available ECGs with reference rankings of TWA content.

1. Introduction

One hundred years after it was first reported by Hering [1], T-wave alternans is widely understood to be an important indicator of risk of sudden cardiac death. Yet for most of that time TWA was believed to be rare, until 1981, when at Computers in Cardiology, Adam, Akselrod, and Cohen reported the existence of microvolt-level TWA, too small in amplitude to be detected visually at standard display scales[2].

T-wave alternans is a pattern in the ECG characterized by two (rarely more) distinct forms of T-waves appearing in alternation, at or above a patient-specific heart rate generally in the range of 90 to 120 beats per minute. Although the mechanisms have not been fully elucidated, a large amount of empirical evidence collected during the past 25 years has demonstrated an association between the amount of TWA, the heart rate at which it appears, and the risk of sudden cardiac death (SCD). In particular, the absence of significant TWA in a patient with congestive heart failure, low ejection fraction, or a recent myocardial infarction is strongly predictive of a low risk of SCD. A positive finding in such a patient, though less specific, may indicate that an implantable cardiac defibrillator would be appropriate, an indication that can be confirmed using invasive testing. A review by Armoundas, Tomaselli, and Esperer discusses mechanisms that may account for the associations among TWA and other risk factors for SCD, as well as clinical applications of TWA[3].

Since TWA analysis is performed on the surface ECG, it is an inexpensive and non-invasive test. In clinical applications, TWA analysis can be done as part of an exercise stress test, but there is interest in the research community in using conventional long-term (Holter) ECG recordings to observe TWA in the context of activities of daily living.

A variety of algorithms for detecting and quantifying TWA have been proposed, employing techniques from linear and nonlinear signal processing such as spectral analysis, complex demodulation, counting zero-crossings in a series of correlation coefficients, periodogram and complex demodulation analysis of T-wave principal components, Capon filtering, Poincaré maps, periodicity transforms, statistical tests, moving averages, maximum likelihood estimators and generalized likelihood ratio tests, and more. For a comprehensive and systematic discussion of methods for TWA detection and analysis, see the review by Martínez and Olmos[4].

It remains very difficult to validate or to compare any of these algorithms, since no generally accepted objective criteria exist for measuring TWA, and no generally available set of validation data exists as a basis for comparison. This Challenge aimed to improve understanding of the strengths and limitations of classic and novel TWA analysis methods; to establish a collection of reference ECGs ranked by severity of TWA, as determined by a preponderance of evidence; and to encourage the development and dissemination of open-source TWA detectors and estimators in order to support and stimulate further research on the properties and implications of TWA.

2. Methods

Interested researchers were invited to nominate ECG recordings for use in the Challenge. On the basis of the nominations, the Challenge data set of 100 two-minute ECGs[5] was posted on PhysioNet[6]. Of these, 56 were ECGs obtained from 26 subjects with known risk factors for sudden cardiac death, including 24 from subjects in the PTB Diagnostic ECG Database[7] who had myocardial infarctions; 12 from subjects in the Long-Term ST Database[8] who had coronary artery disease and transient myocardial ischemia; 10 from subjects in the St.-Petersburg Institute of Cardiological Technics 12-lead Arrhythmia Database[9], including subjects with myocardial ischemia, ventricular tachycardia, and myocardial infarction; and 10 from subjects in the Sudden Cardiac Death Holter Database[10] who experienced sustained ventricular arrhythmias and cardiac arrest within minutes to a few hours after the segments selected for the Challenge data set. In most cases, the Challenge data set contains two or more ECGs per subject to permit followup studies of the evolution of TWA in individuals over time.

The 44 remaining ECGs in the Challenge data include 12 ECGs of healthy subjects (6 from the PTB Database and 6 from the MIT-BIH Normal Sinus Rhythm Database[11]), and a group of 32 synthetic ECGs[12]. Thirty of this final group contained artificial TWA in calibrated amounts (the last two contained small amounts of added noise rather than TWA). The artificial TWA was created by modulating the T-wave loop of the synthetic vectorcardiogram (VCG), then projecting the VCG onto 12 scalar ECG leads using one of 5 individual Dower transform (IDT) matrices derived from 5 subjects in the PTB Database. In this way, the artificial TWA is distributed across the scalar ECG leads. Each IDT was used to produce 6 ECGs for one "subject", with varying TWA amplitudes chosen from 2, 4, 6, 8, 10, 13, 15, 17, 30, 45, and 60 microvolts. The TWA amplitudes were defined as the maximum vector difference between the forms of the T-wave loop in the VCG.

For each record in the challenge data set, participants were asked to detect TWA, or to estimate its peak magnitude, using a fully-automated method. Two participants detected but did not quantify TWA, reporting 1 as the estimate for each record in which TWA was detected, and 0 as the estimate for the remaining records. The remaining participants submitted quantitative estimates of TWA peak magnitude, in varying units.

Nearly thirty participants analyzed part or all of the challenge data set, and 23 participants submitted complete sets of 100 estimates of TWA. Since the various analyses produce incommensurate measurements, they cannot be compared numerically across participants. Within a set of measurements from a single participant, however, they can be used to arrange the ECGs in order, from least to most

amount of TWA as estimated by that participant, and this is the basis of the scoring that was used in the Challenge:

1. For each entry, the ECGs are ranked by the magnitudes of the associated TWA estimates. Thus the ECG with the lowest TWA estimate in a given entry receives the rank of 1 for that entry, the ECG with the second-lowest TWA estimate gets a rank of 2, etc.

2. For each participant who is able to distinguish between the synthetic ECGs with high and low amounts of TWA, one entry is selected for the next step. At the conclusion of the Challenge, 19 participants met this criterion.

3. Each ECG receives a median rank, which is the median of the ranks assigned it by the selected entries, and a reference ranking is made by sorting the median ranks (i.e., the ECG with the lowest median rank gets a reference rank of 1, etc.).

4. The score for each entry is the Kendall (τ) rank correlation coefficient between the entry ranking and the reference ranking, where 1 is perfect agreement and -1 is perfect disagreement[13].

Participants were allowed to revise their entries in order to explore other methods or to improve their scores. The number of entries was limited to reduce the likelihood of obtaining a superior result by chance. Throughout the challenge, preliminary scores, calculated on the basis of preliminary reference rankings, were provided as feedback to participants.

3. Results

The final scores of the top 5 participants were 0.911 (Jubair Sieed), 0.890 (Giovanni Bortolan), 0.881 (Alexander Khaustov), 0.827 (Dingchang Zheng), and 0.779 (Philip Langley). The first- and third-place results were obtained by participants in the open-source division of the Challenge, who contributed the software they developed for the challenge in source form for further study by the research community. These sources, the Challenge data set, the final reference ranking, and final scores for the 19 participants whose entries were used to derive the final reference ranking, are all available at http://physionet.org/-challenge/2008/.

The median score for a set of random measurements, given the final reference ranking, was 0.358; a score above 0.436, achieved by 21 participants, is significant ($p > 0.99$).

Notably, although the top-scoring entries that included T-wave measurements were able to rank the synthetic ECGs for a single "subject" in the correct order by amount of artificial TWA (see figure 1), the final reference ranking contains a small number of incorrectly ordered synthetic ECGs (see figure 2). Furthermore, the estimated TWA amplitudes in synthetic ECGs differ systematically across "subjects".
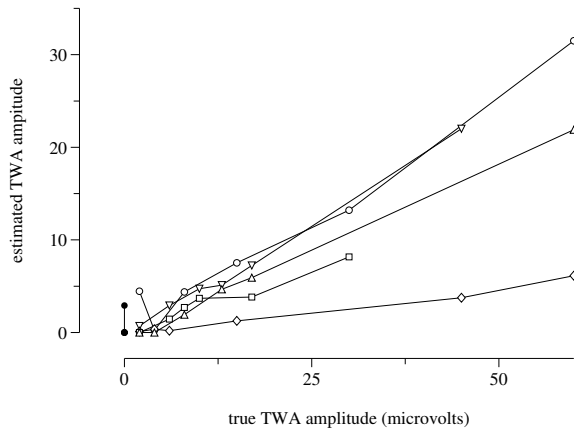
Figure 1. Estimated vs. known TWA amplitudes in the synthetic subset of the Challenge data set, from a top-scoring entry. Data for ECGs generated using the same "subject" (see text) are connected. Ideally, connected sets would lie along coincident straight lines. Any line segments with negative or zero slopes reflect estimates that are inconsistent with known TWA amplitudes. Differences in slope between connected sets reflect systematic inter-subject differences in TWA estimation.
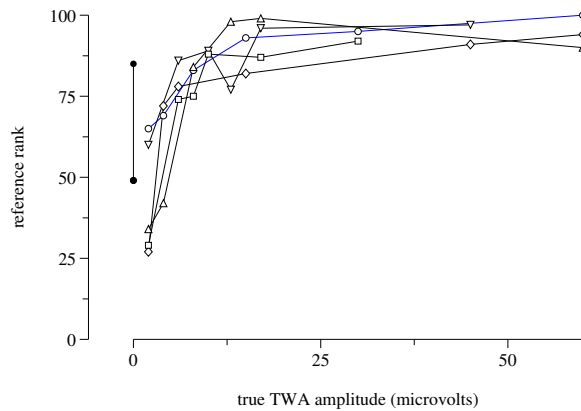

Figure 2. Reference rank vs. known TWA amplitudes in the synthetic subset. As in figure 1, data for ECGs of the same "subject" are connected. Negative slopes reflect estimation inconsistencies that are shared across a majority of the 19 entries used to determine the reference ranks. Since the ordinates are ranks rather than measurements, non-linearity reflects the non-uniform distribution of TWA amplitudes across the entire challenge data set.

These observations reflect both the difficulty of measuring TWA and the limitations of the scoring method. Since the scores indicate how well each participant's assessments match those of the group, if a group ranking of a given ECG is incorrect, then a correct ranking of that ECG by an individual diminishes that individual's score. In the context of this challenge, errors in the reference ranking of the synthetic ECGs did not alter the order of the final scores, as shown by an experiment in which scores were recalculated using an alternate reference ranking that was adjusted to correct these errors.

4.   Discussion and conclusions

A possible objection to the scoring algorithm is that the inclusion of a given entry in the determination of the median introduces a bias in favor of that entry. Given a sufficiently large number of entries, such a bias will be insignificant, but in any case, the scores were recalculated using ten alternative reference rankings that excluded each of the top ten entries, as well as rankings that excluded various combinations of these top-ranked entries. The scores obtained varied slightly but the order of scores of the top ten participants was unaffected by the choice of reference ranking, except that the removal of either of the top two entries reverses the order of the second and third entries.

Another concern is the effect of including the "detector" entries among those used to determine the median rankings, since they do not contain information needed to distinguish low from moderate from high amplitude TWA, and their inclusion may tend to diminish the influence of other entries that do contain such information. Again, however, an experiment showed that exclusion of "detector" entries had only minimal effects on the reference ranking and the final scores, most likely because only 2 of 19 participants submitted such entries.

Interestingly, the top score is only the second-best score achieved by its owner, who received a score of 0.919 for his initial entry. (The substitution did not affect the order of scores.) In all other cases, the final scores are the highest achieved by their owners.

For details of the algorithms used by the participants in this challenge, see their papers in this volume of Computers in Cardiology. They employed a wide variety of time- and frequency-domain approaches, including methods previously described in the literature as well as novel methods developed specifically for this Challenge. It is clear that many of these methods are in general agreement with respect to a subset of ECGs in the Challenge data set that were judged to have high amounts of TWA, but the TWA in some of the high-TWA cases was missed or significantly underestimated in some entries, suggesting that an algorithm making use of multiple approaches is likely to yield better sensitivity even for high-amplitude TWA than sim-

pler approaches. The top-scoring algorithms were able to rank the synthetic ECGs with low-amplitude TWA with many fewer errors than in most entries, suggesting that the best algorithms can detect TWA with an amplitude of as little as 4 microvolts.

Acknowledgements

References

[1] Hering HE. Das Wesen des Herzalternans. Muenchener med Wochenschr 1908;4:1417–1421.

[2] Adam DR, Akselrod S, Cohen RJ. Estimation of ventricular vulnerability to fibrillation through T-wave time series analysis. In Computers in Cardiology 1981. Los Alamitos: IEEE Computer Society Press, 1981; 307–310.

[3] Armoundas AA, Tomaselli GF, Esperer HD. Pathophysiological basis and clinical application of T-wave alternans. JACC 2002;40:207–217.

[4] Martínez JP, Olmos S. Methodological principles of T-wave alternans: a unified framework. IEEE Trans Biomed Eng 2005;52(4):599–613.

[5] T-Wave Alternans Challenge Database. http://physionet.org/pn3/twadb/.

[6] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation 2000 (June 13);101(23):e215–e220. Circulation Electronic Pages: http://circ.ahajournals.org/cgi/content/full/101/23/e215.

[7] PTB Diagnostic ECG Database. http://physionet.org/physiobank/database/ptbdb/.

[8] Long-Term ST Database. http://physionet.org/physiobank/database/ltstdb/.

[9] St.-Petersburg Institute of Cardiological Technics 12-lead Arrhythmia Database. http://physionet.org/physiobank/database/incartdb/.

[10] Sudden Cardiac Death Holter Database. http://physionet.org/physiobank/database/sddb/.

[11] MIT-BIH Normal Sinus Rhythm Database. http://physionet.org/physiobank/database/nsrdb/.

[12] Clifford GD, Sameni R. An artificial multi-channel model for generating abnormal electrocardiographic rhythms. Computers in Cardiology 2008;35.

[13] Abdi H. The Kendall rank correlation coefficient. In Salkind N (ed.), Encyclopedia of Measurement and Statistics. Thousand Oaks (CA): Sage, 2007; 1–7.

Address for correspondence:

George B. Moody
MIT E25-505A, Cambridge, MA 02139 USA
george@mit.edu