RR Interval Time Series Modeling: The PhysioNet/Computers in Cardiology Challenge 2002

GB Moody

Harvard-M.I.T. Division of Health Sciences and Technology
Cambridge, MA, USA

Abstract

Cardiac inter-beat (RR) interval time series contain fluctuations at time scales ranging from a few seconds to many hours. Realistic models of these series are potentially useful to researchers, not only as sources of surrogate data with known properties for evaluating novel analytic methods, but also as sources of insight into the diverse mechanisms underlying heart rate variability. PhysioNet and Computers in Cardiology have sponsored an open on-line competition aimed at stimulating the creation and exchange of high-quality models of RR interval variability, the third in an annual series of challenges for the research community. Participants first created software that was used to generate 24-hour synthetic time series, then attempted to identify the synthetic series within an unlabeled data set that included roughly equal amounts of real and synthetic data. All of the software models and the data used in the challenge are available at http://www.physionet.org/challenge/2002/.

1. Introduction

Heart rate variability has attracted much attention from researchers since the early 1980s. It has long been understood that a metronomic heart rate is pathological, and that the healthy heart is influenced by multiple neural and hormonal inputs that result in variations in inter-beat (RR) intervals, at time scales ranging from less than a second to 24 hours. Even after 20 years of study, new analytic techniques continue to reveal properties of the time series of RR intervals. Much research in this area aims to discover or to explain how specific changes in variability can be related to specific pathologies.

Given how much is known about heart rate variability, it might be thought that simulating a realistic sequence of RR intervals would be an easy task. The intricate interdependencies of variations at different scales, however, make it difficult to create a simulation of sufficient realism to mislead an experienced observer, and it may be even harder to deceive a program designed to quantify these subtle features.

Researchers interested in evaluating new analytic methods benefit when they have access to realistic models that can produce surrogate data with known properties, and to repositories of real data that can serve as a basis for comparison of disparate analyses. Models can also inform basic research when they provide insight into the nature and interactions of hidden mechanisms that may underlie observable phenomena. As a public research resource, PhysioNet aims to provide the research community with freely available data and software that support and frequently bootstrap innovative studies in physiology, biomedical engineering, and medical physics. Consistent with this goal, PhysioNet and Computers in Cardiology have jointly sponsored an annual series of open, on-line challenges[1, 2] designed to stimulate rapid progress on interesting research and clinical questions. The current challenge is the third of this series, and the first to focus explicitly on software models.

2. Organization of the challenge

The challenge was intended to motivate the development of realistic models of RR variability, encompassing fluctuations at all time scales up to and including those related to the 24-hour sleep-wake cycle. The major problem in designing the challenge was to devise an objective method for ranking the models with respect to realism.

The solution adopted for this challenge was to define two challenge events. Entrants in the first event submitted software capable of generating synthetic 24-hour RR interval time series; entrants in the second event classified a set of time series, including some generated by the software submitted for event 1, and others that were real time series. Scores for event 1 were determined by entries in event 2, and vice versa. All participants were required to enter event 2 (to insure that there would be a sufficient basis for ranking the RR interval generators submitted for event 1); event 1 was optional (to encourage the participation in event 2 of clinicians and others outside of the community of researchers who develop models).

0276–6547/02 $17.00 © 2002 IEEE

Computers in Cardiology 2002;29:125–128.

2.1. The reference data set

To encourage the participation of non-specialists in the challenge, a collection of real RR interval time series was provided for study (http://www.physionet.org/challenge/-2002/nsrdb-rr.tar.gz). This set contains the intervals from the 18 records of the MIT-BIH Normal Sinus Rhythm Database, which is also available on PhysioNet.

2.2. RR interval generators

A short example program (available at http://www.physionet.org/challenge/2002/rrgen.c) was provided as a framework for the RR interval generators to be entered in event 1. The example contains an initialization function (with a random seed) that sets any parameters required for a simulation, and a second function that returns the length of the next (simulated) RR interval. The rules of the challenge specified that participants write functions in standard (ANSI/ISO) C to replace these two functions in the example.

Participants were warned that each generator would be used to create two series in the challenge data set (see below), so that an important design constraint is that different initial random seeds should result in different outputs. Entries were allowed to define additional functions, global and local variables; to use other functions from the ANSI/ISO C standard library and math library; and to create temporary files in the current directory (which, however, did not persist between runs). Entries were not permitted to modify the main (control) function provided in the example; to write to the standard output; to change the current directory; to start another program or process; to incorporate real (physiologic) RR interval sequences in the output; or to include code or data that could not be made freely available after the conclusion of the challenge.

Seven teams entered event 1. Since multiple entries were permitted, eight generators were available for use in event 2.

2.3. The challenge data set

The challenge data set (http://www.physionet.org/challenge/2002/dataset.tar.gz) consists of 50 RR interval time series (see figure 1), each between 20 and 24 hours in length, presented in the same format as the reference data set.

Twenty-six of these were obtained by semi-automated analysis of long-term ECG recordings of adults between the ages of 20 and 50 who have no known cardiac abnormalities. This subset was designated as group A. Small numbers of ectopic beats are common in such recordings, as are short intervals of artifacts that may cause false beat detections or missed beat detections. Recordings with significant amounts of noise or ectopy were excluded. Group A was selected from the same population and with the same criteria as those represented in the reference data set, although it is apparent that the reference data set has a higher (though still clinically insignificant) incidence of ectopy than group A.

Each of ten generators (the eight generators entered into event 1, and two unofficial entries) was used to create two synthetic RR interval time series. This set of 20 series was designated as group B. Different random seeds were used to initialize the generators for each run. Two more unofficial generators (described below) were used to produce the last four series, which were designated as group C.

The lengths of the groups B and C series were determined randomly with a distribution that roughly matched that of the lengths of the group A series.

The synthetic and real series were assigned random identification numbers in the challenge data set, which was posted on PhysioNet on Wednesday, 24 April 2002, marking the start of event 2. The exact numbers of real and synthesized series were unknown to participants in event 2, who knew only that roughly equal numbers of real and synthesized series were present in the dataset, and that two series had been created by each generator.

2.4. Scoring

Entrants in event 2 classified each of the 50 series in the challenge data set as real (A), synthetic (B), or unknown (C). Each correct classification of a series in groups A or B earned 2 points, but each incorrect classification resulted in a 1-point penalty. Since the unofficial group C generators did not obey all of the rules of event 1, participants in event 2 were given 2 points for each of the four group C series, no matter how they were classified. Thus the highest possible score in event 2 was 100 points. Participants in event 2 were allowed to submit up to five entries. An autoscorer received entries submitted using a web browser, and returned scores by email to participants.

Scoring of event 1 was somewhat more complex than for event 2. The overall accuracy, a, of each event 2 participant is defined as the number of correct classifications made by that participant divided by the number of series to be classified. Based on a, we can define a weight

$$w = (a - 0.5)^2 + 0.05$$

to be given to that participant's classifications. Each group B series received $2w$ points for each "real" classification if $a > 0.5$, $2w$ points for each "synthetic" classification if $a <= 0.5$, and w points for each "unknown" classification. Each generator received a score that is the sum of the scores

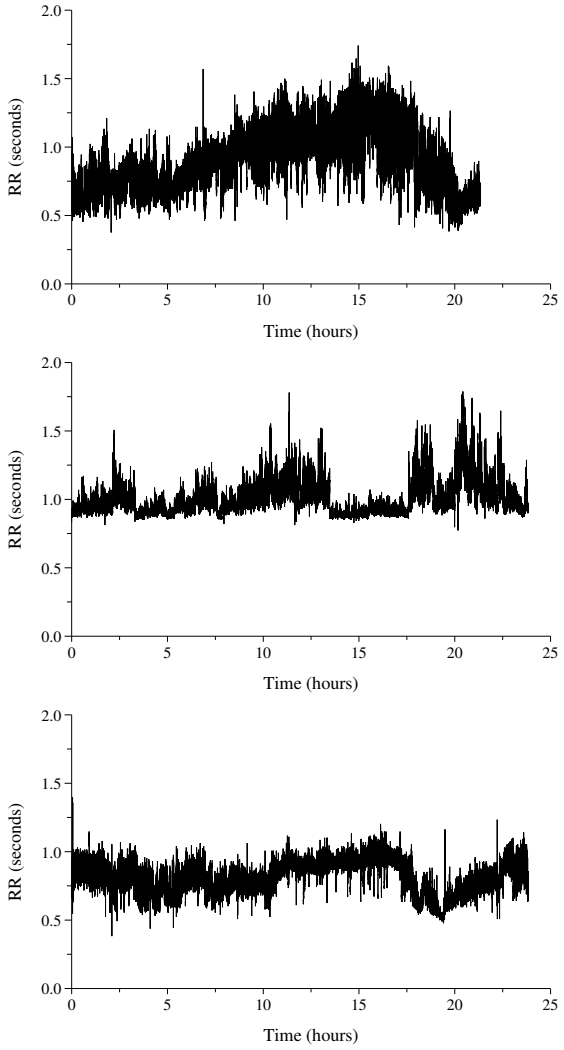

Figure 1. Three series from the challenge data set. The upper panel shows a series from group A (real data). See the last page for information about the center and lower panels.

for its two group B series. Participants who entered more than one generator received a separate score for each generator.

The weighting factor, w, was introduced so that a generator gets significantly more credit for misleading a really good classifier than for misleading one whose classifications are no better than random. The values of w are symmetric about $a = 0.5$ because a classifier who misclassifies everything is clearly able to tell the difference between real and synthesized data, despite a fundamental confusion about which is which. A small positive bias was added to w so that a series that consistently misleads poor classifiers receives a (small) score increment.

3. Results

The final scores for event 1 are summarized in table 1. The top-scoring generator, entered by DC Lin, created the series designated as rr10 and rr37, which were misclassified as real series respectively by 4 and 5 participants in event 2. The other generators attracted smaller numbers of misclassifications, but it is clear that most event 2 entrants were able to identify most of the group B series as synthetic, with almost no false positives (real series classified as synthetic). The details of most of the algorithms used by the generators entered in event 1 can be found in their authors' papers elsewhere in this volume, and are not repeated here.

The rules of event 1 required that generators not incorporate portions of real (physiologic) RR interval sequences in the output, and both of the group C generators violated this rule. The first of these, created by Mohammed Saeed of MIT, combined a model of short-term (beat-to-beat) fluctuations with long-term fluctuations that were determined from smoothed averages of real time series. Although the output of this generator is superficially realistic, most event 2 participants were able to identify its output as synthetic.

The second group C generator simply time-reversed an entire (real) 24-hour series. This outside-the-box strategy would have won event 1 decisively had it been permitted by the rules, since its outputs (rr14 and rr16) received 13 and 15 "real" classifications respectively (of the 17 official event 2 entries).

Six of seventeen entrants in event 2 achieved perfect scores of 100 points within the first week; in order of their entries, these were CC Yang, SH Yi and colleagues, E Bowers and colleagues, N Wessel, T Smuc, and H Malburg. Only two of these entrants needed a second attempt to achieve their perfect scores. Since the outcome of event 2 had been resolved rapidly, event 2 was stopped, with the agreement of the event 1 participants. (Unofficial entries in event 2 continue to be accepted and scored indefinitely, but they do not influence the event 1 results.)

4. Conclusions

The results suggest that it is quite difficult to design a an algorithm for synthesizing RR interval time series with sufficient realism to mislead a careful observer or a well-crafted classification algorithm.

The success of the unofficial time-reversal generator is somewhat surprising, since short-term time asymmetries in RR time series are well-known (e.g., compensatory pauses following ventricular ectopic beats) and might be expected to reveal the nature of these group C series. The input series were chosen because they did not exhibit any of these well-known phenomena, however; the surprise is that features sensitive to long-term asymmetries do not appear to be necessary in order to distinguish groups A and B perfectly.

Table 1. Final results for event 1 (generating RR interval series).

Score	*Entrant*
3.452	DC Lin Ryerson University, Toronto, Canada
1.494	D Gamberger, I Maric, T Smuc, G Bosanac, N Bogunovic, G Krstacic Rudjer Boskovic Institute, Institute for Cardiovascular Prevention and Rehabilitation Zagreb, Croatia
0.689	CC Yang, CH Chang, HW Yien Taipei Veterans General Hospital, School of Medicine, National Yang-Ming University Taipei, Taiwan
0.497	PE McSharry, GD Clifford Dept Maths & Dept Engineering, University of Oxford, UK
0.202	M Roy University of Michigan, Ann Arbor, Michigan, USA
0.202	MA Garcia-González, J Ramos-Castro Instrumentation and Bioengineering Division, Electronic Engineering Department Universitat Politècnica de Catalunya, Barcelona, Spain

It was originally hoped that group A would include pairs of time series obtained from each human subject, so that similarities between pairs of series would not immediately give away the group B series. It was not possible to obtain such pairs within the time available, however, and several event 2 participants commented that they were able to exploit inter-series similarities to simplify their classification task.

The task of simulating long-term RR variability related to the 24-hour sleep-wake cycle added considerably to the difficulty of participating in event 1. Over intervals on the order of an hour or less, the outputs of most of the generators are much more difficult to distinguish from real data than were the 20- to 24-hour series used in the challenge.

A major outcome of the challenge is that a diverse set of software RR interval generators that share a common interface is now available to the research community to support and stimulate future studies. Since all of the models are provided in C source form (at http://www.physionet.org/challenge/2002/), their workings can be studied, and elements of two or more models can be combined. This collection of algorithms will be an important resource for future investigations that require synthetic RR or heart rate time series, and for development of even more realistic generators.

Acknowledgements

PhysioNet/Computers in Cardiology Challenges are conducted using the facilities of PhysioNet, a public service of the Research Resource for Complex Physiologic Signals, which is supported by a grant from the National Center for Research Resources of the US National Institutes of Health (P41 RR13622). The author and his colleagues at PhysioNet thank Gerold Porenta and the board of Computers in Cardiology for their continuing and enthusiastic support of this series of challenges. Thanks also to Mohammed Saeed for an early suggestion that developed into the topic of this challenge, and for contributing two of the series in group C; to Ramakrishna Mukkalama, Roger Mark, C-K Peng, and Ary Goldberger for their suggestions and encouragement; to Diane Perry, Isaac Henry, and Joseph Mietus for their assistance in assembling the group A data; and to all those who participated in the challenge, many of whom report their results elsewhere in this volume. Finally, special thanks are due to those who have contributed their software models from event 1 for use by the research community at large: Dragan Gamberger and colleagues, MA Garcia-González and J Ramos-Castro, Philip Langley and colleagues, DC Lin, PE McSharry and GD Clifford, M Roy, and CC Yang.

In figure 1, the center and lower panels show series from groups B (synthetic data) and A (real data) respectively.

References

[1] Moody GB, Mark RG, Goldberger AL, Penzel T. Stimulating rapid research advances via focused competition: The Computers in Cardiology Challenge 2000. Computers in Cardiology 2000;27:207–210.

[2] Moody GB, Goldberger AL, McClennen S, Swiryn S. Predicting the onset of paroxysmal atrial fi brillation: The Computers in Cardiology Challenge 2001. Computers in Cardiology 2001;28:113–116.

Address for correspondence:

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