ARIC Manuscript Proposal # 1655

PC Reviewed: 6/8/10   Status: A   Priority: 2
SC Reviewed: _________   Status: _____   Priority: ____

1.a. Full Title: Sudden Cardiac Death Prediction from Multimodal Data

b. Abbreviated Title (Length 26 characters):

   SCD Prediction

2. Writing Group:

   Writing group members:

   Donald Geman
   Daniel Naiman
   Bruno Jedynak
   Wendy Post
   Dan Arking
   Alvaro Alonso

   I, the first author, confirm that all the coauthors have given their approval for this manuscript proposal. _DQN_ [please confirm with your initials electronically or in writing]

   First author: Daniel Naiman
   Address:

   Department of Applied Mathematics and Statistics
   Whiting School of Engineering
   Johns Hopkins University
   3400 North Charles St.
   Baltimore, MD 21218

   Phone: 410-516-7203   Fax: 410-516-7459
   E-mail: daniel.naiman@jhu.edu

   ARIC author to be contacted if there are questions about the manuscript and the first author does not respond or cannot be located (this must be an ARIC investigator).

   Name: Dan Arking
   Address:
3. **Timeline:**

Our aim is to complete the analysis by Dec 31, 2010.

4. **Rationale:**

We are currently leading a study to develop statistical tools for integrating available patient data from several modalities and scales in order to classify patients according to their risk of either sudden cardiac death (SCD) or overall death, and to make these tools available to researchers in the cardiovascular community for analyzing their own data. The aim of this proposal is to evaluate these methods for assessing the risk of death based on multi-modal data and to produce additional methodology relating various other outcomes e.g. CHD, high blood cholesterol, high blood pressure, diabetes to SCD.

Our team has developed new classification methodology using data from the Reynolds PROSE-ICD, which is a cohort of patients who have had implantable cardioverter defibrillators (ICDs) implanted as a primary prevention for SCD. The available features include ECG measurements, genotypes at various marker loci, proteomic analyses, MRI and CT images, age, gender, ethnicity, and results of additional clinical evaluation. It is evidently desirable to determine for which patients, if any, implantation can be safely avoided.

The main limitation we face is woefully insufficient data. Since we are doing multi-modal prediction, it is necessary to merge data on subjects from several different modalities, including genetic data (SNP), ECG, IMAGE, and PROTEOMIC measurements. Whereas data from at least one such modality is available for most of the subjects, there are only a small number of subjects with data for all modalities; consequently, missing data are "structural", i.e., more the rule than exception. More seriously, the numbers of sudden cardiac deaths among the subjects is relatively small, resulting in the need for us to use "overall death" as the endpoint phenotype.
We have developed a new decision-tree approach to addressing the structurally missing data problem. At each node in the tree, an additional third NA (missing) branch is added to address those observations for which the splitting rule cannot be applied, and in addition, the NA-node contains subjects for which the rule does apply. This method has the benefit of making use of all of the available observations for making a splitting decision at the NA-node.

We have applied the methodology to the Reynolds data but that data set is relatively small. As a result it would be of enormous benefit for us to have access to the ARIC data, for which there are 276 cases of SCD determined through adjudication and there is also extensive GWAS, clinical, ECG and imaging data. A data set of this size and diversity would allow us to properly test the methodology we have developed for multi-modal prediction with structurally missing data. In addition to using the ARIC data to develop predictors of SCD for the general population, we would determine which predictor variables are available in both studies (ARIC and Reynolds), build a decision tree using the Reynolds data and validate it using the ARIC data, and vice versa.

5. Main Hypothesis/Study Questions:

Can we move beyond establishing individual associations between predictor variables and outcomes to structured, multi-modal prediction applying modern high-dimensional classification methods in machine learning?

Can we use decision trees built from data for the Reynolds population for predicting risk of death or SCD to draw similar conclusions about the ARIC population, and vice versa.

Can we build an informative classification scheme for predicting outcomes related to SCD in a hierarchical fashion?

6. Design and analysis (study design, inclusion/exclusion, outcome and other variables of interest with specific reference to the time of their collection, summary of data analysis, and any anticipated methodologic limitations or challenges if present).

Study sample
All subjects who gave informed consent to be involved in genetic analyses will be included in this study.

Exposures of interest
We plan to use as exposures data in the following categories:
- Sociodemographic variables: age, gender, race
- Anthropometric variables: body mass index, height, waist circumference
- Biochemical variables: blood lipids (subtypes), glucose, insulin, creatinine
- Systolic and diastolic blood pressure
- Use of medication: antihypertensive meds, lipid lowering drugs, antidiabetic meds
Outcome

We will consider both overall death and SCD as the main outcomes of interest. Definition of SCD: All cases of fatal MI and fatal CHD (both inpatient and out of hospital) have been reviewed to determine if they meet criteria for SCD. Sudden cardiac death is defined as a sudden unexpected death that appears to be related to an arrhythmic etiology. Cases of SCD are being identified as definite, or possible, which includes cases complicated by other co-morbidities, such as ESRD, CHF or liver failure. For these cases, the patient must have been clinically stable prior to a sudden cardiac arrest. For all events, the individual must have been seen alive within 24 hours and had symptoms for less than one hour.

Analysis

We will build classification trees for predicting death and SCD using the exposure/feature data described above. Classifiers will produce estimates of conditional probabilities of death SCD given a subject’s profile. These will employ recently developed tools by the proposers for building classification trees in the presence of censored survival and structurally missing data. It is anticipated that many of the available features for predicting death and SCD will be highly correlated. This is certainly the case for the considerable number of electrocardiogram parameters. Thus, some strategies for reducing the feature set to a more manageable size while continuing to capture most of the predictive capabilities will be required. To address this issue, we will introduce novel feature selection tools such as those based on conditional mutual information into the analysis.

Various statistical learning methods have been developed for addressing the missing data issue. These are discussed in the references below.

The classification tree method we have developed for addressing the missing data issue, builds on the standard classification tree methodology for predicting a binary response variable $Y$ from training data consisting of observations with feature variables $x_1, ..., x_k$ and the (known) value of $Y$. Like their non-missing data counterparts, trees are built recursively with all observations included at the root node of the decision tree. The building process consists of attempting to identify a suitable rule for splitting each node to create three child nodes, followed by moving observations at the node along to child nodes. A splitting rule is a pair $(x, (L, R))$ consisting of a feature $x$ and a partitioning of the range of its values into two components $L$ and $R$. For such a pair, the observations at the node fall into two groups according to which of the two conditions $x \in L$ or $x \in R$ is satisfied. A pair is considered optimal if it is the one for which the conditional distribution of the response variable $Y$ for observations at the node, is as different as possible depending on which of the two conditions holds. When an optimal pair is
found, three children of the current node, labeled $L$, $R$, and $NA$, are created. The observations for which $x_i$ is available at the node, are passed along to the appropriate child node according to which condition $x_i \in L$ or $x_i \in R$ holds. In addition, all observations are passed to the NA node, including the observations for which feature $x_i$ is not available (hence the NA label for the child). A node is not split when either there are insufficiently many observations in it, or there is not a sufficiently informative splitting rule available. This node then has no children and is referred to as a leaf. Splitting continues until each observation has been the part of a sufficiently number of non-NA splitting decisions.

References


7.a. Will the data be used for non-CVD analysis in this manuscript?  ____ Yes  _X_ No

b. If Yes, is the author aware that the file ICTDER03 must be used to exclude persons with a value RES_OTH = “CVD Research” for non-DNA analysis, and for DNA analysis RES_DNA = “CVD Research” would be used?  ____ Yes  ____ No

(This file ICTDER03 has been distributed to ARIC PIs, and contains the responses to consent updates related to stored sample use for research.)

8.a. Will the DNA data be used in this manuscript?  _X_ Yes  ____ No

8.b. If yes, is the author aware that either DNA data distributed by the Coordinating Center must be used, or the file ICTDER03 must be used to exclude those with value RES_DNA = “No use/storage DNA”?  _X_ Yes  ____ No
9. The lead author of this manuscript proposal has reviewed the list of existing ARIC Study manuscript proposals and has found no overlap between this proposal and previously approved manuscript proposals either published or still in active status. ARIC Investigators have access to the publications lists under the Study Members Area of the web site at: http://www.cscc.unc.edu/ARIC/search.php

   x   Yes     No

One of our key goals is to investigate death and SCD in conjunction with other possible outcome variables. We are attempting to determine whether a broad set of methods can be useful to the cardiovascular community, and to develop additional such tools. Consequently, it is possible that some particular outcome variable for which we develop predictors as a byproduct of our analytic methods will turn out to be one that is currently under investigation by other ARIC investigators, and we recognize the potential for overlap with other ARIC work. We will work out a plan to prevent such overlap, and, as our work progresses, we will welcome additional investigators who are studying SCD and related outcome variables to join our team.

10. What are the most related manuscript proposals in ARIC (authors are encouraged to contact lead authors of these proposals for comments on the new proposal or collaboration)?

11. a. Is this manuscript proposal associated with any ARIC ancillary studies or use any ancillary study data?   x Yes     No

   2005.07, 2006.03, 2008.09,

11.b. If yes, is the proposal

   A. primarily the result of an ancillary study (list number* _________)

   B. primarily based on ARIC data with ancillary data playing a minor role (usually control variables; list number(s)* __________

*ancillary studies are listed by number at http://www.cscc.unc.edu/aric/forms/

12. Manuscript preparation is expected to be completed in one to three years. If a manuscript is not submitted for ARIC review at the end of the 3-years from the date of the approval, the manuscript proposal will expire.