```
Publication Request:
   >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
   This file describes the contents of the heart-disease directory.

   This directory contains 4 databases concerning heart disease diagnosis.
   All attributes are numeric-valued.  The data was collected from the
   four following locations:

     1. Cleveland Clinic Foundation (cleveland.data)
     2. Hungarian Institute of Cardiology, Budapest (hungarian.data)
     3. V.A. Medical Center, Long Beach, CA (long-beach-va.data)
     4. University Hospital, Zurich, Switzerland (switzerland.data)

   Each database has the same instance format.  While the databases have 76
   raw attributes, only 14 of them are actually used.  Thus I've taken the
   liberty of making 2 copies of each database: one with all the attributes
   and 1 with the 14 attributes actually used in past experiments.

   The authors of the databases have requested:

      ...that any publications resulting from the use of the data include the
      names of the principal investigator responsible for the data collection
      at each institution.  They would be:

       1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
       2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
       3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
       4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation:
          Robert Detrano, M.D., Ph.D.

   Thanks in advance for abiding by this request.

   David Aha
   July 22, 1988
   >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

1. Title: Heart Disease Databases

2. Source Information:
   (a) Creators:
       -- 1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
       -- 2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
       -- 3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
       -- 4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation:
             Robert Detrano, M.D., Ph.D.
   (b) Donor: David W. Aha (aha@ics.uci.edu) (714) 856-8779
   (c) Date: July, 1988

3. Past Usage:
    1. Detrano,~R., Janosi,~A., Steinbrunn,~W., Pfisterer,~M., Schmid,~J.,
       Sandhu,~S., Guppy,~K., Lee,~S., \& Froelicher,~V. (1989).  {\it
       International application of a new probability algorithm for the
       diagnosis of coronary artery disease.}  {\it American Journal of
       Cardiology}, {\it 64},304--310.
       -- International Probability Analysis
       -- Address: Robert Detrano, M.D.
                   Cardiology 111-C
                   V.A. Medical Center
                   5901 E. 7th Street
                   Long Beach, CA 90028
       -- Results in percent accuracy: (for 0.5 probability threshold)
             Data Name:  CDF    CADENZA
          -- Hungarian   77     74
             Long beach  79     77
```

```
            Swiss       81     81
         -- Approximately a 77% correct classification accuracy with a
            logistic-regression-derived discriminant function
    2. David W. Aha & Dennis Kibler
       --


         -- Instance-based prediction of heart-disease presence with the
            Cleveland database
            -- NTgrowth: 77.0% accuracy
            --       C4: 74.8% accuracy
    3. John Gennari
       -- Gennari, J.~H., Langley, P, \& Fisher, D. (1989). Models of
          incremental concept formation. {\it Artificial Intelligence, 40},
          11--61.
       -- Results:
          -- The CLASSIT conceptual clustering system achieved a 78.9% accuracy
             on the Cleveland database.

 4. Relevant Information:
     This database contains 76 attributes, but all published experiments
     refer to using a subset of 14 of them.  In particular, the Cleveland
     database is the only one that has been used by ML researchers to
     this date.  The "goal" field refers to the presence of heart disease
     in the patient.  It is integer valued from 0 (no presence) to 4.
     Experiments with the Cleveland database have concentrated on simply
     attempting to distinguish presence (values 1,2,3,4) from absence (value
     0).

     The names and social security numbers of the patients were recently
     removed from the database, replaced with dummy values.

     One file has been "processed", that one containing the Cleveland
     database.  All four unprocessed files also exist in this directory.

 5. Number of Instances:
        Database:    # of instances:
          Cleveland: 303
          Hungarian: 294
        Switzerland: 123
      Long Beach VA: 200

 6. Number of Attributes: 76 (including the predicted attribute)

 7. Attribute Information:
    -- Only 14 used
       -- 1. #3  (age)
       -- 2. #4  (sex)
       -- 3. #9  (cp)
       -- 4. #10 (trestbps)
       -- 5. #12 (chol)
       -- 6. #16 (fbs)
       -- 7. #19 (restecg)
       -- 8. #32 (thalach)
       -- 9. #38 (exang)
       -- 10. #40 (oldpeak)
       -- 11. #41 (slope)
       -- 12. #44 (ca)
       -- 13. #51 (thal)
       -- 14. #58 (num)       (the predicted attribute)

    -- Complete attribute documentation:
       1 id: patient identification number
       2 ccf: social security number (I replaced this with a dummy value of 0)
       3 age: age in years
```

```
 4 sex: sex (1 = male; 0 = female)
 5 painloc: chest pain location (1 = substernal; 0 = otherwise)
 6 painexer (1 = provoked by exertion; 0 = otherwise)
 7 relrest (1 = relieved after rest; 0 = otherwise)
 8 pncaden (sum of 5, 6, and 7)
 9 cp: chest pain type
   -- Value 1: typical angina
   -- Value 2: atypical angina
   -- Value 3: non-anginal pain
   -- Value 4: asymptomatic
10 trestbps: resting blood pressure (in mm Hg on admission to the
   hospital)
11 htn
12 chol: serum cholestoral in mg/dl
13 smoke: I believe this is 1 = yes; 0 = no (is or is not a smoker)
14 cigs (cigarettes per day)
15 years (number of years as a smoker)
16 fbs: (fasting blood sugar > 120 mg/dl)  (1 = true; 0 = false)
17 dm (1 = history of diabetes; 0 = no such history)
18 famhist: family history of coronary artery disease (1 = yes; 0 = no)
19 restecg: resting electrocardiographic results
   -- Value 0: normal
   -- Value 1: having ST-T wave abnormality (T wave inversions and/or ST
               elevation or depression of > 0.05 mV)
   -- Value 2: showing probable or definite left ventricular hypertrophy
               by Estes' criteria
20 ekgmo (month of exercise ECG reading)
21 ekgday(day of exercise ECG reading)
22 ekgyr (year of exercise ECG reading)
23 dig (digitalis used furing exercise ECG: 1 = yes; 0 = no)
24 prop (Beta blocker used during exercise ECG: 1 = yes; 0 = no)
25 nitr (nitrates used during exercise ECG: 1 = yes; 0 = no)
26 pro (calcium channel blocker used during exercise ECG: 1 = yes; 0 = no)
27 diuretic (diuretic used used during exercise ECG: 1 = yes; 0 = no)
28 proto: exercise protocol
     1 = Bruce
     2 = Kottus
     3 = McHenry
     4 = fast Balke
     5 = Balke
     6 = Noughton
     7 = bike 150 kpa min/min  (Not sure if "kpa min/min" is what was
         written!)
     8 = bike 125 kpa min/min
     9 = bike 100 kpa min/min
    10 = bike 75 kpa min/min
    11 = bike 50 kpa min/min
    12 = arm ergometer
29 thaldur: duration of exercise test in minutes
30 thaltime: time when ST measure depression was noted
31 met: mets achieved
32 thalach: maximum heart rate achieved
33 thalrest: resting heart rate
34 tpeakbps: peak exercise blood pressure (first of 2 parts)
35 tpeakbpd: peak exercise blood pressure (second of 2 parts)
36 dummy
37 trestbpd: resting blood pressure
38 exang: exercise induced angina (1 = yes; 0 = no)
39 xhypo: (1 = yes; 0 = no)
40 oldpeak = ST depression induced by exercise relative to rest
41 slope: the slope of the peak exercise ST segment
   -- Value 1: upsloping
   -- Value 2: flat
   -- Value 3: downsloping
42 rldv5: height at rest
```

```
      43 rldv5e: height at peak exercise
      44 ca: number of major vessels (0-3) colored by flourosopy
      45 restckm: irrelevant
      46 exerckm: irrelevant
      47 restef: rest raidonuclid (sp?) ejection fraction
      48 restwm: rest wall (sp?) motion abnormality
         0 = none
         1 = mild or moderate
         2 = moderate or severe
         3 = akinesis or dyskmem (sp?)
      49 exeref: exercise radinalid (sp?) ejection fraction
      50 exerwm: exercise wall (sp?) motion
      51 thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
      52 thalsev: not used
      53 thalpul: not used
      54 earlobe: not used
      55 cmo: month of cardiac cath (sp?)  (perhaps "call")
      56 cday: day of cardiac cath (sp?)
      57 cyr: year of cardiac cath (sp?)
      58 num: diagnosis of heart disease (angiographic disease status)
         -- Value 0: < 50% diameter narrowing
         -- Value 1: > 50% diameter narrowing
         (in any major vessel: attributes 59 through 68 are vessels)
      59 lmt
      60 ladprox
      61 laddist
      62 diag
      63 cxmain
      64 ramus
      65 om1
      66 om2
      67 rcaprox
      68 rcadist
      69 lvx1: not used
      70 lvx2: not used
      71 lvx3: not used
      72 lvx4: not used
      73 lvf: not used
      74 cathef: not used
      75 junk: not used
      76 name: last name of patient
         (I replaced this with the dummy string "name")

 9. Missing Attribute Values: Several.  Distinguished with value -9.0.

 10. Class Distribution:
         Database:      0   1   2   3   4 Total
           Cleveland: 164  55  36  35  13   303
           Hungarian: 188  37  26  28  15   294
         Switzerland:   8  48  32  30   5   123
       Long Beach VA:  51  56  41  42  10   200
```