Project CardI-HACK

Building the challenge database

Important : This document details the content of the columns in the database used for the CardI-HACK data challenge.
For your information, the database consists of a synthetic dataset reconstructed for 1000 patients, built from the data of several hundred patients.
However, the representativeness/usability of the dataset is guaranteed.

The database consists of 4 parts.

- One section contains baseline clinical data, which describes the patient and will be used to predict the two outcomes sought in this challenge.
- A genetic component that also describes the patient at baseline is also used to predict the two outcomes sought in this challenge
- Two columns labeled "outcome"
    - Severity at a binary "baseline", to be predicted using a fraction of the data
    - Occurrence of major clinical events during follow-up after baseline in 4 classes, to be predicted using various baseline data (clinical and genetic)

Part 1 – Baseline Clinical Data

Nine variables were chosen at baseline for their high completeness in the initial database. They allow for a very precise and relevant definition of patients.

"ID" - unit: no unit

Automatically generated random identifier, does not correspond to any other database.

"Age_at_baseline" – unit: year

This refers to the age at which the patient begins their care in the department , with a significant battery of tests. (This is important to note, as care sometimes follows initial cardiac clinical events that highlight the existence of the pathology in the patient).

"Age_at_diag" – unit: year

This is the age at which ***the diagnosis of HCM is made*** ; it may be different from that of the baseline (Age_Baseline column).

"BMI": Body Mass Index - unit: no unit

The Body Mass Index (BMI) is a tool used to estimate a person's body size and degree of overweight.

"BSA": Body Surface Area - unit: m^2

Body surface area (BSA) is an estimate of the total surface area of skin covering the human body.

It is used in medicine to index various parameters such as cardiac output, but also to adjust drug doses, particularly in oncology (chemotherapy), to assess nutritional needs or basal metabolism, or to estimate the burned surface area in the case of burns.

"Genre" - unit: no unit

Patient's gender. Information to define which gender corresponds to 1 and which gender corresponds to 2 is not revealed.

"Max Thickness" - unit: millimeter

This is the maximum measured value of the left ventricle wall (in red). The normal left ventricle wall thickness is measured at end-diastole. The larger the value of this variable, the more pathological the condition is for the patient.

“Gradient” – BINARY 0/1

The **left ventricular (LV) gradient** , often measured by Doppler echocardiography, is the pressure difference between the left ventricle and the aorta during systole. This gradient is particularly important in the context of hypertrophic obstructive cardiomyopathy (HCM) => the challenge pathology.

These values are used to assess the severity of the obstruction and to guide treatment, which may include medication or invasive procedures such as septal ablation (see below). In this column, zero represents a normal value, and 1 represents a pathological value.

“TVNS” – BINARY 0/1

Nonsustained ventricular tachycardia (NSVT) is a type of cardiac arrhythmia characterized by episodes of rapid and often irregular ventricular beats that last no more than 30 seconds. These episodes can occur during or outside of physical exertion and cause the heart to beat at more than 100 beats per minute. However, this arrhythmia may be asymptomatic or cause palpitations, dizziness, or even faintness.

In our database, 0 corresponds to the absence of TVNS, and 1, the presence of TVNS in the patient.

“ FEVG” – BINARY 0/1

Left ventricular ejection fraction (LVEF) is a key parameter for assessing cardiac function during systole. It represents the percentage of blood ejected by the left ventricle with each contraction. In other words, it measures how efficiently the heart pumps blood throughout the body.

In our database, 0 corresponds to a normal LVEF (≥50%) and 1 to an abnormal frequency (<50%).

“ATCD_MS”: History of sudden death in relatives – BINARY 0/1

This column reports the history of sudden death among the patient's relatives (parents, grandparents, uncles/aunts, brothers/sisters, children) who died before the age of 40.

In our database, 0 corresponds to no deceased person among relatives, and 1, at least one relative who died suddenly before the age of 40.

“SYNCOPE” – BINARY 0/1

This column reports the patient's history of syncope in the two years preceding the baseline.

In our database, 0 corresponds to the absence of syncope, and 1, at least one syncope in the two years preceding the baseline.

“Diameter_OG” - millimeter

This refers to the diameter of the left atrium. The left atrium (LA) plays a central role in cardiac mechanics. Its systole contributes approximately 20% to ventricular filling. It modulates left ventricular filling pressures. The left atrium is often described as a barometer of left ventricular (LV) diastolic function in various pathologies, as it is directly exposed to LV diastolic pressure during mitral valve opening. However, its role in valvular heart disease, cardiomyopathies, and supraventricular arrhythmias is just as important as in heart failure with preserved ejection fraction.

Part 2 – Genetic data

Part 2 contains genetic data that we ask you to use primarily to predict the outcome. Genetics is the science that studies the inheritance and transmission of biological traits from one generation to the next. It encompasses simple mechanisms, such as those described by Mendel, but also complex interactions between multiple genes and the environment, as in multifactorial diseases .

The database contains, on the one hand, a column from Mendelian genetics (which is based on the laws of heredity, concerns rare genetic variants that are pathogenic or probably pathogenic, with simple modes of transmission: Dominant: a single copy of the mutated gene is sufficient for the trait to appear, Recessive: two copies of the mutated gene are necessary).

The database also contains multiple columns from multifactorial genetics (which is based on frequent genetic variants, also called genetic polymorphisms and usually of the SNP single nucleotide polymorphism type mechanism), and chosen because of their interest in cardiomyopathies or in the physiological function of the left ventricle.

“Variant.Pathogene” – no unit

This information was obtained after sequencing the genes (usually sarcomeric) involved in hypertrophic cardiomyopathy.

0 indicates that no mutations with a high level of pathogenicity (pathogenic or probably pathogenic) were found in the patient.

1 indicates that a mutation with a high level of pathogenicity has been found in the patient

2 indicates that several mutations with a high level of pathogenicity were found in this patient.

"SNP *x* "

This information was obtained after genome-wide association study (GWAS) and selection of polymorphisms of potential interest in hypertrophic cardiomyopathy.

Part 2 consists of two sub-parts, a set of 75 SNP polymorphisms to be considered as a priority (SNP1 to SNP75), and 213 other SNP polymorphisms (SNP76 to SNP288), to be considered later or optionally, if they improve the prediction of the Outcome.

OUTCOME #1 –

OUTCOME SEVERITY: Characterization of patient severity at Baseline

This outcome characterizes the baseline level of pathology severity in each patient. Participants are asked to predict whether a patient was classified as non-severe or severe at baseline.
In our database,

- 0 corresponds to a non-severe patient and
- 1 to a patient presenting at least one severity criterion.

The variable OUTCOME SEVERITY was derived from the presence of one or more of the following clinical features at baseline:

- Abnormal LVEF (<50%)
- History of unplanned hospitalization for heart failure
- History of atrial fibrillation
- History of atrial flutter
- History of stroke (cerebrovascular accident) or TIA (transient ischemic attack)
- History of sudden death recovered
- History of appropriate therapeutic shock in patients with implantable defibrillator
- History of septal alcohol ablation (interventional treatment of obstruction or gradient)
- History of myomectomy (surgical treatment of obstruction or gradient)

Constraint for predicting Outcome Severity:

The outcome should only be predicted with baseline age, patient sex, and genetic data (Mendelian and multifactorial).

OUTCOME #2 –
OUTCOME MACE: Occurrence of major clinical events

This outcome represents the occurrence and timing of major adverse cardiovascular events during follow-up (after the "baseline"), which you are asked to predict as a priority.

In our database,
- 0 corresponds to **the absence** of any major clinical events during the patient follow-up period.
- 1 corresponds to the occurrence of a **major clinical event within 1000 days** after baseline
- 2 corresponds to the occurrence of a **major clinical event within 3000 days** after the baseline

The OUTCOME MACE variable was derived from the occurrence of at least one of the following events during follow-up:

- Death from cardiovascular or cardiac causes during follow-up
- Unplanned hospitalization for heart failure during follow-up
- Sustained Ventricular Tachycardia or Sustained Ventricular Fibrillation during follow-up
- Atrial fibrillation during follow-up
- Atrial flutter during follow-up
- Stroke/TIA during follow-up
- Sudden Death recovered during follow-up
- Appropriate shock during follow-up in defibrillator wearers
- Heart transplantation during follow-up

The time to occurrence of each event was taken into account and the first event occurring during the follow-up was considered, and the date considered to introduce the temporality of occurrence of the event during the post-baseline follow-up, according to two time periods as described above.

Constraint for predicting the outcome of a major clinical event:

The outcome can be predicted with the baseline data BUT must necessarily include **at least one of the priority SNPs, and with a maximum of 100 SNPs in total in the final proposed model (supplementary list included).**
Predicting this outcome is a priority.

Additional constraint

The algorithm must be able to handle missing data.