Masterclass in AI and Cardiovascular Health

**Data-Driven Cardiovascular Care: Using AI and Big Data to Save Lives**

Hands-on Workshop 1: Use of large-language models in clinical cardiology

Date: 27 June 2025

Time: 11:30 - 12:30

## Welcome to this hands-on workshop!

Today, you’ll be introduced to an AI-powered virtual assistant designed to help you explore and analyze clinical data. **No coding** or data science background is required.

We’ll be working with a **synthetic cardiology dataset** that includes a variety of clinical parameters related to cardiovascular health. Using simple language and natural questions, you’ll learn how to interact with the AI assistant to uncover patterns, gain insights, and better understand the data.

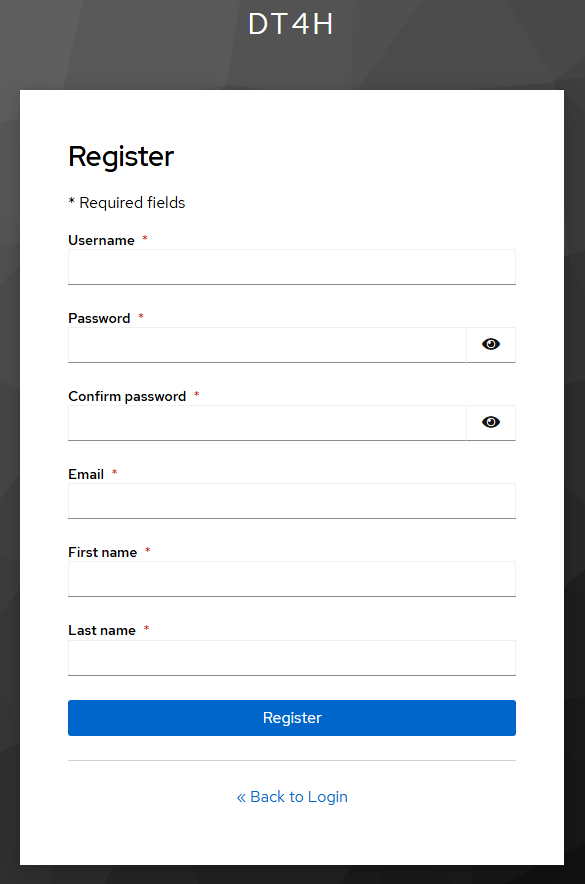
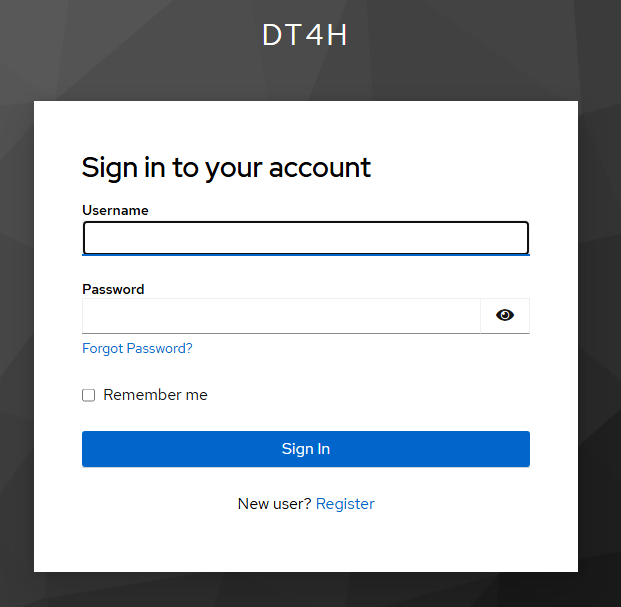
By the end of this session, you’ll feel confident using the assistant to:

* Ask meaningful questions about the dataset
* Identify key clinical factors and trends
* Interpret the AI’s responses in a clinical context

## Registration

Before you begin interacting with the cardiology AI assistant, you need to register for access.

* Open the AI virtual assistant page located @ <https://chomsky.ilsp.gr:8643/>
* Click on “Register” link just below the “Sign In” button
* Fill in the registration form, and click register
* After registration, you will be automatically redirected to the user interface of the AI virtual assistant



## Synthetic Clinical Data

In this workshop, we will be working with a **synthetic clinical dataset** modeled after real patient records from individuals admitted for **heart failure** at the **Amsterdam University Medical Centers (AUMC)**. This data has been **synthetically generated**, meaning it does not contain any actual patient-identifiable information, but retains the **statistical properties and structure** of the original dataset.

### What is Synthetic Clinical Data?

Synthetic data is artificially generated data that simulates the characteristics of real-world data. In the clinical context, it replicates important relationships and distributions seen in real patient data—such as comorbidities, lab values, and treatment outcomes—without exposing any real patient’s private health information.

### Why Use Synthetic Data?

There are several compelling reasons to use synthetic data in healthcare training and AI development:

* **Privacy preservation**: Since synthetic data is not linked to real patients, it complies with data protection regulations such as GDPR and HIPAA.
* **Safe experimentation**: It allows clinicians, data scientists, and students to explore datasets, build tools, and test hypotheses without ethical concerns or legal restrictions.
* **Reproducibility**: Synthetic datasets can be shared openly, making it easier to reproduce analyses and collaborate across institutions.

## Dataset Overview

The dataset used in this workshop contains **23 columns**, each representing a **clinical variable** related to patients who were admitted to the hospital due to **heart failure**. Each **row** in the dataset corresponds to a **single patient admission**, capturing key details about their condition, risk factors, and clinical findings.

In other words:

* **Columns** = variables such as age, blood pressure, laboratory results, medications, and outcomes.
* **Rows** = individual patient records.

The table below provides an overview of the 23 variables included in the dataset:

|  |  |  |
| --- | --- | --- |
| Column Name | Description | Type |
| sex | Sex | Binary 0=Female and 1=Male |
| age | Age | Numeric |
| rf\_smoking | Smoking status | Binary 0=No and 1=Yes |
| bmi | Body mass index | Numeric |
| sbp | Systolic blood pressure | Numeric |
| dbp | Diastolic blood pressure | Numeric |
| hdl\_cholesterol | High-density lipoprotein cholesterol level | Numeric |
| total\_cholesterol | Total cholesterol level | Numeric |
| creatinine | Serum creatinine level | Numeric |
| ace | Use of ACE inhibitors | Binary 0=No and 1=Yes |
| beta | Use of beta-blockers | Binary 0=No and 1=Yes |
| t2dm\_threshold | Type 2 diabetes mellitus status | Binary 0=No and 1=Yes |
| chronic\_ischemic\_hd | Chronic ischemic heart disease | Binary 0=No and 1=Yes |
| acute\_mi | Acute myocardial infarction | Binary 0=No and 1=Yes |
| af | Atrial fibrillation | Binary 0=No and 1=Yes |
| ht\_medication | Use of antihypertensive | Binary 0=No and 1=Yes |
| diabetes\_threshold | Diabetes status | Binary 0=No and 1=Yes |
| hf\_threshold | Heart failure status | Binary 0=No and 1=Yes |
| fam\_history | Family history of heart disease | Binary 0=No and 1=Yes |
| ht\_threshold | Hypertension status | Binary 0=No and 1=Yes |
| arthritis\_threshold | Arthritis status | Binary 0=No and 1=Yes |
| townsend\_deprivation\_index | Socioeconomic deprivation index | Numeric |
| rf\_ethnicity | Ethnicity | Categorical (6 categories, from 0 to 5) |

## Important

While the tool is designed to be user-friendly, there may be occasional errors or instances where the AI becomes unresponsive. If that happens, you can simply refresh the page or start a new chat conversation to continue.

## Icon Legend

 — Represents a query or prompt that you can copy and paste directly into the AI virtual assistant to get results quickly.

## Step 1: Familiarize Yourself with the Dataset

Objective: Get to know the dataset’s structure, variables, data types, and size. This helps you understand what information is available and how it is organized.

**Tips:**

* Use clear, simple sentences.
* Ask about specific columns to get focused answers.
* You can request summary statistics like mean, median, or counts.
* If you like, you can request the AI virtual assistant to comment on summary statistics.
* **View the first patient**

Show me the data of the first patient

* **Get the number of rows (patients) in the dataset**  
  How many patient records are there in total?
* **List all variables (columns) with their descriptions**  
  List all variables in the dataset with their descriptions
* **Check variable types**  
   Create a table of the available columns and provide little description on each column including the data type (numeric, categorical)
* **Check for missing data**  
  Are there any missing values in the dataset? If yes, which columns have them?
* **Basic summary statistics for some variables**  
  Provide summary statistics (mean, median, std) for age, BMI, and blood pressure.

How many smokers are there in the dataset and what is the prevalence?

Does the dataset contain more male or female patients?

What is the distribution of HDL values?

What is the distribution of total cholesterol?

List the number of patients taking ACE inhibitors and beta-blockers, and those taking both medications simultaneously.

What is the prevalence of smokers among males and females?

## Step 2: Exploring Risk Factors

**Objective:** Understand how different risk factors relate to each other.

 What is the prevalence of chronic ischemic heart disease in smokers vs non-smokers?

 Show the average HDL cholesterol levels for patients with and without atrial fibrillation.

Compare blood pressure values for patients with and without hypertension medication.

What is the average Townsend deprivation index for patients with acute myocardial infarction?

Summarize the impact of smoking and family history of HF on AF.

## Step 3: Clinical Parameter Correlations

**Objective:** Understand the relationship between the risk factors (variables).

Is there a correlation between BMI and systolic blood pressure?

Are there differences in creatinine levels across ethnic groups?

How do arthritis and hypertension correlate?

What is the relationship between age and diastolic blood pressure?

What is the relationsjip between age and systolic blood pressure for patients with diabetes?

## Step 4: Risk calculation

Objective: Explore how the AI virtual assistant may help in calculating cardiovascular related risk scores.

### Framingham risk score

There are different versions of the Framingham risk to calculate 10-year CHD risk. We need to specify to the AI assistant which variables (columns) it should use to calculate the risk. Furthermore, we need to instruct the AI assistant on how to calculate the risk score. Although Large Language Models (LLMs) have demonstrated excellent performance across many tasks involving natural language processing, they inherently may not work properly for calculations.

We will try two approaches to calculate a risk score for a particular patient in our dataset. One by providing the variables we want to involve in the calculation and we let the AI virtual assistant do the calculation based on its internal knowledge about the Framingham score (i.e., finding out the formula to calculate the points for each risk factor). The second approach is to provide further instructions to the LLM on how to calculate the risk using fictional risk score (not Framingham).

* 1. **Let’s try first the Framingham risk score, and see if we get correct results**

**Please execute the following queries sequentially**

Extract the data of the first patient.

Calculate the Framingham risk score given the following variables for the first patient: Age, Sex, Total Cholesterol, HDL Cholesterol, Systolic Blood Pressure (SBP) Treatment for Hypertension, and Smoking Status. Show your calculation in details.

**Compare the correctness of the calculate score (points) with the one you can obtain by using the following web calculator:**

[**https://www.mcw.edu/calculators/coronary-heart-disease-risk**](https://www.mcw.edu/calculators/coronary-heart-disease-risk)

* 1. **Let’s now create our own risk score calculation criteria. We will use the same variables mentioned above. In this approach, we will provide the AI assistant extra instructions on how to calculate the risk.**
* **The table below will be the criteria to score the risk factors for a particular outcome:**

| **Variable** | **Criteria** | **Points** |
| --- | --- | --- |
| Age (years) | 30-39 | 1 |
|  | 40-49 | 2 |
|  | 50-59 | 3 |
|  | 60+ | 4 |
| Sex | Male | 2 |
|  | Female | 0 |
| Total Cholesterol | < 5 mmol/L | 0 |
|  | ≥ 5 mmol/L | 2 |
| HDL Cholesterol | ≥ 1.3 mmol/L | 0 |
|  | < 1.3 mmol/L | 1 |
| Systolic Blood Pressure | < 140 mmHg | 0 |
|  | ≥ 140 mmHg | 2 |
| Treatment for Hypertension | No | 0 |
|  | Yes | 2 |
| Smoking Status | No | 0 |
|  | Yes | 2 |

**Please execute the following queries sequentially (please copy the entire query including the table)**

Extract the data of the first patient.

Use the following criteria to calculate a risk score using the data you’ve just extracted:

| **Variable** | **Criteria** | **Points** |
| --- | --- | --- |
| Age (years) | 30-39 | 1 |
|  | 40-49 | 2 |
|  | 50-59 | 3 |
|  | 60+ | 4 |
| Sex | Male | 2 |
|  | Female | 0 |
| Total Cholesterol | < 5 mmol/L | 0 |
|  | ≥ 5 mmol/L | 2 |
| HDL Cholesterol | ≥ 1.3 mmol/L | 0 |
|  | < 1.3 mmol/L | 1 |
| Systolic Blood Pressure | < 140 mmHg | 0 |
|  | ≥ 140 mmHg | 2 |
| Treatment for Hypertension | No | 0 |
|  | Yes | 2 |
| Smoking Status | No | 0 |
|  | Yes | 2 |

**How accurate are the results? You may ask the AI assistant to recalculate or confirm the results.**

## Step 5: Cohort selection

Objective: Filter patients on certain criteria’s, simulating the inclusion and exclusion criteria commonly used in clinical trials.

Identify patients older than 60 who have high systolic blood pressure and are on diabetes. Then, create a table describing their characteristics and risk factors from the dataset.

 Find patients with a family history of heart disease who are smokers and have a BMI over 30. Then create a table with summary statistics for these individuals where you describe SBP. DBP, creatinine, Sex and Age.

## Step 6: Natural Language Summary Generation from Patient Data

Objective: Demonstrate how the AI assistant can generate clinical summaries in natural language based on structured patient data. This helps clinicians quickly interpret and communicate patient profiles without needing to manually synthesize rows of data.

**Potential applications in the clinic:**

* Clinical Summaries: Quickly generate overviews of a patient’s risk factors and health status.
* Discharge Note Drafts: Automatically create structured narratives for diagnoses, treatments, and follow-up.
* Patient Education: Provide clear, plain-language explanations of lab results and care plans.
* Triage & Referrals: Summarize clinical reasons for referral based on patient data.

---

Extract the data of the third patient in the dataset

Write a clinical summary of this patient's cardiovascular profile.

---

---

Extract the data of the fifth patient in the dataset

Please describe this data in natural language

Summarize your description

---

---

Write a referral letter to a cardiologist describing the first patient in the dataset.

---

---

extract the data of patient number 50

Write a referral letter to a dietitian describing this patient and focusing on the importance of dietary evaluation and management.

---