# Clinical Data Extraction and Summarization

## Introduction:

The goal of this technical evaluation is to create a comprehensive solution for extracting information from healthcare datasets, especially text documents. Autism Spectrum Disorder (ASD) is a developmental disorder that affects communication and behavior. It is crucial to detect ASD early so that children can receive the support they need to thrive. During the intervention of autism, clinicians access several electronic healthcare records of treatment progression. Developing a tool assists clinicians to quickly get relevant information and produce insights, saving time and increasing their capacity.

## Problem statement:

Access to organized medical information in clinical datasets is limited, preventing effective data exploitation. The challenge is to create a unified solution that uses Named Entity Recognition (NER) for entity extraction, document search based on a query, and clinical document summarization using Large Language Models (LLMs) to allow healthcare professionals to retrieve and comprehend relevant medical data quickly.

The following steps should be followed for developing your solution:

1. Develop an **NER system to extract important medical entities**.
2. Enable users **to search a clinical document using a text query** to better retrieve relevant information.
3. **Pretrain or fine-tune a small lightweight deep model (with a model size not more than 3GB, or a very light model that can efficiently fit on most hardware with CPUs/GPUs) on a dataset** relevant to healthcare to improve the contextual understanding of clinical notes/electronic health records and for the better extraction of medical entities. The **embeddings generated by the model should be leveraged** to extract more meaningful representations of words or sentences, capturing the context. The proposed technique can use transformer models to adapt to the specifics of healthcare data.
4. Automatically **generate a concise and informative summary of the clinical document utilizing the learned representations** from the previous step 3 and using LLMs.
5. **Integrate** the NER, Document Search, and Summarization components into a unified solution.

Upon running your code on a clinical document, the code should print the entities recognized by the NER technique, ask the user for a query to search within the document and retrieve relevant results to display from the document, and finally print a short summary of the clinical document.

## Dataset:

You can collect and construct the solution from any existing open-source healthcare datasets. **It is encouraged that you use a dataset related to mental health and treatment with audio/speech if available**. You can use speech signals along with text documents in step 3 to leverage more meaningful representations.

## Deliverables:

1. Source code files: Data preparation (if any), training/inference pipeline scripts (if any), and requirements files.
2. README.md: A short description of the approach to the solution and a screenshot of the results. The descriptions should be easy to follow and help the evaluator easily reproduce the results.
3. All the above are archived in one single .zip or .tar file. Either send a mail or share the link to download.

## Time requirements and submission timeline:

1. You can expect this assignment to take around 12 hours to complete.

2. Please submit your work within 7 days of receiving the assignment.