**A blue and black text

Description automatically generated**

**TECHNICAL EXERCISE**

**Designed by: Tim Reason**

**Candidate: XXX**

**Version 1.0**

**September 2025**

**Why this technical exercise?**

The technical exercise mimics the type of AI-related problem you might be asked to work on at Estima. The exercise gives you the opportunity to demonstrate your independent problem-solving skills, one of the core skills for the role. We are looking for individuals who can provide a well-thought-out first approach, whilst recognizing that it may not be possible to solve every aspect of the problem. Please be ready to discuss your solution during a follow-up interview.

**Task Description**

**Context**

Network meta-analyses (NMAs) are regularly conducted as part of health technology assessments (HTAs) and provide input data for health economic models, which produce results that compare the costs and clinical effects of health interventions. To conduct an NMA, firstly, a systematic literature review (SLR) is conducted, which identifies all available data in the disease area being considered, and then relevant data is extracted from relevant publications. This includes baseline characteristic for patients, and clinical outcome data.

Data extraction is very time consuming, and errors are frequently observed. Automation of data extraction may help to reduce these errors.

A client has funded a project to investigate whether large language models (LLMs) can facilitate automation of data extraction. The client’s vision is to see if it is possible for an LLM to process relevant publications[[1]](#footnote-1) and extract relevant data into a data extraction form (a formatted Excel or CSV file).

You are working as part of a team with a technical lead to deliver this project. You are tasked with scoping out the problem and designing a first approach. You will later discuss this with the technical lead and develop the approach further.

**Task**

You are to create a data extraction form in Excel, which has columns for required data. These need to be populated by data extracted from two relevant publications in the area of ulcerative colitis, which have been provided as PDFs.

You have also been informed by the client that the following baseline characteristics need to be extracted:

* Number of patients recruited by trial per treatment arm
* Patient gender: Male (number, proportion), Female (number, proportion)
* Age (mean or median, measure of data spread e.g., standard deviation)
* Body mass index (mean or median, measure of data spread e.g., range)
* Patient race (number/proportion): White; Native American or Alaska Native; Asian; Black or African American; Native Hawaiian or Other Pacific Islander
* Disease duration (mean or median, measure of spread, units)
* Location and extent of disease (number/proportion):Left side; Extensive or pancolitis; Limited to rectum
* Number of acute UC episodes in the past year (mean or median, measure of spread)
* Adapted Mayo score (mean or median, measure of spread)
* Adapted Mayo score category (number, proportion): ≤ 7, > 7
* Previous medication use (number, proportion): Immunosuppressants; Aminosalicylates; Corticosteroids.

And that the following clinical outcomes need to be extracted:

* Proportion of patients[[2]](#footnote-2) with a clinical response at all reported timepoints between 8 and 26 weeks
* Proportion of patients with a clinical and endoscopic response at all reported timepoints between 8 and 26 weeks
* Proportion of patients[[3]](#footnote-3) with any adverse event (AE)
* Proportion of patients with any treatment-emergent adverse event (TEAE).

We would like you to think about how this task might be achieved, and to make a start on delivering it. We would like you to spend between 3 and 6 hours working on this (more should you wish).

For this, we would like you to provide the following:

* A bullet list of the steps likely to be needed to achieve the automation process (in Word)
* Pseudocode or code to conduct the data extraction (Python)
* The populated (as far as possible) data extraction form (Excel)

The data extraction form only needs to include data that you have managed to automatically extract.

You are not expected to provide a complete working solution for this but to develop an approach for the project, and to start implementing some of your ideas.

If you would like to ask further clarifying questions about this exercise, please contact Emma Benbow at [emma.benbow@estima-sci.com](mailto:emma.benbow@estima-sci.com)

OpenAI API key:

DsPzbRoxQ4BAgRW44iXyCAJ1jPWB7ZQLKAAmGU01KoA5CYlNgpwRjWNHS4d0NjTpsfTK73pFliT3BlbkFJrFIjqVlSl4NT8VsiJyOJsFyjvbrWapLhzkP8vwtErtFhnhN9hf0c5ni3ato9MKjg1-qpDPswsA

1. Publications with the relevant copyright to allow scraping by AI [↑](#footnote-ref-1)
2. number of patients in each treatment arm; number of patients or percentage of patients achieving response in each arm, timepoint at which outcome is reported [↑](#footnote-ref-2)
3. number of patients in each treatment arm; number of patients or percentage of patients experiencing an AE in each arm [↑](#footnote-ref-3)