Title: Leveraging Artificial Intelligence(AI) in Individual Patient Data Meta-Analysis(IPD-MAs) in resource-limited settings

Authors: N Brink, C Parker, P Marsh, T Makanga

Introduction: Individual Patient Data Meta-Analyses(IPD-MAs) rely on multiple clinical trials to pool data for analysis. This allows researchers to investigate research questions and reach conclusions not possible with smaller sample sizes. The HE2AT centre has ambitious goals to uncover important heat-health interactions by creating a rich database of clinical data, spatially and temporally mapped to climate and socio-economic data, using studies conducted in Sub-Saharan Africa. Study heterogeneity, however, complicates this process. Advances in artificial intelligence(AI) promise an opportunity to leverage the power of automation in conducting IPD-MAs.

Objectives, and justifications: To develop an approach to automate time-consuming tasks associated with data harmonisation to reduce costs and improve efficiency, thus increasing accessibility of IPD-MAs in resource-limited settings. Large-Language-Models(LLMs), like those pioneered by OpenAI and others, rely on an underlying framework of tokenisation and embeddings to allow for a machine-readable understanding of semantic meanings of text. The complex but patterned nature of data harmonisation across studies allows for the potential for this technology to be applied to complete time, financial and human-intensive processes.

Methodology: A master codebook was manually developed using standardised ontology terms to facilitate the harmonisation of clinical datasets to the HE2AT Centre database. This information is utilised to automatically match variables from each study to their counterpart in the master codebook using embeddings. Following this, instructions are generated using LLMs to harmonise from the study dataset. This instruction is then carried out using LLM-generated Python script. A human expert approves each step. The resulting harmonised data is stored in a master database, augmented by additional climate and socioeconomic data.

Anticipated results: Automated workflow, with human experts in the loop to validate outputs to render large-scale data harmonisation within reach of traditionally-resource limited contexts by optimising efficiency to produce rich datasets for analysis. Feedback loops allow for continuous improvements while ensuring that experts in their respective fields validate all outputs.

Next Steps: Large-scale fine-tuning of open-source LLMs and embedding models to provide a generalisable, semi-automated assistant to aid clinical data harmonisation.