Project Details	
Project Code	MRCPHS24Br Liu
Title	Automating knowledge synthesis in biomedical literature using AI and language models
Research Theme	Population Health Sciences
Summary	Knowledge synthesis approaches can be used to extract and summarise large amounts of information from research articles to generate broader perspectives and new insights e.g. to comprehensively summarize the effect of obesity and other factors on the risk of breast cancer using all relevant research articles. This PhD will explore using artificial intelligence approaches, such as language models to automate knowledge synthesis on a large scale.
Description	Background: Knowledge synthesis can be a slow and cumbersome process but is an essential tool for medical and public health policy-makers. Formal systematic reviews require rigid protocols and extensive human effort from trained professionals, whereas a massive volume of research evidence emerges every year. Recent advances in computational approaches in artificial intelligence (AI) and natural language processing (NLP), such as named entity recognition and text summarization powered by large language models (LLMs, such as BERT or GPT-4), have substantially improved the efficiency and accuracy of information extraction at a massive scale. These AI and NLP methods could potentially offer the opportunity to complement manual systematic reviews with automated knowledge synthesis reports that are updated in real-time as new literature emerges, and therefore improve the efficiency and responsiveness of decision-making of systematic reviewers, public health professionals and clinicians. This project aims to substantially improve the rapid review of literature by developing methods to automate knowledge synthesis from published research articles. The student will explore use of language models and NLP methods to automate: 1) the identification and extraction of key research information, and 2) the assessment of research quality and risk of bias from a published or pre-print article. Objective 1: To investigate the performance of various automated approaches for text summarization. The student will evaluate and compare several NLP methods including rule-based approaches, small-scale language models as well as LLMs for the summarization of literature text as available from research article databases/online sources (e.g. PubMed, and preprint servers such as medRxiv) regarding the overall research objectives, key findings etc. The student will also evaluate the respective functionalities and performance, bias, as well as their potential environmental impact from these approaches. This objective aims to train the student regarding the current widely-used methods and underlying biomedical literature data, as well as to further lead to the adequate combination and balance between multiple methods for subsequent objectives. The student will be able to define their own evaluation criteria after consultation with stakeholders and experts. Objective 2: To develop efficient and robust methods to extract and harmonize structural information from the literature text. The student will develop methods to identify and extract key information from the scientific text regarding its research and findings, such as the involved objects / subjects,

	research methods and study designs, as well as quantitative information such as effect sizes, etc. In addition, the student will integrate methods combining rule-based approaches and LLMs to coherently harmonize the extracted individual information across heterogenous study types (such as randomized trials, non-randomized interventions, observational studies) to enable comprehensive triangulation of evidence. The student will have the opportunity to decide on the details of information extraction and integration of methods. Objective 3: To develop a novel framework to automate the assessment of risk of bias from research articles. Risk of bias refers to the risk that the research findings reported by a study are inaccurate due to limitations of the study, such as selection bias, measurement error or missing information. The student will build on methods developed in Obj. 2 to robustly assess the risk of bias of a study, and automate such assessment in published and pre-print research articles as well as contrast and validate the results from their developed methods with established frameworks (e.g. RobotReviewer, ROBINS-E, etc). The student will steer the choice of methods, which could involve developing new components or repurposing existing ones.
Supervisory Team	
Lead Supervisor	
Name	Dr Yi Liu
Affiliation	Bristol
College/Faculty	Faculty of Health Sciences
Department/School	MRC Integrative Epidemiology Unit, Population Health Sciences, Bristol Medical School
Email Address	yi6240.liu@bristol.ac.uk
Co-Supervisor 1	
Name	Dr Louise Millard
Affiliation	Bristol
College/Faculty	Faculty of Health Sciences
Department/School	Population Health Sciences, Bristol Medical School
Co-Supervisor 2	
Name	Dr Luis Espinosa-Anke
Affiliation	Cardiff
College/Faculty	College of Physical Sciences and Engineering
Department/School	School of Computer Science and Informatics
Co-Supervisor 3	
Name	Professor Tom Gaunt
Affiliation	Bristol
College/Faculty	Faculty of Health Sciences
Department/School	MRC Integrative Epidemiology Unit, Population Health Sciences, Bristol Medical School