Assisting Study Selection in Systematic Reviews Using a Large Language Model

Zhaozhen Xu, Senior Research Associate

HDRN workshop

July 2025

MRC Integrative Epidemiology Unit

University of BRISTOL

About Me

PhD in Computer Science:
Natural language processing (NLP)
Large language model (LLM)
Sentence Embedding
Question (a special type of text)

MSc in Advanced Computing:
Data mining, Machine learning

Postdoc in MRC IEU:
Applying machine learning and NLP in the medical text
Text mining
Systematic review automation

MRC Integrative Epidemiology Unit

University of BRISTOL

Background

The World Cancer Research Fund has an ongoing research programme called Global Cancer Update Programme (CUP Global) since 2007

- Provide up-to-date systematic reviews to analyse the evidence linking diet, nutrition and physical activity to the risk of, and survival from, cancer

A team of researchers at Imperial College produce at least two high-quality systematic reviews a year, and they are constantly collecting

- **new evidence for existing topics**
- **evidence for new future topics**

The process of conducting a systematic review is labour and time intensive

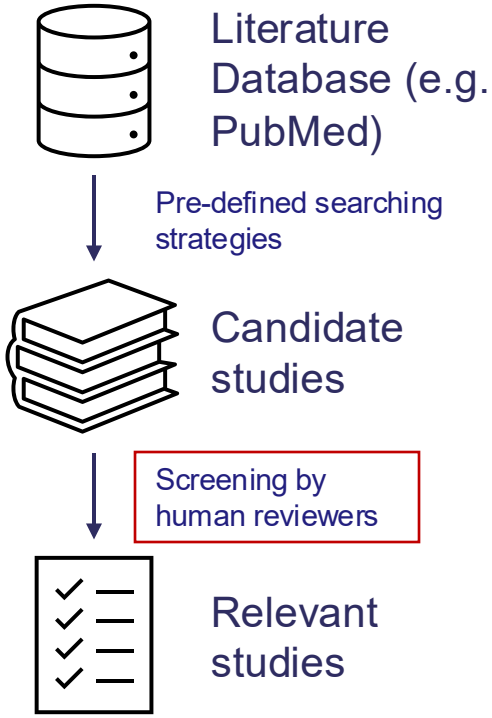
MRC Integrative Epidemiology Unit

University of BRISTOL

Study Screening (Semi) Automation

Aim: Increase the efficiency of conducting systematic reviews by assisting the human reviewers to identify relevant studies from the search results retrieved from the literature database

- Train a study screening model powered by a large language model to (semi) automate the process of screening
- The trained model can provide reliable predictions on existing topics and can be extended for future topics

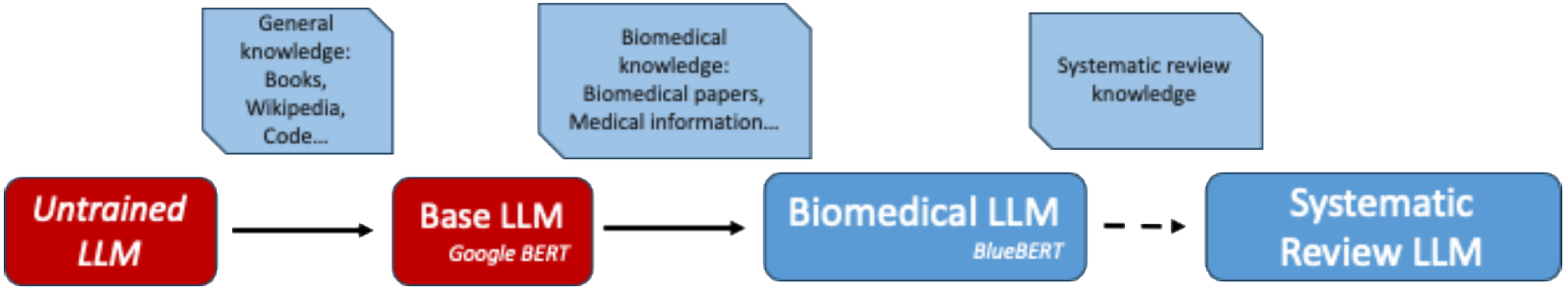


MRC Integrative Epidemiology Unit

University of BRISTOL

Large Language Model (LLM)

- **LLMs** are a new generation of machine learning methods, and has been increasingly widely used (e.g. OpenAI ChatGPT)
- Morden methods: Attention mechanism, Transformer architecture
- Large parameter size (340 millions for BERT, 1.8 trillions for GPT-4)
- More powerful in natural language understanding compared to traditional machine learning models, but more difficult to train and apply
- Pre-trained with general and domain specific (e.g. biomedical) knowledge



MRC Integrative Epidemiology Unit

University of BRISTOL

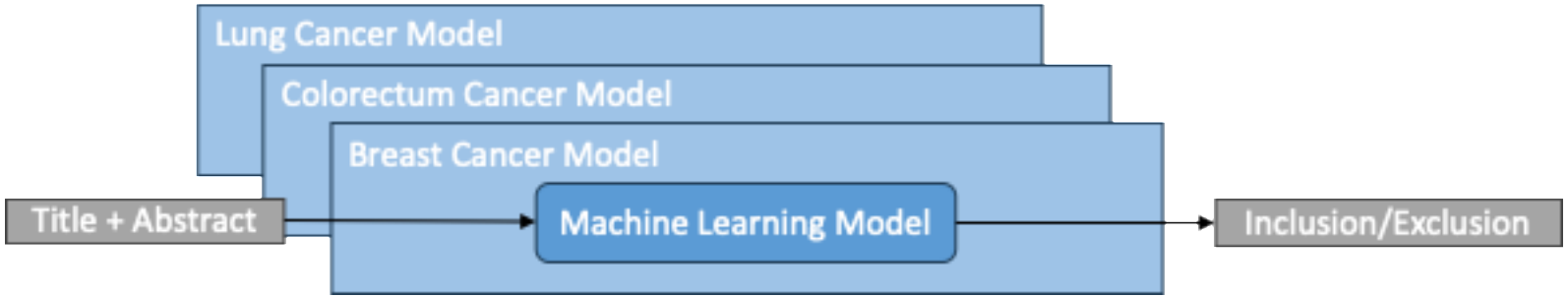
Study Screening Data

Topic	Study Title	Study Abstract	Included/Excluded
Breast Cancer	Cardiometabolic factors and breast cancer risk in U.S. black women.	Previous studies have suggested that metabolic syndrome …	Included
Breast Cancer	Factors associated with the use of preventive services by women in Greece.	BACKGROUND: The purpose of the current work was …	ExcludedTitleAbstract
Bladder Cancer	Urethral caruncle in a 9-year-old girl: a case report and review of the literature	INTRODUCTION: Urethral caruncles are the most frequent benign tumors of …	ExcludedFullText

- 17 topics in cancer incidence
- Around 96,000 candidate studies in total
- Less than 8.6% of the studies are included in the CUP-Global systematic review database
- Inclusion and exclusion criteria are given for each topic

Topic-Specific Study Screening Model

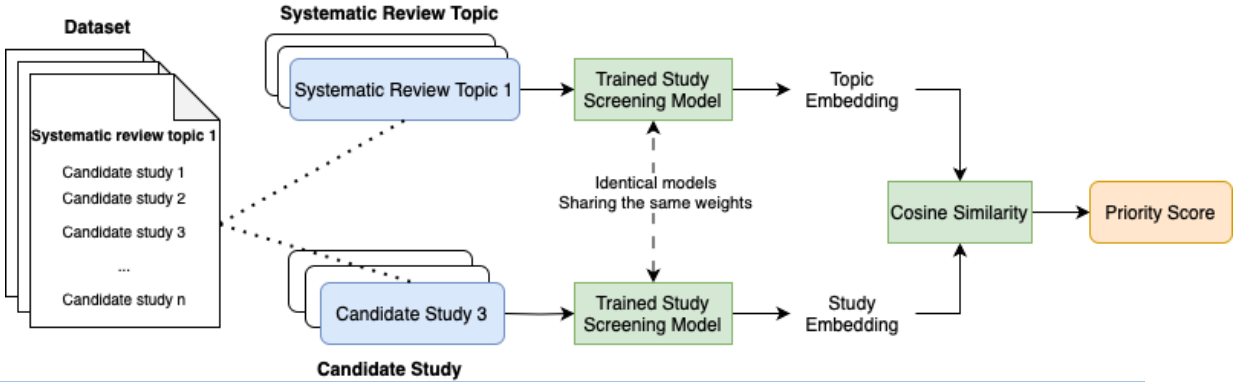
- Screening studies by titles and abstracts
- Customised to maximise the prediction performance for each topic
- Limitations:
  - Each topic needs to train a separate model -> computational cost, data sufficiency
  - Cannot be extended to new topics

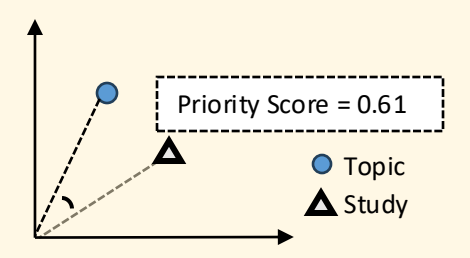


M-PreSS: A Model Pre-training Approach for Study Screening (powered by BlueBERT)

- Opposite to topic-specific models, a general model can provide reliable predictions on multiple topics with one trained model
- Can be extended to new topics, with/without data from the new topics
- However, it might trade off some performance for generalisation

Query: Breast cancer incidence. Criteria: ….





Title: Cardiometabolic factors and breast cancer risk in U.S. black women. Abstract: …

Figure source: "M-PreSS: A **M**odel **Pre**-training Approach for **S**tudy **S**creening in Systematic Reviews" (Preprint)

Threshold for Decision Making

- Indicate where a reviewer can stop reviewing
- Suggest an inclusion/exclusion decision to assist human reviewer
- Compare the performance between the trained model and human reviewers

A specific threshold was calculated from the training set for each topic that includes all the relevant study with least false positives (i.e. best precision at 100% recall)

Threshold = (Score_4 + Score_5) / 2
= (0.550 + 0.548) / 2

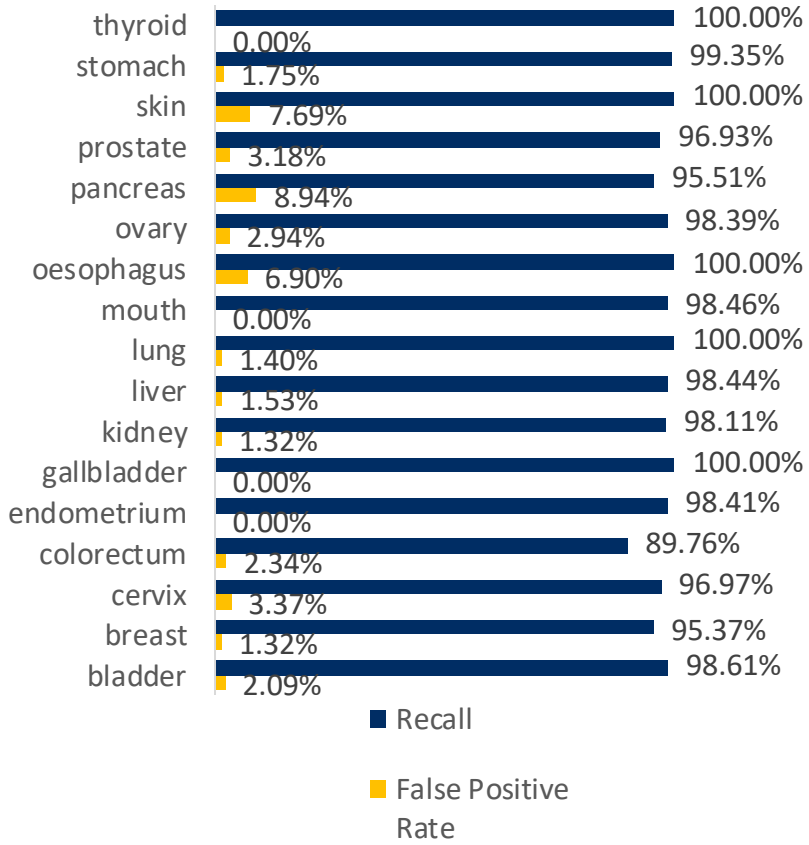
Rank	Candidate Studies	General Model Output: Priority Score	*Prediction with Threshold*	Human decision
1	Study 1	0.66251105	included	included
2	Study 2	0.5602507	included	excluded
3	Study 3	0.55684996	included	included
4	Study 4	0.5505986	included	included
5	Study 5	0.54793453	excluded	excluded
6	Study 6	0.5449597	excluded	excluded
7	Study 7	0.5416114	excluded	excluded

Cancer Incidence Study Screening Results

The model has more than **90% of Recall** among most of the topics (apart from colorectum). In the meanwhile, the **false positive rate is under 10%** for all topics.

- Identify most of the primary studies without including too many false positive studies to further review and analysis
- Excluded the studies on title abstract level, whereas humans need to read the full text to exclude

➢ The model can make reliable prediction when there is new evidence collected for existing topics



MRC Integrative Epidemiology Unit

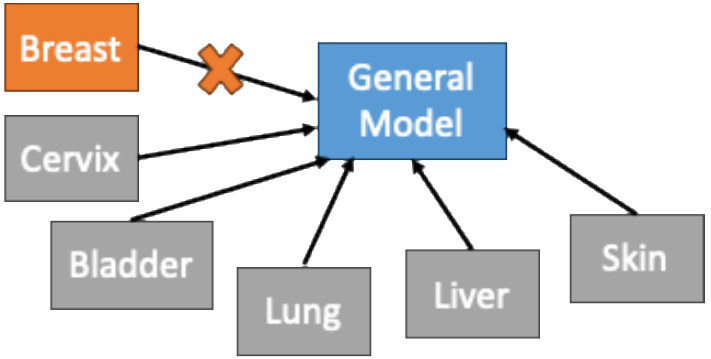
University of BRISTOL

Leave-One-Out Cross-Validation

One of the advantages of training a general model is that it can be applied to new topics without extra training

Evaluate the study screening model's ability to transfer the study screening knowledge/patterns learned to a new topic:

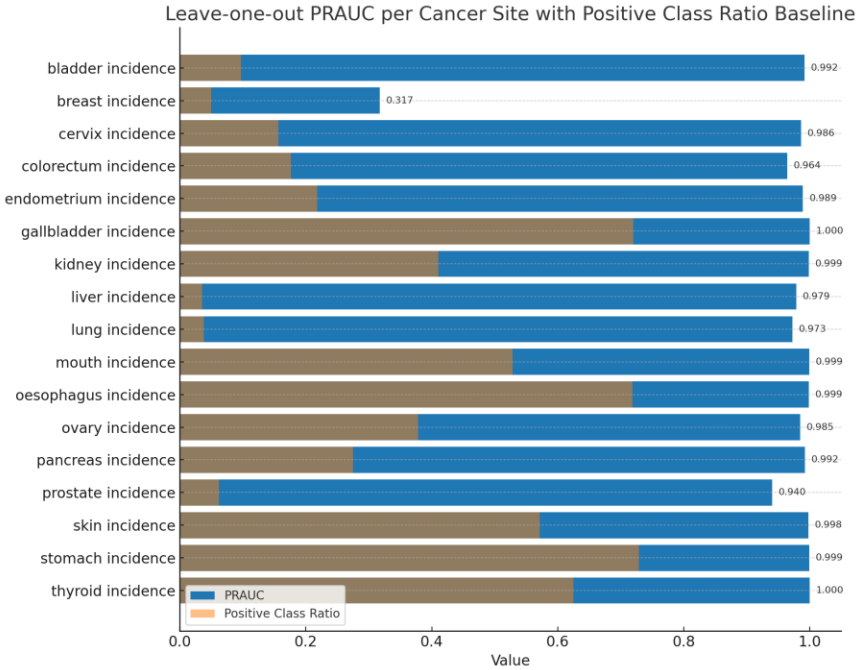
1. Remove the target topic (e.g. breast cancer) from the training data
2. Train the model
3. Evaluate the trained model (i.e. a general model without previous knowledge on the breast cancer study screening) on the target topic



Leave-One-Topic-Out Results

- Repeat the leave-one-topic-out validation for all the topics in CUP Global dataset
- The model has a good generalisation ability for the unseen topics

➢ The trained study screening model with existing topics can make reliable prediction when there are new topics added to the project



MRC Integrative Epidemiology Unit

University of BRISTOL

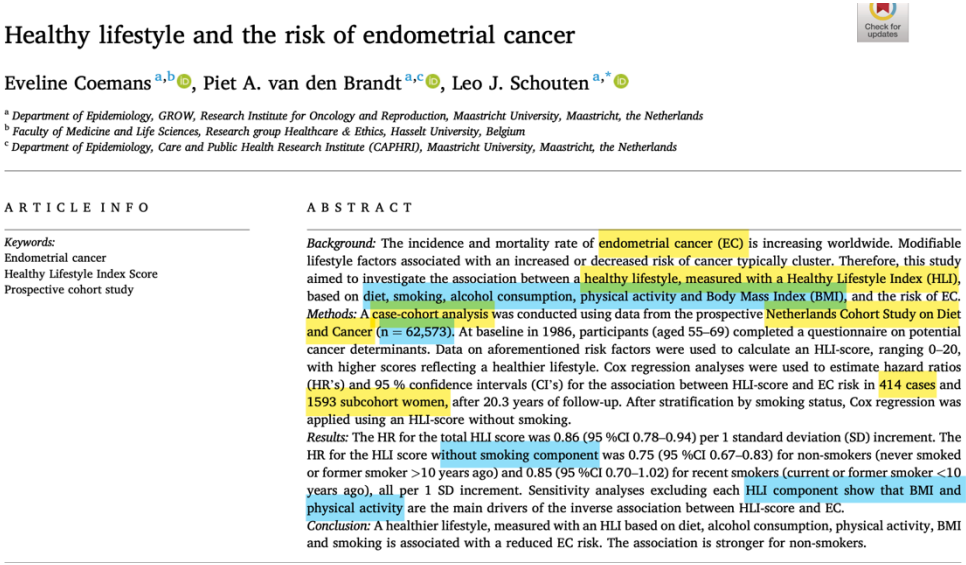
Ongoing Work: Key Information Extraction

Some key information can also assist in study screening decision-making

- Exposure
- Outcome, outcome type
- Study design
- …

With the information extracted

- Help the reviewer quickly locate the key information of a study
- Add the information extracted into the inputs general model to provide a more accurate prediction



Healthy lifestyle and the risk of endometrial cancer

Eveline Coemans[a,b], Piet A. van den Brandt[a,c], Leo J. Schouten[a,*]

[a] Department of Epidemiology, GROW, Research Institute for Oncology and Reproduction, Maastricht University, Maastricht, the Netherlands
[b] Faculty of Medicine and Life Sciences, Research group Healthcare & Ethics, Hasselt University, Belgium
[c] Department of Epidemiology, Care and Public Health Research Institute (CAPHRI), Maastricht University, Maastricht, the Netherlands

ARTICLE INFO

Keywords:
Endometrial cancer
Healthy Lifestyle Index Score
Prospective cohort study

ABSTRACT

Background: The incidence and mortality rate of endometrial cancer (EC) is increasing worldwide. Modifiable lifestyle factors associated with an increased or decreased risk of cancer typically cluster. Therefore, this study aimed to investigate the association between a healthy lifestyle, measured with a Healthy Lifestyle Index (HLI), based on diet, smoking, alcohol consumption, physical activity and Body Mass Index (BMI), and the risk of EC.
Methods: A case-cohort analysis was conducted using data from the prospective Netherlands Cohort Study on Diet and Cancer (n = 62,573). At baseline in 1986, participants (aged 55–69) completed a questionnaire on potential cancer determinants. Data on aforementioned risk factors were used to calculate an HLI-score, ranging 0–20, with higher scores reflecting a healthier lifestyle. Cox regression analyses were used to estimate hazard ratios (HR's) and 95 % confidence intervals (CI's) for the association between HLI-score and EC risk in 414 cases and 1593 subcohort women, after 20.3 years of follow-up. After stratification by smoking status, Cox regression was applied using an HLI-score without smoking.
Results: The HR for the total HLI score was 0.86 (95 %CI 0.78–0.94) per 1 standard deviation (SD) increment. The HR for the HLI score without smoking component was 0.75 (95 %CI 0.67–0.83) for non-smokers (never smoked or former smoker >10 years ago) and 0.85 (95 %CI 0.70–1.02) for recent smokers (current or former smoker <10 years ago), all per 1 SD increment. Sensitivity analyses excluding each HLI component show that BMI and physical activity are the main drivers of the inverse association between HLI-score and EC.
Conclusion: A healthier lifestyle, measured with an HLI based on diet, alcohol consumption, physical activity, BMI and smoking is associated with a reduced EC risk. The association is stronger for non-smokers.

MRC Integrative Epidemiology Unit

University of BRISTOL

Summary

- We trained a general study screening model to identify relevant studies for CUP Global to conduct systematic reviews related to cancer incidence
- The model achieved 90% recall and 10% false positive rate on most of the cancer sites
- The trained general model was able to provide reliable predictions on new topics without further training on new data

Acknowledgement

University of Bristol team

- Core team: Zhaozhen Xu, Tom Gaunt, Yi Liu (PI)
- Advisor: Louise Millard, Richard Martin, Julian Higgins, Philippa Davies, Maria Sobczyk-Barad

Imperial College London team

- Doris Chan, Kostas Tsilidis, Eduardo Seleiro, Lam Teng, Georgios Markozannes

World Cancer Research Fund

World Cancer Research Fund

M-PreSS: A **M**odel **Pre**-training Approach for **S**tudy **S**creening in Systematic Reviews (Preprint)

Zhaozhen Xu: zhaozhen.xu@bristol.ac.uk

medRxiv
THE PREPRINT SERVER FOR HEALTH SCIENCES

UKRI MRC Integrative Epidemiology Unit

University of BRISTOL