Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240427

Towards Autonomous Living Meta-Analyses: A Framework for Automation of Systematic Reviews and Meta-Analyses

Anna GÓRSKA [a,1] and Evelina TACCONELLI [a]
[a] *Infectious Diseases Section, Department of Diagnostics and Public Health, University of Verona, Verona, Italy*
ORCiD ID: Anna Górska https://orcid.org/0000-0003-3305-8711

Abstract. Systematic review and meta-analysis constitute a staple of evidence-based medicine, an obligatory step in developing the guideline and recommendation document. It is a formalized process aiming at extracting and summarizing knowledge from the published work, grading, and considering the quality of the included studies. It is very laborious and time-consuming. Therefore, the meta-analyses are rarely updated and seldom living, decreasing their utility with time. Here, we present a framework for integrating the large language models and natural language processing techniques applied to the previously published systematic review and meta-analysis of the diagnostic test accuracy of the point of care tests. We show that the framework can be used to automate the screening step of the existing meta-analyses with minimal costs to quality and, to a large extent, the extraction step while maintaining the strict nature of the systematic review process.

Keywords. Meta-analysis, Large Language Models, Fine-tuning, Systematic review, Guidelines, Workflow automation

1. Introduction

Systematic review and meta-analysis (SRM) constitute a staple of evidence-driven medicine, and it is the most reliable methodology for developing guidelines and recommendations. Most SRM processes are manual and extremely time and labor-intensive. Therefore, every meta-analysis quickly becomes outdated. Due to the time requirements, less than 1% (estimated based on a PubMed search) of published SRMs are regularly maintained. Natural language processing (NLP) tools have been utilized to automate and streamline SRM [1-3]. However, the recent advancements in the large-language models (LLMs) provide novel, powerful capabilities and constitute a fast-growing field [2,4-6]. Nonetheless, up to now, none have been applied to infectious diseases, provided the desired flexibility, and proposed a comprehensive framework to maintain and utilize multiple performed SRMs.

The SRM is a strictly defined multi-step process, comprising: (1) defining the question and search strategy, (2) performing the search, (3) abstract and title screening, (4) full-text screening and extraction, (5) meta-analysis (deriving summary statistic). Here, we present a novel, light-weight, modular, flexible, reproducible, and monitorable

[1] Corresponding Author: Anna Górska; E-mail: anna.gorska@univr.it.

automation framework for the two most time-consuming steps: (3) screening and (4) extraction, applied to our systematic review and meta-analysis of the diagnostic test accuracy of point-of-care tests in acute respiratory tract infections, published in January 2022 [7,8]. Briefly, the meta-analysis aimed at summarizing performance statistics of various clinical tests applicable in community settings (primary care or emergency room) to determine if the respiratory tract infection (influenza-like-illness, flu, pneumonia, etc.) was caused by the viral or bacterial pathogen, to prevent the inappropriate antibiotic prescribing.

2. Methods

2.1. Meta-analysis

Overall, 10,086 publications were screened for the systematic review, with 625 papers accepted in the full-text screening. Since 99% of the publications were present in PubMed, the current framework is limited to PubMed and utilizes BioPython [9]. The publication details, extraction values, screening, and all other information were deposited in the relational database.

2.2. Screening

The screening was formulated as a supervised classification learning problem. The dataset consisted of the merged publications' titles and abstracts and the screening label constituting if the publication participated in the ValueDx meta-analysis, i.e., was positively screened based on both the abstract and full text. The dataset was split into train/test sub-sets (70%-30%). Four LLM models: biobert-base-cased-v1.1 [10], BiomedNLP-PubMedELECTRA-base-uncased-abstract [11], BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext [12], BioM-BERT-PubMed-PMC-Large [13] were independently fine-tuned for this task. The LLMs were downloaded from huggingface.co and handled using the Python transformers library [14]. Different inputs, i.e., portions of the abstract and title, were tested as well: title and abstract, only title, and only the last sentence, as the sanity check. Loss, precision, and recall performance statistics were computed. For the best-performing model, the rarefaction curves were computed.

2.3 Extraction

Each extraction column calls for a specifically tailored pipeline. Broadly, the extraction was based on the three main methods. The first method (NER) utilized a simple Named Entity Recognition with spaCy [15], then mapping to the manually curated and structured reference dataset. Each item in the reference consisted of the entity, synonym list, and extraction list. Therefore, a single entity could participate in several extraction variables. This pipeline was applied to each sentence in the text and later summarized. The second method (LLM) relied on the supervised LLM finetuning, analogous to the screening, developed either in-house using the extraction variables shared across multiple reviews or developed by other parties. The third method (GPT) relied on querying the OpenAI's gpt-3.5-turbo-1106 (with cl100k_base encoding) via the Python API with a structured query: "You are a helpful assistant. This was a title and abstract of the scientific study

[TITLE + ABSTRACT].", followed by the extraction question, e.g., "Does this study use white blood count. Answer with yes or no.". Finally, the country was extracted with the geography3 package (16). In the current version, the extraction can be applied to the abstracts.

Although the original ValueDX extraction table included over a hundred columns, only a few entered the final analysis. Here, we present results for the Population, across all methods. The LLM-based method was tested by splitting the dataset into training and testing. The GTP and NER were tested against the manual full-text extraction. The crucial part of the extraction, which was not automated at this stage, was the confusion matrix, describing how well the point of care test performed and the QUADAS-2-based quality criteria [17].

3. Results

3.1. Screening

The fine-tuning for the screening classification task worked well. As expected, the best models were fine-tuned using the title and the beginning of the abstract (Table 1). The rarefaction curves plateaued before 20% of the dataset, suggesting the dataset was sufficiently large.

Table 1. The four best combinations for screening the Value-DX.

Model	Loss	Precision	Recall
BiomedNLP-PubMedELECTRA-base-uncased-abstract	0.110	0.960	0.960
biobert-base-cased-v1.1	0.130	0.960	0.960
BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext	0.170	0.950	0.950
BioM-BERT-PubMed-PMC-Large	0.230	0.940	0.940

3.2. Extraction

Both NER and GTP methods were able to produce the answers for the great majority of the extracted papers. **Table 2** presents the results of the automated extraction. The first method (NER) performed well in detecting children but worse when detecting adults and mixed than both the GTP and LLM methods. It is understandable, as pediatric populations were probably clearly defined in the abstract. The LLM-based method performed much worse than the analogous screening exercise. The models were characterized by a larger loss and poorer precision or recall. This performance was most likely due to the smaller amount of data available for this training, as they can only be done for the extracted publications and as it performed much better in the binary classification task (Viral or Bacterial).

Table 2 The best runs for the three variables extracted with the second method, i.e., in-house finetuning.

Variable	Method/Model (Values)	Loss	Precision	Recall
Population	NER (Adults/Children/Mix)	-	0.60/0.98/0.1	0.50/0.73/1.0
Population	LLM: BioM-BERT-PubMed-PMC-Large	0.590	0.770	0.770
Population	GTP (Adults/Children/Mix)	-	0.97/0.93/0.06	0.36/0.76/0.75

Viral or Bacterial disease	LLM: BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext	0.390	0.970	0.970

4. Discussion and Conclusions

Here, we introduced a framework for automating systematic reviews and meta-analyses and showed its application to real-life SRM. It is a supervised, transparent, and explainable method that allows for minimal loss of quality while maintaining the strict nature of the SRM process. Even at this stage of the framework development, the screening step can be automated without losing quality. The extraction step can be semi-automated, enabling a significant speedup to the extraction process. All these factors allow the framework to be included in developing medical guidelines.

The current framework has limitations, as it can interact only with PubMed. The framework does not attempt to assign the quality of the studies, as we believe it must remain the clinical scientist's manual work. It might not be applicable to other systematic reviews, especially those with complex inclusion criteria for which the information might not be included directly in the abstract. Therefore, the framework should be carefully adapted to each new review.

The underlying database enables seamless interaction and maintenance of the reviews, easy no-loss updates, and comparisons between meta-analysis and the searches performed on other dates. It also enables the reuse of extractions performed for the overlapping publications between related systematic reviews. It is also crucial when applying the LLM-based extraction pipeline, as the common extraction columns can be leveraged to get more data for the LLM-finetuning. The database retains the extraction author; therefore, the automatic extractions will not pollute the training sets.

The modular structure of the extraction allows for any number of extensions, e.g., adding modules specialized in extracting particular variables. Additionally, the automated extraction from the abstracts can be used to construct a pre-screening for the new systematic review, to either guide the reviewer in selecting the most informative publications so that the remaining publications can be automated or to construct the screening logic that might enable to entirely skip the title and abstract screening step, and therefore provide a continuously updated meta-analysis, delivering transferability of new scientific results directly to the patient care.

Funding

Innovative Medicines Initiative-2 Joint Undertaking, grant agreement No 820755 (Value-Dx) This Joint Undertaking receives support from the EuropeUnion's Horizon 2020 research and innovation programme and EFPIA and bio- Marieux SA, Janssen Pharmaceutica NV, Accelerate Diagnostics SL, Abbott, Bio-Rad Laboratories, BD Switzerland Sarl, and The Wellcome Trust Limited The commercial companies had no part in the design, analysis, writing or decision to publish the results.

References

[1] D'Ambrosio A, Grundmann H, Donker T. An open-source integrated framework for the automation of citation collection and screening in systematic reviews. 2022; Available from: http://arxiv.org/abs/2202.10033

[2] Thomas J, Noel-Storr A, Marshall I, Wallace B, McDonald S, Mavergames C, et al. Living systematic reviews: 2. Combining human and machine effort. J Clin Epidemiol. 2017;91:31–7.

[3] van de Schoot R, de Bruin J, Schram R, Zahedi P, de Boer J, Weijdema F, et al. An open source machine learning framework for efficient and transparent systematic reviews. Nat Mach Intell. 2021 Feb 1;3(2):125–33.

[4] Alshami A, Elsayed M, Ali E, Eltoukhy AEE, Zayed T. Harnessing the Power of ChatGPT for Automating Systematic Review Process: Methodology, Case Study, Limitations, and Future Directions. Systems. 2023 Jul 1;11(7).

[5] Khraisha Q, Put S, Kappenberg J, Warraitch A, Hadfield K. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. Res Synth Methods. 2024;

[6] Van Dijk SHB, Brusse-Keizer MGJ, Bucsán CC, Van Der Palen J, Doggen CJM, Lenferink A. Artificial intelligence in systematic reviews: promising when appropriately used. Vol. 13, BMJ Open. BMJ Publishing Group; 2023.

[7] Hellou MM, Górska A, Mazzaferri F, Cremonini E, Gentilotti E, De Nardo P, et al. Nucleic-acid-amplification tests from respiratory samples for the diagnosis of coronavirus infections: systematic review and meta-analysis. Clin Microbiol Infect [Internet]. 2020;27(3):341–51. Available from: https://doi.org/10.1016/j.scitotenv.2019.135577

[8] Gentilotti E, De Nardo P, Cremonini E, Górska A, Mazzaferri F, Canziani LM, et al. Diagnostic accuracy of point-of-care tests in acute community-acquired lower respiratory tract infections. A systematic review and meta-analysis. Clinical Microbiology and Infection [Internet]. 2022;28(1):13–22. Available from: https://doi.org/10.1016/j.jaci.2021.02.040

[9] Buchmann JP, Holmes EC. Entrezpy: A Python library to dynamically interact with the NCBI Entrez databases. Bioinformatics. 2019;35(21):4511–4.

[10] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234–40.

[11] Tinn R, Cheng H, Gu Y, Usuyama N, Liu X, Naumann T, et al. Fine-Tuning Large Neural Language Models for Biomedical Natural Language Processing. ArXiv. 2021 Aug 1;33(16):1–7.

[12] Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. ACM Trans Comput Healthc. 2022 Jan 1;3(1).

[13] Sultan Alrowili E, Vijay-Shanker K. BioM-Transformers: Building Large Biomedical Language Models with BERT, ALBERT and ELECTRA [Internet]. 2021. Available from: https://github.com/salrowili/

[14] Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. 2019 Oct 8; Available from: http://arxiv.org/abs/1910.03771

[15] Honnibal, Matthew and Montani I. spaCy2: Natural language understanding with {B}loom embeddings, convolutional neural networks and incremental parsing. unpublished. 2017;

[16] Georgaphy3. Available from: https://github.com/somnathrakshit/geograpy3?tab=readme-ov-file

[17] Whiting PF, Rutjes AW d., Westwood ME, Malllett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. Ann Intern Med. 2011;155(4):529–161.