**Multi-LLM Framework for Automated Validation of RECORD Guideline Compliance in Observational Studies**

**Chenyu Li MS1, Seohu Lee BS MsE2, Yanshan Wang PhD1, Michael Becich MD PhD1,**

**Harold P Lehmann MD PhD2**

**1University of Pittsburgh School of Medicine Department of Biomedical Informatics, PA,**

**2Johns Hopkins University School of Medicine Biomedical Informatics and Data Science, Baltimore MD**

Introduction

Clinical epidemiologists and other methodologists have advocated for the use of methodology checklists for the past 40 years for improving the quality of published articles. These checklists result from consensus of experts in their respective methodology fields and are assembled in equator-network.org, for dissemination. However, too many articles are published with no such checklists provided, even in supplementary material. For instance, about half of observational-study manuscripts do not submit such checklists, 1such as STROBE2; the current checklist for EHR-based observational studies is RECORD (REporting of studies Conducted using Observational Routinely-collected Data).3

A survey of authors of observational studies suggested a number of barriers to adoption of these checklists, chief among them, the time taken to complete the checklist.4 Manual assessment of compliance is time-consuming and inconsistent, constituting a key barrier to adoption.4 We developed an LLM-based framework to automate validation of research papers against RECORD guidelines, enhancing transparency and reproducibility of observational research.

Methods

We designed a three-agent architecture leveraging complementary LLM capabilities:

* Reasoner (LLM1): extracts guideline items from RECORD documents and generates validation prompts
* Extractor (LLM2): identifies relevant information from research papers
* Validator (LLM3): validates extracted information against guidelines

The framework processes PDFs of research papers and automatically evaluates compliance with all 35 RECORD+STROBE items. Each agent performs a specialized task, with downstream agents building on previous outputs. We evaluated the system on randomly sampled 30 open-access observational studies published in medical journals from our previous study.5 We comparing its assessments against expert manual review. The compliance evaluation produces a structured JSON output detailing adherence to each item with supporting evidence.



Figure Framework Design

Results

Our framework demonstrated XXX accuracy compared to manual assessment across all 35 RECORD+STROBE items, with sensitivity (XXX) and specificity (XXX) for detecting non-compliant items. Papers achieved an average compliance rate of XXX%, with common deficiencies in data cleaning methods (item 12.2), code validation (item 6.2), and database linkage details (item 12.3). Automated processing required 5.8 minutes per paper versus 30 minutes for manual review, an XXX% time reduction.

The multi-LLM approach provided more robust assessments than single-LLM implementations, with LLM3 (Validator) identifying nuanced compliance issues that LLM2 extractor missed, particularly for methodology items. Inter-model agreement reached XXX of items, while human-LLM agreement was observed for XXX of validations.

*User Interface*

A screenshot of a black screen

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

Discussion and Conclusions

Our framework addresses a significant barrier to RECORD guideline adoption by drastically reducing assessment time while maintaining high accuracy. The triangulation approach using multiple LLMs improves robustness through complementary strengths of different models. Unlike previous approaches focused on checklist creation,6our system directly evaluates paper compliance and provides targeted improvement recommendations.

We have released our implementation as open-source software with a web-based user interface at https://github.com/ChenyuL/RWE\_LLM\_validator. The UI allows users to upload guideline PDFs and research papers, with a flexible architecture that supports the creation of new guideline folders, enabling application to various reporting standards using the same infrastructure. While we used RECORD guidelines to demonstrate the framework's effectiveness, the system is designed to work with any structured reporting guideline. The structured outputs facilitate conversion to FHIR-based representations, enabling integration with evidence systems like EBMonFHIR and supporting efforts to create computable representations of biomedical evidence.

References

1. Wiehn E, Ricci C, Alvarez-Perea A, et al. Adherence to the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) checklist in articles published in EAACI Journals: a bibliographic study. Published online March 12, 2021. doi:10.22541/au.161553910.02760640/v1

2. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies. *Int J Surg*. 2014;12(12):1495-1499. doi:10.1016/j.ijsu.2014.07.013

3. Benchimol EI, Smeeth L, Guttmann A, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *Z Evid Fortbild Qual Gesundhwes*. 2016;115-116:33-48. doi:10.1016/j.zefq.2016.07.010

4. Sharp MK, Glonti K, Hren D. Online survey about the STROBE statement highlighted diverging views about its content, purpose, and value. *J Clin Epidemiol*. 2020;123:100-106. doi:10.1016/j.jclinepi.2020.03.025

5. Li C, Alsheikh AM, Robinson KA, Lehmann HP. Use of recommended Real-World Methods for Electronic Health Record data analysis has not improved over 10 years. *medRxiv*. Published online June 22, 2023:2023.06.21.23291706. doi:10.1101/2023.06.21.23291706

6. Li C, Li S, Lehmann HP. An LLM-based Application to Create Methodology Checklists from Full-Text Articles. In: *Proceedings of the Annual Symposium on Artificial Intelligence for Medical Diagnosis and Healthcare*. ; 2023.

Abstract