

AI-Assisted and Human-Driven Biostatistical Workflows: A Task-Based Comparison Using Finnish Health Registry Data

Camilla Torriani⁽¹⁾, Juha Karjalainen⁽²⁾, Maria Cristina Monti⁽¹⁾, Tuomo Hartonen⁽²⁾, Andrea Ganna⁽²⁾

 Camilla Torriani 0009-0006-5879-6795

(1) Department of Public Health, Experimental and Forensic Medicine, Unit of Biostatistics and Clinical Epidemiology, University of Pavia, Pavia, Italy

(2) Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland

CORRESPONDING AUTHOR: ICamilla Torriani, Department of Public Health, Experimental and Forensic Medicine, Unit of Biostatistics and Clinical Epidemiology, University of Pavia, Pavia, Italy. E-Mail: camilla.torriani01@universitadipavia.it

INTRODUCTION

Recent advances in generative and agentic artificial intelligence are rapidly changing how biomedical research workflows are designed and conducted. AI-based assistants increasingly support coding, workflow automation, data analysis and scientific reporting. These tools could reduce time for routine analytical tasks and help researchers generate reproducible statistical pipelines. However, in biomedical and population-health research, analytical outputs must be not only fast and executable, but also methodologically appropriate, reproducible, interpretable and safe. A key open question is whether agentic AI tools can support biostatisticians without compromising scientific validity. We will address this question within ML4Health at FIMM, University of Helsinki, which combines health data with other information from nearly the whole Finnish population, including data from approximately 7 million individuals. The research data include diagnoses, disease treatments, medications, laboratory values, medical notes, home and institutional care, personal information, residential and marital history, pregnancies and births, education, job position, social assistance, and dates and causes of death. In this setting, Goose will be used as the agentic workflow interface, while Gemini will represent the generative engine driving code generation and reasoning support.

OBJECTIVE

We aim to develop and apply a task-based comparative framework to quantify the methodological validity, reproducibility and efficiency of an AI-assisted biostatistical workflow using Finnish health registry data.

METHODS

We will compare the AI-assisted workflow with a human-driven biostatistical workflow. We will define approximately 20 biostatistical tasks relevant to ML4Health and classify them into domains of increasing complexity. For each task, we will prepare an a priori methodological assessment sheet specifying input information, analytical objective, expected output and key sub-components required for a valid response. An expert-validated reference solution will also be prepared. Both workflows will receive the same input data or metadata, task description and analytical objective. For the AI-assisted workflow, we will use a standardized prompt chain and document prompts, outputs, corrections and decisions. Each output will be compared with the reference solution using a predefined scoring grid. Each sub-component will be scored as 2 = correct and complete or equivalent to the reference solution, 1 = partially correct or requiring minor correction, and 0 = incorrect, incomplete or requiring major correction. For each task, we will calculate total and percentage performance scores. Outputs will be classified according to predefined validity thresholds, considering both percentage score and major validity-threatening errors. Errors will be categorized by type and severity, including data-handling, coding, recoding, modelling, visualization, interpretation, reporting and reproducibility errors. The time required to generate, review and correct each output will also be recorded. Results will be summarized overall and by analytical domain.

RESULTS

We expect to identify domains in which AI-assisted workflows can reduce time and resource use while maintaining acceptable analytical quality, and those requiring substantial expert supervision. Expected outputs include task-level quality scores, error profiles and time-effectiveness summaries.

CONCLUSIONS

This evaluation will clarify the potential role of agentic AI as a supervised support tool for biostatisticians and may inform practical recommendations for transparent, reproducible and efficient AI-assisted analyses in large-scale health data research.