Statistics Innovation

ARTICLE

Open Access

https://doi.org/10.48130/stati-0026-0014

Statistics Innovation **2026**, 3: e017

Agentic T-STROBE: a front-shifted STROBE paradigm for observational research

Chunxin Bo[1,2], Ruimin Li[1,2], Jinghua Zhao[3*], Shiqing Feng[4*], Xiaokang Ji[1,2*] and Fuzhong Xue[1,2*]

[1] *Department of Medical Dataology, School of Public Health, Cheeloo College of Medicine, Shandong University, Jinan 250012, China*
[2] *National Institute of Health and Medical Big Data, Jinan 250003, China*
[3] *Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, Victor Phillip Dahdaleh Heart & Lung Research Institute, University of Cambridge, Cambridge, UK*
[4] *Department of Orthopaedics, The Second Qilu Hospital of Shandong University, Cheeloo College of Medicine, Shandong University, Jinan 250033, China*
* Correspondence: jhz22@cam.ac.uk (Zhao J); shiqingfeng@sdu.edu.cn (Feng S); jxk@sdu.edu.cn (Ji X); xuefzh@sdu.edu.cn (Xue F)

Abstract

The reproducibility crisis is a structural consequence of research complexity surpassing human capacity—a limitation of the 'artisanal science' paradigm. Current guidelines like STROBE function as retrospective *post hoc* checklists. We introduce Agentic T-STROBE, a framework that reframes methodological completeness and procedural rigor as structural constraints intrinsic to the research workflow, rather than criteria applied in *post hoc* evaluation. It employs (i) front-shifting STROBE to pre-emptive design constraints; (ii) a large language model (LLM)-driven dual-agent system programmatically executing workflows; and (iii) the Predictability–Computability–Stability (PCS) framework as an organizing evaluation lens. Agent-generated protocols (instantiated as AgenMed_CSS) demonstrated methodological completeness superior to 300 peer-reviewed studies (mean percent of maximum possible (POMP) score: 98.3% vs 68.3%; $p < 0.001$). This advantage was consistently observed across four LLM backends (mean POMP [95% CI]: 99.2 (98.5–99.8), 98.7 (98.0–99.3), 99.4 (99.0–99.8), and 97.2 (96.0–98.3)), suggesting that the observed methodological completeness is more contingent on constraint architecture than on the capabilities of the underlying model. The statistical execution agent achieved > 92% fidelity. Conversely, unconstrained human-AI collaboration yielded inferior quality (mean POMP: 63.1 [56.9–69.2]). Agentic T-STROBE establishes a prospective, auditable oversight framework for methodological completeness and protocol transparency, without substituting scientific judgment.

Keywords: AI agents, Large language model, Statistical workflow automation, Research reproducibility, STROBE guideline, Observational study

Citation: Bo C, Li R, Zhao J, Feng S, Ji X, et al. 2026. Agentic T-STROBE: a front-shifted STROBE paradigm for observational research. *Statistics Innovation* 3: e017 https://doi.org/10.48130/stati-0026-0014

Introduction

The discourse regarding the reproducibility crisis has dominated metascientific inquiry for more than a decade, driven by the alarming frequency with which published findings fail independent verification[1–4]. While conventional diagnoses often focus on individual researcher negligence, p-hacking, or the perverse incentives of the publish-or-perish culture, these explanations remain superficial. A deeper, structural pathology exists: the researcher-dependent workflow[5–7] that currently underpins scientific production has reached its cognitive limit. In this 'artisanal' mode, complex, multistage methodological tasks are contingent upon the tacit knowledge, cognitive stability, and memory of individual researchers[8]. As the complexity of modern scientific inquiry grows exponentially, this 'human-centric variability' has solidified as the fundamental bottleneck restricting research reliability and cumulativity[8].

To mitigate this variability, the academic community has invested heavily in standardization, most notably through the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement[9]. STROBE and its extensions provide a comprehensive checklist covering critical aspects of research design, execution, and analysis, intended to enhance transparency and completeness. While STROBE represents a milestone in methodological dissemination[10,11], its utility is inherently limited by its nature as a descriptive, *post hoc* reporting guideline[12,13]. It delineates how to describe research clearly, not how to execute it correctly. It functions as a map rather than a navigation system; a blueprint rather than a construction robot. Consequently, a significant practice-norm gap persists, in which adherence to guidelines remains voluntary and inconsistent.

The advent of artificial intelligence has introduced the concept of the Large Language Model (LLM) as a research 'co-pilot'[14–16]. This prevailing model typically relies on unstructured, free-form interactions between researchers and general-purpose models to accelerate specific tasks such as literature review[17], code generation[18], or manuscript refinement[19]. However, this mode fails to address the core issue of methodological reliability. In the absence of structured methodological frameworks and rigid constraint mechanisms, unconstrained human–AI interaction cannot guarantee methodological correctness[20–22]. Indeed, by introducing model hallucinations[23–25], exacerbating user knowledge blind spots, and adding the stochasticity of conversational interaction, this model risks amplifying rather than reducing human-centric variability.

The emergence of LLM-based agentic systems—capable of autonomous planning, tool invocation, multi-step reasoning, and self-correction—represents a qualitative advance beyond the co-pilot paradigm[26,27]. In the biomedical domain, agentic frameworks have demonstrated early promise across a range of tasks: BioMedAgent autonomously chains bioinformatics tools into executable workflows for cross-omics analysis and machine-learning modelling[27]; agentic systems in radiology automate report generation and protocol optimization[28]; and multi-agent LLMs have been applied to collaborative hypothesis generation in drug discovery[29]. Comprehensive reviews synthesizing over 60 emerging agentic systems consistently identify reproducibility and methodological grounding as the central unresolved challenges[26].

© The Author(s)

https://doi.org/10.48130/stati-0026-0014

Critically, however, existing agentic systems share a common limitation: they are designed for task execution within established analytical pipelines, not for *a priori* methodological constraint enforcement in the research design phase. None of the existing systems encodes domain-specific reporting guidelines (such as STROBE) as executable design rules, nor do they generate a complete, auditable computational lineage spanning from research question formulation to statistical analysis execution. The gap between automated task completion and guaranteed methodological rigor in observational epidemiology therefore remains unaddressed.

We argue that the necessary evolution in scientific rigor demands a paradigm shift from descriptive guidelines to an Epistemological Operating System. Such a system must transform methodological best practices from passive text into active, executable contracts. To this end, we propose and validate Agentic T-STROBE (Agentic Target-STROBE Emulation), instantiated here as AgenMed_CSS for cross-sectional studies (CSS). To our knowledge, this system represents the first agentic framework specifically designed to enforce *a priori* methodological rigor in observational research design—distinguished from existing systems by its integration of front-shifted reporting guidelines as executable constraints, its dual-agent planning-execution architecture, and its generation of a complete, auditable computational lineage. It is not merely a tool but an engineered paradigm designed to make methodological rigor an *a priori* system property.

The Agentic T-STROBE paradigm is founded on three synergistic principles:

1. Agentic Target-STROBE Emulation (Agentic T-STROBE): This principle involves front-shifting the STROBE checklist. Instead of serving as a retrospective check on a finished manuscript, STROBE items are reconstructed into a set of mandatory, executable design rules and deliverables—specifically, the research protocol, statistical analysis plan (SAP), and simulation tables—that govern the research lifecycle from its inception.

2. Agentic Execution: We construct a dual-agent system that programs these rules and complex methodological decision trees into a researcher-supervised automated workflow. Within this workflow, structured user input and protocol review precede automated statistical analysis, and complete auditable logs are generated throughout, supporting high-fidelity implementation of methodological requirements.

3. PCS (Predictability–Computability–Stability) as a metric of veracity: Drawing upon the theoretical framework of veridical data science[30], we integrate PCS as an organizing framework for process-level reliability assessment. This facilitates a shift from subjective peer review to quantitative quality assurance.

Through comparative experimentation, we evaluate whether protocol completeness and execution fidelity are primarily supported by the structured workflow rather than by a specific LLM backend or unconstrained human-AI collaboration. The results indicate that methodological completeness functions as a property of the workflow that can be specified, audited, and reproduced across different models and users, rather than as an attribute contingent on individual memory or *ad hoc* judgment.

Materials and methods

AgenMed_CSS system architecture

Overall design

The AgenMed_CSS system is architected as an end-to-end 'planning–execution' dual-agent cognitive workflow. This framework is explicitly designed to decompose the complexity of scientific inquiry into two distinct, specialized operational phases: a high-level strategic planning phase, governed by the Planning Design Agent, which is responsible for the formulation of rigorous research protocols; and a low-level implementation phase, governed by the Execution Analysis Agent, which is tasked with precise code generation and local execution.

Planning design agent

To anchor the LLM's inferential reasoning within strict methodological boundaries—thereby neutralizing the risk of hallucination—we encoded the front-shifted STROBE workflow into a granular decision topology. This knowledge graph serves as a deterministic scaffold, strictly confining the LLM's execution trajectory to validated standard operating procedures throughout the entire research lifecycle (see Supplementary File 1, Material S1).

We developed AgenMed_CSS as a multi-module collaborative system comprising modular prompt chains (see Video S1 for the workflow structure and node-level prompt configuration) built on the n8n agent-workflow platform (https://github.com/n8n-io/n8n). In this agent-workflow, the decision tree (see Supplementary File 1, Material S1) was decomposed into 52 nodes, each with a precisely specified LLM role, task, methodological guidance, and standardized JSON outputs stored in PostgreSQL for real-time retrieval. This architecture enables high-fidelity, end-to-end cross-sectional study design and analysis while ensuring full traceability, auditability, and transparency.

For a given cross-sectional study design and analysis project (e.g., a population-based cross-sectional survey of bone mineral density among adults aged ≥ 65 years in Shandong Province, China), the workflow begins with researcher input rather than autonomous question generation. Users first select an appropriate question-structuring framework according to their research objective: PECO (population, exposure, comparison, outcome) for association-oriented questions; CoCoPop (condition, context, population) for prevalence estimation; and SPIDER (sample, phenomenon of interest, design, evaluation, research type) for qualitative or mixed-methods research. A combined PECO–CoCoPop structure is also available for studies that pursue both objectives. Users then complete the corresponding structured fields, and these researcher-specified inputs anchor the planning design agent's subsequent protocol generation (see Video S2 for a demonstration of protocol generation and statistical analysis, and Supplementary File 1, Material S4 for the research protocol).

The generated protocol is organized into nine chapters: Title and abstract, Background and objectives, Study design and setting; Participants; Variable definitions and data sources/measurement; Bias control and sample size; Statistical methods; Ethical statement and funding; Research results publication and dissemination plan. Within these chapters, the 22 STROBE_CSS items are expanded into 32 scorable protocol elements (see Supplementary File 1, Material S2 and S8). Specifically, the above full-process decision tree (see Supplementary File 1, Material S1) encompasses the entire research lifecycle—including research question framework selection, definition of target/source populations and sampling frames, sampling design and sample size calculation, variable measurement and tool selection, bias control, missing data handling, association/regression modeling, and results reporting. Furthermore, predefined statistical decision tree rules, constructed based on the principles of statistical description and inference for cross-sectional studies[31], are embedded within the relevant modules (see Supplementary File 1, Material S5). These rules enforce mandatory specifications, such as assumptions and methods for sample size calculation, strategies for missing data handling, schemes for complex survey weight applica-

tion, and protocols for model specification and diagnosis. The decision tree serves as the default methodological pathway, while justified deviations required by the study question, data structure, sampling design, or model assumptions are permitted with documented rationale and audit-trail preservation.

By encoding methodological constraints into reusable decision rules, this structured design aims to maximize the reduction of methodological fluctuations arising from natural language ambiguity, thereby ensuring a predictable baseline for the generated protocol regarding terminology usage, content coverage, and methodological consistency. Importantly, this constraint structure does not remove the researcher from the design process. The generated protocol functions as a reviewable artifact through which investigators can refine population definitions, variable specifications, covariate assumptions, and the SAP before statistical execution begins.

Execution analysis agent

Operating within a user-localized R/Python runtime environment, this agent executes a closed-loop 'Parse-Generate-Execute-Repair' workflow by invoking external large language models (LLMs) via real-time API calls to parse protocols and generate code.

It accepts the researcher-reviewed statistical analysis plan (SAP), embedded within the protocol generated by the planning design agent, and the local dataset as primary inputs. First, the system parses the structured constraints of the SAP to generate hierarchical sub-task prompts. Subsequently, R code is generated based on these prompts and metadata (e.g., variable names, types, and missing data profiles). Crucially, to mitigate randomness in code generation and ensure consistency with the SAP, the code generation phase employs a low temperature setting (temperature = 0.1) for the LLMs (e.g., Claude-Sonnet-4, ChatGPT-5)[32,33].

Prior to execution, the system enforces a pre-check gateway defined by protocol-specified methodological constraints, including any approved deviations recorded in the SAP. Code passing this pre-check is executed within the local environment. In the event of execution failure, error logs are captured and categorized, triggering an automated repair loop (with a maximum of 10 iterations). The system retries using a minimal modification strategy until successful execution is achieved (see Supplementary File 1, Material S6). The resulting code, tables, figures, repair traces, and execution logs are then available for investigator review before the statistical report is finalized.

Data security and provenance mechanism

The entire workflow of the execution analysis agent strictly adheres to a 'Data Does Not Leave the Domain' principle. External LLM APIs receive only desensitized metadata (e.g., variable lists) and error logs; raw data remains permanently resident within the local environment. All artifacts generated during the process—including multi-level prompts, model versions and parameters, generated code scripts, execution results (tables and figures), accepted methodological deviations, and comprehensive system logs—are automatically archived into independent working directories named according to chapter numbering. This creates a complete, immutable chain of evidence suitable for peer review and future reproduction (see Supplementary File 1, Material S7).

PCS framework for veridicality assessment

To evaluate the reliability of the research outputs, this study adopts the PCS (predictability, computability, stability) framework from veridical data science as its theoretical foundation.

Predictability evaluation

Predictability is operationalized as a reality check in three dimensions.

First, regarding the predictability evaluation of the intelligent design protocol produced by the planning design agent, we initially refined the STROBE_CSS Statement into a 32-item quantifiable scoring rubric[34,35] (see Supplementary File 1, Material S8). Using this rubric as the comparative benchmark, we uploaded both the rubric and the target text—either the original published articles or the intelligent design protocol produced for a specific query—to the LLM (e.g., Gemini 2.5 Pro). Guided by the established prompt engineering (see Supplementary File 1, Material S9), the system conducted a quantitative performance evaluation of the intelligent design protocols derived from original published articles (or those generated *de novo* for specific queries).

We emphasize that the STROBE_CSS-based POMP score was designed to quantify checklist-based methodological completeness and the correctness of predefined methodological steps, not to assess intrinsic scientific validity. The 32-item rubric evaluates whether key design and analytical elements are explicitly addressed and methodologically appropriate, including sample-size justification, bias-control strategy, missing-data handling, and statistical method specification. A high POMP score, however, does not indicate that the research question is clinically important, biologically plausible, or free from unmeasured confounding. These dimensions necessarily depend on domain expertise, causal reasoning, data quality, and external evidence, and lie beyond the scope of checklist-based evaluation. Accordingly, POMP is best understood as a process-level quality indicator, not a comprehensive measure of scientific merit.

Second, regarding the predictability evaluation of the execution analysis agent, we utilized the 22-item SAP evaluation framework (see Supplementary File 1, Material S10)—operationalized from the statistical analysis section (see Supplementary File 1, Material S4, Chapter 7) of the planning design agent's protocol—as the comparative scoring rubric. We uploaded this rubric alongside the code and statistical results output by the execution analysis agent to the LLM (e.g., Gemini 2.5 Pro). Based on the established prompt engineering (see Supplementary File 1, Material S11), the system performed a quantitative performance evaluation of the agent's code and statistical outputs.

Crucially, to assess the objectivity and validity of the evaluation results produced by the LLM (e.g., Gemini 2.5 Pro), we selected 10 senior experts in epidemiology or biostatistics independent of the study team. We paired these experts 1:1 based on professional background (epidemiology or biostatistics) and similarity in age (within 1–5 years), resulting in two groups of five experts each. For each pair, we generated two random numbers using a random number table and ranked them; the expert with Rank 1 was assigned to the experimental group, and the expert with Rank 2 to the control group.

Regarding the expert-validation sample size, we performed an exact noncentral-t power calculation for a paired t-test with two-sided $\alpha = 0.05$, $1 - \beta = 0.95$, and a conservative anticipated paired standardized effect size of Cohen's dz = 1.50, chosen below the effect observed in the 300-pair benchmark comparison (overall dz = 2.50; stratum-specific dz = 2.30, 3.20, and 3.59). The minimum required sample was eight article–protocol pairs; we selected nine pairs to support balanced stratified random sampling, with three articles independently and randomly selected from each JCR quartile stratum (Q1, Q2, and Q3). With $n = 9$ and dz = 1.50, the resultant power was approximately 0.975.

https://doi.org/10.48130/stati-0026-0014

1. The five experts in the experimental group evaluated the performance of the intelligent design protocols (generated for the original published articles), using the original published articles as background reference and strictly adhering to the aforementioned 32-item STROBE_CSS criteria.

2. Conversely, the control group experts directly evaluated the research design and analytical performance of the original published articles, using the same criteria and the original published articles as background reference.

For the comparison of evaluation scores between paired experts, we employed paired t-tests. For reliability analysis, we used ICC(A,5) (two-way random, average measures) to assess Inter-Rater Reliability (IRR) among the five experts within the experimental (or control) group. To assess automated–expert calibration, we compared each object's automated POMP score (Gemini 2.5 Pro) with the corresponding five-rater expert-consensus POMP score across 18 objects. Agreement was summarized using Pearson r, Lin's CCC, mean bias, and MAE. For the IRR analysis, ICC (A,5) was used, and relaxed item-level agreement was additionally reported because protocol scores approached the scale ceiling.

Computability evaluation

To assess the operational feasibility of AgenMed_CSS under standard computational resource constraints (MacBook Air, 13-inch, 2025; Apple M4; 16 GB RAM; 512 GB SSD), the system was deployed on a standard laptop configuration. The system was run using real-time API calls to LLMs such as Claude-Sonnet-4 and ChatGPT-5. Operational logs were monitored to track metrics including the first-pass correctness rate of code generation for major analysis stages, the efficiency of automatic repair, the accuracy of key method code calls, the overall process success rate, and total execution time. These metrics were used to evaluate the transparency, auditability, and engineering robustness of the computational process, verifying its practical utility (see Supplementary File 1, Material S13).

Output consistency evaluation

This analysis was inspired by the PCS perturbation principle but was not a full PCS stability analysis: the underlying dataset, variable definitions, model specifications, and analytical decisions were not perturbed. Inferential tests were used only as exploratory comparisons and were not interpreted as evidence of stability or equivalence.

1. Planning output consistency: For a single standardized observational research project (see Supplementary File 1, Material S14), the planning design agent iteratively called four widely used LLMs (Claude-Sonnet-4, Gemini 2.5 Pro, ChatGPT-5, and DeepSeek-V3), 50 times each. Welch's ANOVA was conducted to compare POMP scores across backends, with Games–Howell post-hoc tests for pairwise comparisons. Results are reported descriptively as group means, 95% CIs, and absolute between-backend differences.

2. Execution output consistency: Targeting the statistical analysis section of the generated protocol (see Supplementary File 1, Material S4, Chapter 7), the execution analysis agent iteratively called two LLMs (Claude-Sonnet-4 and ChatGPT-5) 50 times each. Welch's t-test was used to compare the performance of the output statistical analysis code and results based on the 22 SAP evaluation metrics.

Validation research design

The validation study links each comparison object with a corresponding evaluation metric and a pre-specified statistical method. The published-article comparison assesses whether agent-generated protocols achieve higher STROBE_CSS completeness than matched peer-reviewed cross-sectional studies, evaluated using POMP scores and paired t-tests. The cross-LLM comparison assesses whether outputs remain consistent across different LLM backends, evaluated using POMP scores, execution-fidelity scores, and Welch's ANOVA. The unconstrained human-AI collaboration comparison assesses whether the structured agentic workflow outperforms general-purpose LLM use without the Agentic T-STROBE framework, evaluated using POMP scores and independent-sample t-tests. These components are described below.

Comparison objects

Published articles

A systematic search was conducted in the PubMed database for cross-sectional study literature published from 2020 to 2025 in JCR Q1, Q2, and Q3 journals (see Supplementary File 1, Material S15). Using simple random sampling, 100 articles were selected from each partition (total n = 300). Based on the background, objectives, and materials and methods of each article, the planning design agent generated a complete intelligent design protocol under the constraints of the Agentic_STROBE_CSS prompt engineering framework developed in this study (see Supplementary File 1, Materials S1 and S2).

Cross-LLM output consistency

On the one hand, for the same observational study project (see Supplementary File 1, Material S14), the planning design agent generated 50 intelligent design protocols each (total n = 200) by iterating through calls to four LLMs: Claude-Sonnet-4, Gemini 2.5 Pro, ChatGPT-5, and DeepSeek-V3, under the Agentic_STROBE_CSS prompt constraints (see Supplementary File 1, Material S1).

On the other hand, for the statistical analysis section of the generated protocols (see Supplementary File 1, Material S4), the execution analysis agent generated statistical analysis code and results 50 times each (total n = 100) by iterating through calls to Claude-Sonnet-4 and ChatGPT-5.

Unconstrained human-AI collaboration

Thirty-two students majoring in medical dataology (19 seniors and 13 graduate students), trained in epidemiology, biostatistics, and data science, were recruited. They were tasked with independently generating a design protocol for the same observational study project (see Supplementary File 1, Material S14) using any general-purpose LLM (platform and model unconstrained). No Agentic_STROBE_CSS prompt engineering framework (see Supplementary File 1, Material S1) or any templates, scoring criteria, or example schemes were provided. The resulting schemes served as the baseline comparator representing the currently prevalent unconstrained human-AI collaboration mode.

Evaluation metrics

STROBE_CSS methodological completeness

Each original published article (or intelligent design protocol) was evaluated using the 32-item STROBE_CSS scoring rubric (see Supplementary File 1, Material S8). This rubric assesses reporting completeness and the methodological appropriateness of predefined design and analysis elements, rather than the intrinsic scientific merit of the research question. Each item was scored on a three-point scale: 0 = missing or inappropriate; 1 = partially achieved; 2 = fully achieved and methodologically correct. The raw score R_i for the i^{th} article (or protocol) is defined as: $R_i = \sum_{j=1}^{32} s_{ij}$,

where $s_{ij} \in \{0,1,2\}$, with a maximum possible score of 64. To facilitate comparison, raw scores were linearly standardized to a 0–100 scale using the POMP (percent of maximum possible) method[36]:

$$POMP_i = 100 \times \frac{R_i}{64} \quad (1)$$

Higher values indicate superior performance by the planning design agent. Scoring was executed by an LLM (e.g., Gemini 2.5 Pro) based on established prompt engineering (see Supplementary File 1, Material S9).

Statistical execution compliance

The execution fidelity was quantitatively evaluated using 'achieved/not achieved' ratings against the 22 evaluation metrics of the SAP (see Supplementary File 1, Material S10) derived from the statistical analysis section of the intelligent design protocol. The compliance rate (on a 100-point scale) for a single run r is defined as:

$$A_r = 100 \times \frac{1}{22} \sum_{k=1}^{22} c_{rk} \quad (2)$$

where $c_{rk} \in \{0,1\}$ represents the judgment result for the k^{th} checkpoint in the r^{th} run. Higher values indicate better performance by the execution analysis agent. Scoring was executed by an LLM (e.g., Gemini 2.5 Pro) based on established prompt engineering (see Supplementary File 1, Material S11).

Statistical analysis

All statistical comparisons were two-sided with a significance threshold of $\alpha = 0.05$.

Paired comparisons (intelligent design protocol vs. published articles)

For the published-article comparison, paired t-tests were used to compare the intelligent design protocols with the original published articles. Effect sizes (Cohen's dz) and 95% confidence intervals (CI) for the mean difference were calculated. Forest plots were generated to present item-level mean scores and their 95% CIs.

Cross-model consistency

For cross-LLM output consistency, due to equal sample sizes but potential heterogeneity of variance, Welch's ANOVA was used for global testing, with Games–Howell tests for post-hoc pairwise comparisons. Results reported include *p*-values, maximum absolute inter-group differences, group means, and 95% CIs.

Execution performance

For the execution analysis agent, Welch's t-test was used to compare performance on the 22 SAP metrics between the two LLMs (Claude-Sonnet-4 vs ChatGPT-5). Item-level achievement proportions and 95% CIs were reported.

System vs human-AI

Independent sample t-tests were used to compare differences between the four LLM-generated protocol groups and the unconstrained human-AI collaboration group, reporting POMP scores and 95% CIs for both groups.

Computability

Descriptive statistics were reported for the AgenMed_CSS computability metrics, including first-pass code generation correctness, automatic repair efficiency, key method code call accuracy, process success rate, and execution time (see Supplementary File 1, Material S13).

Results

AgenMed_CSS system overview

The AgenMed_CSS system is engineered as an end-to-end dual-agent workflow, accessible via a minimalist user interface (Fig. 1). The cognitive process follows a strict planning-execution dichotomy (see Supplementary File 1, Material S3).

The workflow commences with the user providing a natural language description of their research concept. The planning design agent (Fig. 2a) engages first, parsing the user's intent through established problem frameworks (e.g., PECO or CoCoPop). It generates a structured draft of design information, which is then iteratively refined into a final, highly normative intelligent design protocol comprising nine chapters. Critically, Chapter 7 of this protocol, 'Statistical Analysis', functions as a precise, machine-readable instruction set for the subsequent phase (see Supplementary File 1, Material S4).

Upon protocol finalization, the user submits the intelligent design protocol and the local dataset to the execution analysis agent (Fig. 2b). This agent operates strictly within the user's local computing environment to ensure data privacy. It executes a closed-loop task stream: first, it parses the instructions in Chapter 7 (see Supplementary File 1, Material S7), decomposing them into independent sub-task modules (including data preprocessing, missing value treatment, descriptive statistics, and regression modeling). Subsequently, it generates the corresponding R code for each sub-task and executes it locally. The agent incorporates a self-repair loop; if code execution fails, the system captures the error log and iterates (up to a pre-set limit) to correct the syntax or logic. Throughout this process, every prompt, generated code snippet, execution result (tables and figures), and error log is automatically archived, establishing a complete, traceable chain of evidence (see Supplementary File 1, Material S7). Raw data is never transmitted over the external network (see Supplementary File 1, Material S16).

Higher STROBE_CSS methodological completeness than peer-reviewed literature

To quantify STROBE_CSS-concordant methodological completeness, we conducted a benchmark comparison between agent-generated protocols and peer-reviewed cross-sectional studies. We randomly extracted 300 cross-sectional studies published in journals ranked JCR Q1, Q2, and Q3 from the PubMed database. The planning design agent was tasked with generating a *de novo* intelligent design protocol for each study, based strictly on the background, objectives, and materials and methods descriptions extracted from the original articles.

We evaluated both the 300 original published articles and the 300 system-generated protocols using the STROBE_CSS extended checklist, which comprises 32 distinct, quantifiable items (see Supplementary File 1, Material S8). Evaluations were performed by Gemini 2.5 Pro, standardized using the percent of maximum possible (POMP) scoring method.

The results revealed a substantial gap in STROBE_CSS-concordant methodological completeness (Fig. 3a). Under identical conditions of background and objective, the protocols generated by the planning design agent achieved markedly higher STROBE_CSS-concordant completeness than the published articles evaluated. The system-generated protocols achieved a mean POMP score of 98.3%, whereas the published articles averaged only 68.3%. A paired t-test confirmed this difference to be statistically significant ($p < 0.001$).

Analysis at the item level indicates that the intelligent design protocols achieved near-perfect scores across almost all inspection

https://doi.org/10.48130/stati-0026-0014

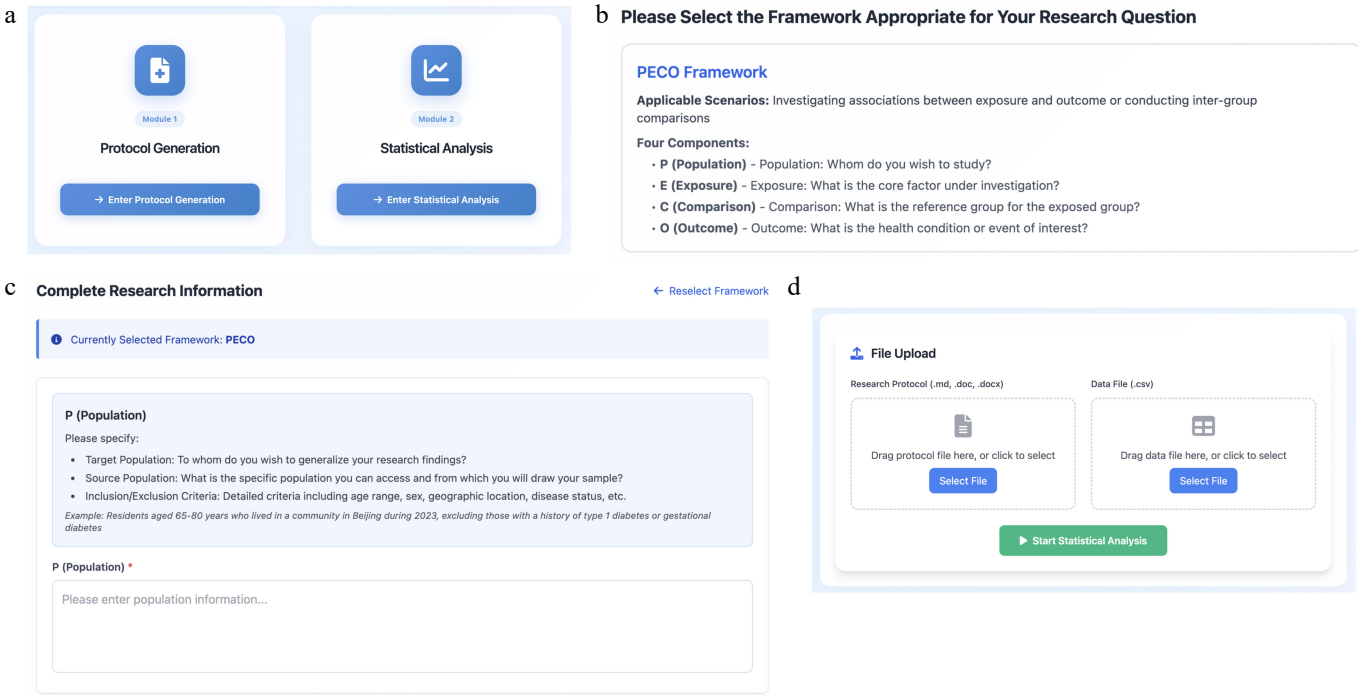


Fig. 1 Minimalist user interface and end-to-end workflow of AgenMed_CSS. (a) Landing page where users select either Module 1 (Protocol Generation) or Module 2 (Statistical Analysis), reflecting the system's 'planning–execution' dichotomy. (b) In Module 1, the Planning Design Agent starts by guiding users to select an appropriate question framework (e.g., PECO) to structure the research intent. (c) The user then completes structured research information according to the chosen framework, which is used to generate and iteratively refine a standardized intelligent design protocol. (d) In Module 2, the Execution Analysis Agent accepts the finalized protocol and a local dataset for analysis, initiating a closed-loop analysis that ensures data privacy and full traceability.

points. In contrast, the published articles displayed systematic deficiencies in reporting, particularly regarding specific research design details, strategies for bias control, detailed exposition of statistical methods, handling of missing data, and the basis for sample size calculations (Fig. 3a). Subgroup analysis by journal rank (Fig. 3b) indicated that while high-impact (Q1) journals exhibited marginally better reporting quality than lower-quartile journals, the published articles across all partitions remained significantly inferior to the output of the agent. This finding quantifies the 'practice-norm gap' inherent in the human-dependent scientific production system.

Validation of automated evaluation

To verify the objectivity and validity of the LLM-based evaluation, we engaged 10 senior experts in epidemiology and biostatistics. These experts were paired 1:1 based on professional background and randomly assigned to an experimental group or a control group. They reviewed nine article-protocol pairs selected via stratified random sampling, with three pairs drawn from each JCR quartile stratum (Q1, Q2, and Q3).

The expert consensus POMP scores corroborated the automated evaluation (Table 1). The intelligent design protocols received a mean expert score of 95.10 ± 2.39, exceeding the corresponding published articles (75.03 ± 12.11). The mean paired difference was 20.07 POMP points (SD of paired differences = 12.05; observed Cohen's dz = 1.67; $t(8) = 5.00$, $p < 0.01$), substantially exceeding the conservative anticipated effect size used in the *a priori* power calculation (dz = 1.50).

Inter-rater reliability (IRR) and automated–expert calibration are summarized in Table 1. Published articles showed high five-rater average-measures reliability (ICC [A,5] = 0.92; 95% CI 0.79–0.98). For agent-generated protocols, ICC (A,5) was not interpretable (–0.25; 95% CI –0.99–0.55) because near-ceiling scores restricted between-object variance. Relaxed item-level agreement was higher for protocols than for articles (91.0% vs 76.7%). Across all 18 objects, automated and expert-consensus POMP scores showed strong association (Pearson r = 0.91, $p < 0.001$) and good concordance (Lin's CCC = 0.83). Mean bias was –3.30 POMP points overall, –9.58 for articles, and +2.99 for protocols, indicating object-type-specific calibration differences. Thus, the main conclusion is supported by independent expert consensus rather than automated scoring alone.

Architecture over intelligence: system robustness

A critical question regarding AI systems is whether performance is an artifact of a specific model's intelligence or a result of the system's design. To address this, we assessed the output consistency of the planning design agent by interchanging its underlying LLM backend. We employed four leading large language models: Claude-Sonnet-4, Gemini 2.5 Pro, ChatGPT-5, and DeepSeek-V3. Each model was tasked with generating 50 intelligent design protocols for a single standardized observational research project (see Supplementary File 1, Material S14).

All backends achieved high protocol completeness (Fig. 4), with mean POMP scores of 99.2 (95% CI, 98.5–99.8) for Gemini 2.5 Pro, 98.7 (98.0–99.3) for Claude-Sonnet-4, 99.4 (99.0–99.8) for ChatGPT-5, and 97.2 (96.0–98.3) for DeepSeek-V3. Welch's ANOVA detected a statistically significant between-backend difference ($F[3, 103.37] = 5.18$, $p = 0.002$), driven primarily by the slightly lower and more variable DeepSeek-V3 outputs, though all backends maintained completeness above 97%. No backend exhibited systematic methodological defects at the item level.

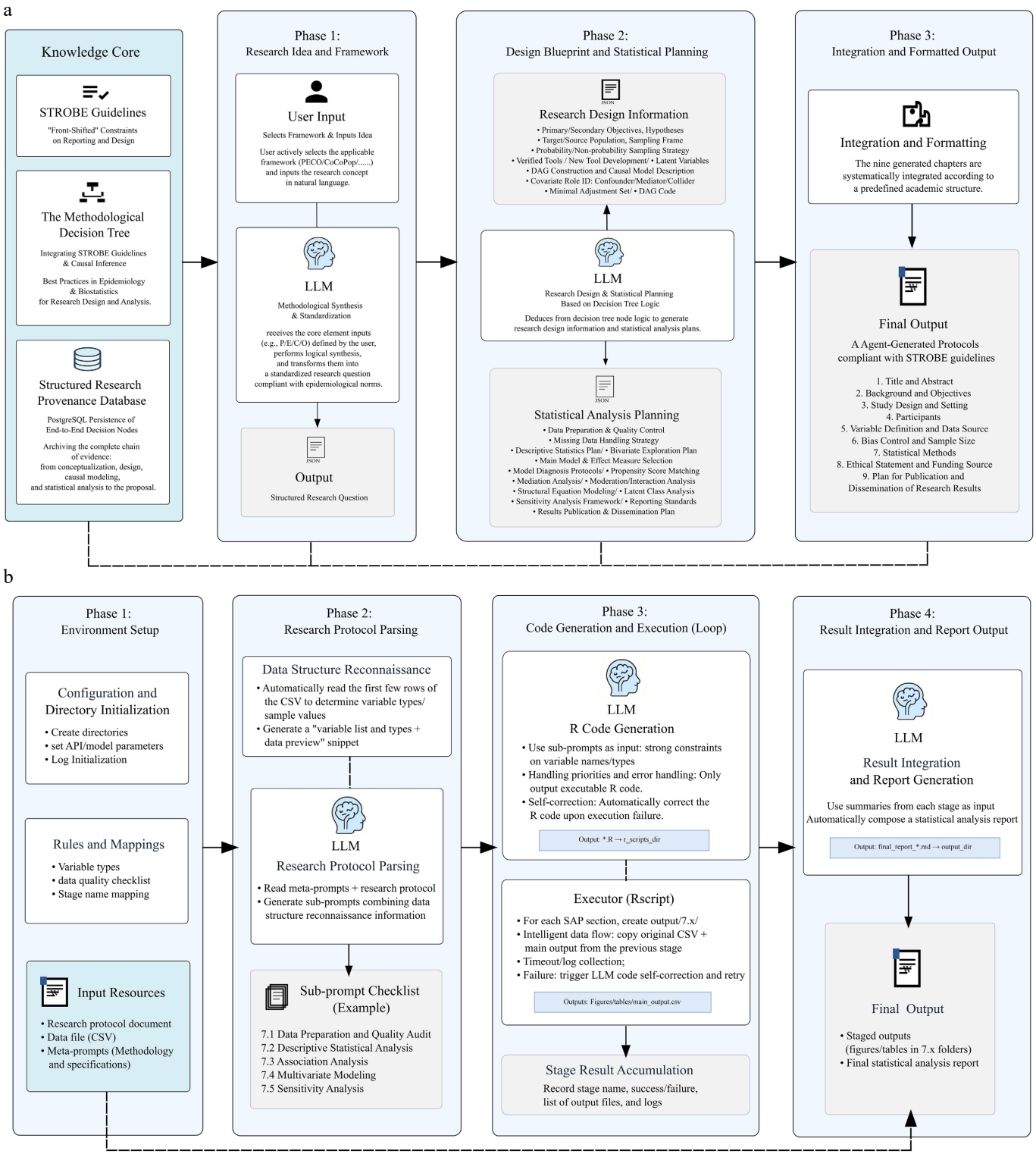


Fig. 2 Dual-agent architecture for reproducible, privacy-preserving epidemiological research. (a) Planning design agent. User input is structured via a question framework (e.g., PECO/CoCoPop) and translated into a standardized intelligent design protocol; the statistical analysis chapter (see Supplementary File 1, Material S7, Chapter 7) serves as the executable analysis specification. (b) Execution analysis agent. The protocol and a local dataset are ingested to run a staged analysis: protocol parsing, iterative R code generation/execution with error-driven self-repair, and integration of figures/tables into the final report.

This finding provides evidence for architectural robustness. The system's methodological rigor is not derived from the intrinsic capabilities of any single LLM but from the front-shifted STROBE rules and the modular prompt chains that constitute the system's constraint framework. This suggests that structure is a more critical determinant of research rigor than model intelligence.

https://doi.org/10.48130/stati-0026-0014

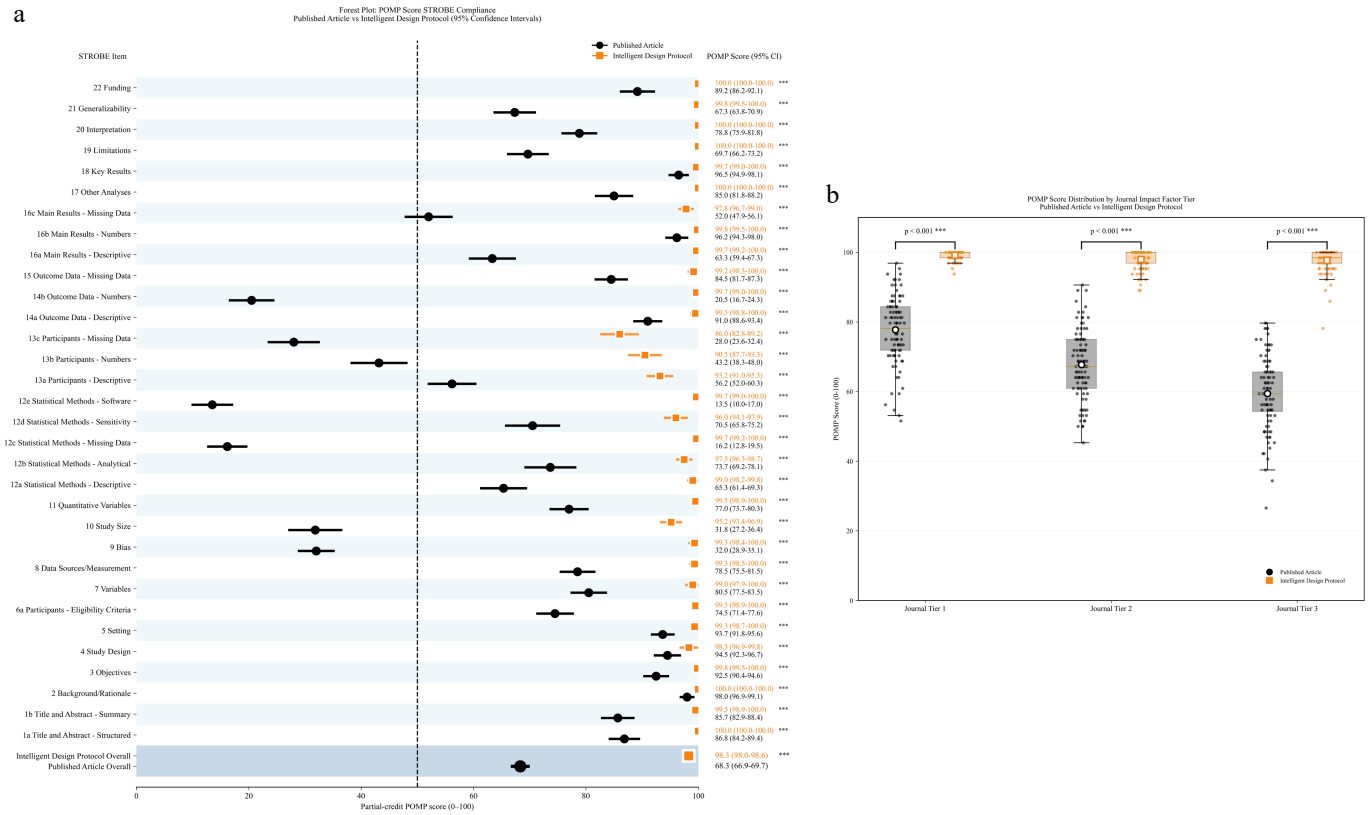


Fig. 3 Benchmarking methodological rigor: agent-generated protocols vs. peer-reviewed articles. (a) Forest plot displaying item-level compliance with the STROBE_CSS checklist. Data points represent the mean percent of maximum possible (POMP) score, and error bars indicate 95% confidence intervals (CIs). The comparison contrasts the planning design agent's output (orange) with the original matched peer-reviewed studies (black; $n = 300$). The bottom rows present the aggregate methodological quality scores for both groups. (b) Distribution of overall POMP scores stratified by journal ranking (tiers 1, 2, and 3 correspond to JCR Q1, Q2, and Q3). Box plots show the median (center line), upper and lower quartiles (box limits), and 1.5× interquartile range (whiskers). *p*-values were calculated using two-sided paired t-tests ($p < 0.001$). The source data is provided in Supplementary File 2.

Table 1. Expert validation of protocol completeness and automated scoring calibration.

Section	Metric	All ($n = 18$)	Articles ($n = 9$)	Protocols ($n = 9$)
A. Expert consensus POMP scores	Mean ± SD (%)	–	75.03 ± 12.11	95.10 ± 2.39
	95% CI	–	65.72–84.35	93.27–96.94
B. Inter-rater reliability	ICC (A,5) (95% CI)	–	0.92 (0.79, 0.98)	−0.25 (−0.99, 0.55)
–	Relaxed agreement	–	76.70%	91.00%
C. Automated–expert calibration	Pearson r	0.91 ($p < 0.001$)	–	–
–	Lin's CCC	0.83	–	–
–	Mean bias	−3.3	−9.58	2.99
–	MAE	6.74	10.49	2.99

Data are presented as mean ± SD or point estimates with 95% confidence intervals (CIs). Section A quantifies the methodological completeness gap, where five-rater expert consensus confirms that agent-generated protocols outperform matched peer-reviewed articles. Section B reports inter-rater reliability and relaxed item-level agreement. The anomalous ICC value for the protocol group reflects a ceiling effect caused by restricted between-object variance, as experts consistently rated these protocols near the maximum score; it should therefore be interpreted together with the relaxed agreement metric. Section C reports automated–expert calibration at the object level using Pearson r, Lin's concordance correlation coefficient (CCC), mean bias, and MAE. Abbreviations: CCC, Lin's concordance correlation coefficient; ICC, intraclass correlation coefficient; MAE, mean absolute error; POMP, percent of maximum possible score. The source data is provided in Supplementary File 2.

Performance of unconstrained AI collaboration

We explicitly contrasted the Agentic T-STROBE paradigm with the currently popular unstructured AI co-pilot mode. We recruited 32 students (19 seniors, 13 graduate students) trained in epidemiology, biostatistics, and data science. This group utilized general-purpose LLMs without the specific constraints of AgenMed_CSS to design protocols for the same standardized project (see Supplementary File 1, Material S14).

The disparity in outcomes was stark (Fig. 5). The AgenMed_CSS system (pooled across four models, $n = 200$) achieved a robust POMP score of 98.6 (95% CI, 98.2–99.0). In contrast, the unconstrained human-AI collaboration group achieved a mean POMP score of only 63.1 (95% CI, 56.9–69.2). Furthermore, the human-AI group's results were characterized by high variance and the frequent omission of critical methodological elements. This confirms that providing a powerful tool (a generic LLM) to a competent but non-expert user is insufficient to guarantee high-quality science. Without a methodological 'scaffold', the co-pilot mode fails to overcome human cognitive blind spots and inconsistencies.

Process as proof: execution fidelity

A rigorously designed protocol is futile if not faithfully executed. We evaluated the execution analysis agent's ability to translate the

https://doi.org/10.48130/stati-0026-0014



Fig. 4 Structural constraints attenuate intrinsic model stochasticity. Forest plots display item-level STROBE compliance scores across four distinct large language model backends integrated into the planning design agent: Gemini 2.5 Pro (top left), Claude-Sonnet-4 (top right), ChatGPT-5 (bottom left), and DeepSeek-V3 (bottom right). Center points represent the mean percent of maximum possible (POMP) score derived from $n = 50$ independent protocol generations for a single standardized reference project. Error bars denote 95% confidence intervals. The 'overall POMP score' (bottom row of each panel) quantifies the global methodological completeness, illustrating high backend-level output consistency under the same constraint architecture. The source data is provided in Supplementary File 2

abstract logic of a design protocol into concrete, executable code. Using the statistical analysis plan (see Supplementary File 1, Material S7, Chapter 7) of the standardized project as a 'gold standard', we commanded the agent to perform the analysis 50 times using Claude-Sonnet-4 and 50 times using ChatGPT-5.

Adherence was quantified using 22 binary checkpoints derived from the analysis plan (see Supplementary File 1, Material S10). Both models demonstrated near-perfect fidelity (Fig. 6). The execution adherence rates were 93.5% (SD = 10.4%; 95% CI, 90.6–96.3) for the Claude-driven agent and 92.3% (SD = 10.2%; 95% CI, 89.5–95.1) for the ChatGPT-driven agent. The difference was not statistically significant ($t = 0.58$, $df = 98$, $p = 0.57$).

Analysis at the item level revealed consistent success in complex tasks such as 'declaring complex sampling designs and variance estimation', 'constructing main/sensitivity models', and 'presenting effect sizes with 95% CIs'. This confirms that AgenMed_CSS effectively translates abstract methodological constraints into verifiable, reproducible computational actions, realizing the core concept of 'process as proof'.

Empirical validation: bridging the 'protocol-to-reality' gap with real-world data

To validate the full-process performance of AgenMed_CSS in a real-world research setting, an empirical test was conducted based on the research project titled 'A population-based cross-sectional survey of bone mineral density among adults aged ≥ 65 years in Shandong Province, China'. The research team first completed a field epidemiological investigation across the three major economic circles of Shandong Province and established a foundational database by matching field data with health archives from the basic Public Health Service system using a deterministic data linkage algorithm. Subsequently, a specific structured data subset covering demographic characteristics, lifestyle factors, and biochemical indicators was extracted from the database. After de-identification, this subset served as the final analysis dataset. This curated dataset was then input into the execution analysis agent alongside the system-generated intelligent design protocol. The agent, without human intervention, produced a statistical analysis report (see Supplementary File 1, Material S17) by strictly adhering to the complete statistical procedures outlined in the protocol (see Supplementary File 1, Material S4, Chapter 7).

https://doi.org/10.48130/stati-0026-0014

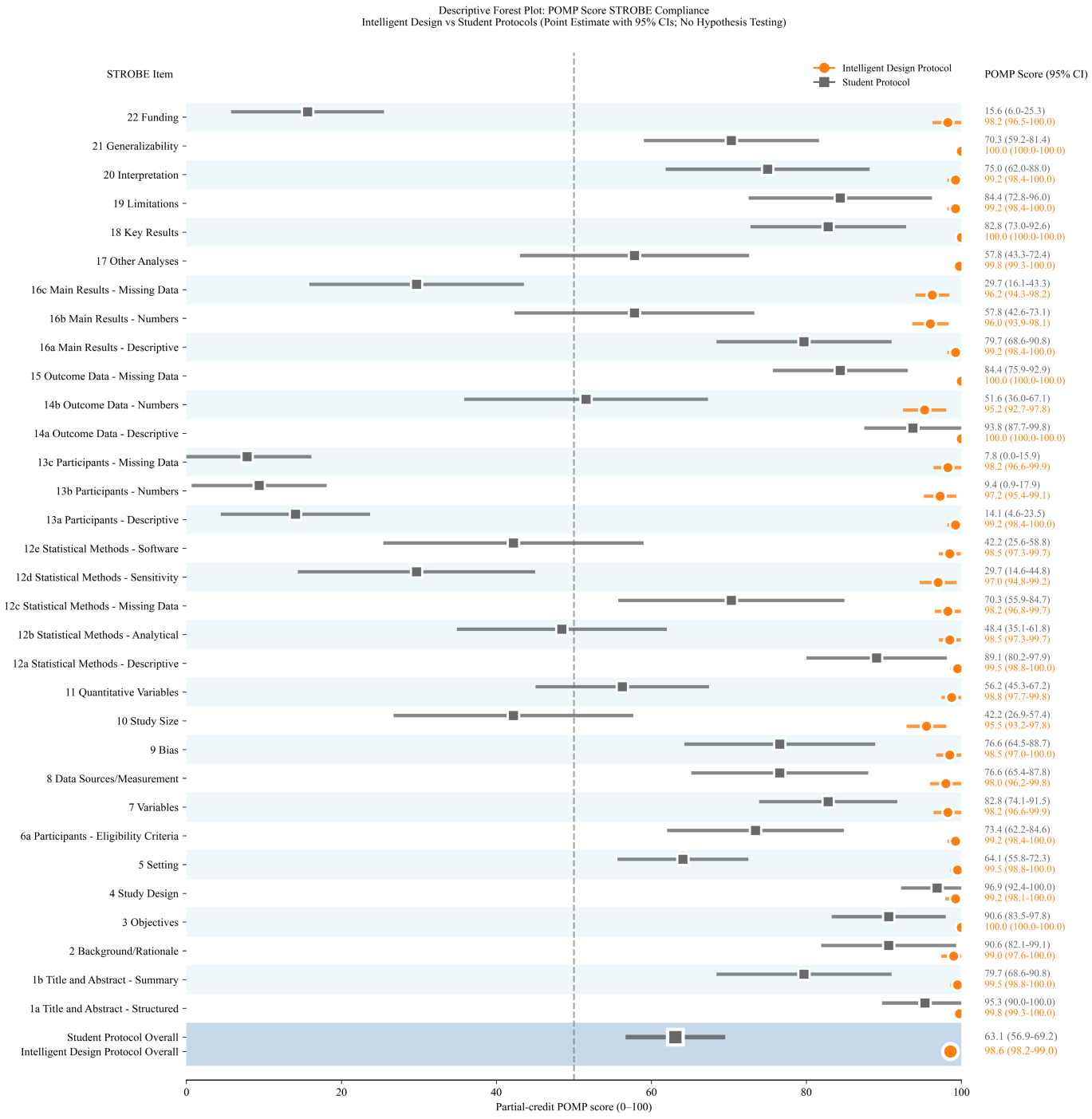


Fig. 5 Methodological scaffolding prevents the failure of unconstrained human-AI collaboration. Forest plot contrasting itemized STROBE compliance between the planning design agent (orange circles; n = 200 pooled from four models) and an unconstrained human-AI collaboration group comprising students using generic large language models (grey squares; n = 32). Data points represent mean percent of maximum possible (POMP) scores; horizontal error bars denote 95% confidence intervals. The significant gap in the aggregate 'overall' score (bottom row), combined with the substantial variance observed in the control group (wide error bars), visually quantifies the 'practice-norm gap'. This demonstrates that without the rigid constraints of the agentic framework, access to advanced AI tools alone fails to mitigate human cognitive blind spots or ensure checklist-based methodological completeness. The source data is provided in Supplementary File 2.

Discussion

The empirical results presented in this study not only validate the efficacy of a novel AI system but also pose profound questions regarding current research paradigms, the foundations of scientific trust, and the future evolution of inquiry.

From descriptive guidelines to executable contracts

Our findings support the value of front-shifting STROBE guidelines from a passive reporting checklist to a set of active design constraints. When these constraints are embedded within an agentic execution framework, methodological norms are transformed

https://doi.org/10.48130/stati-0026-0014

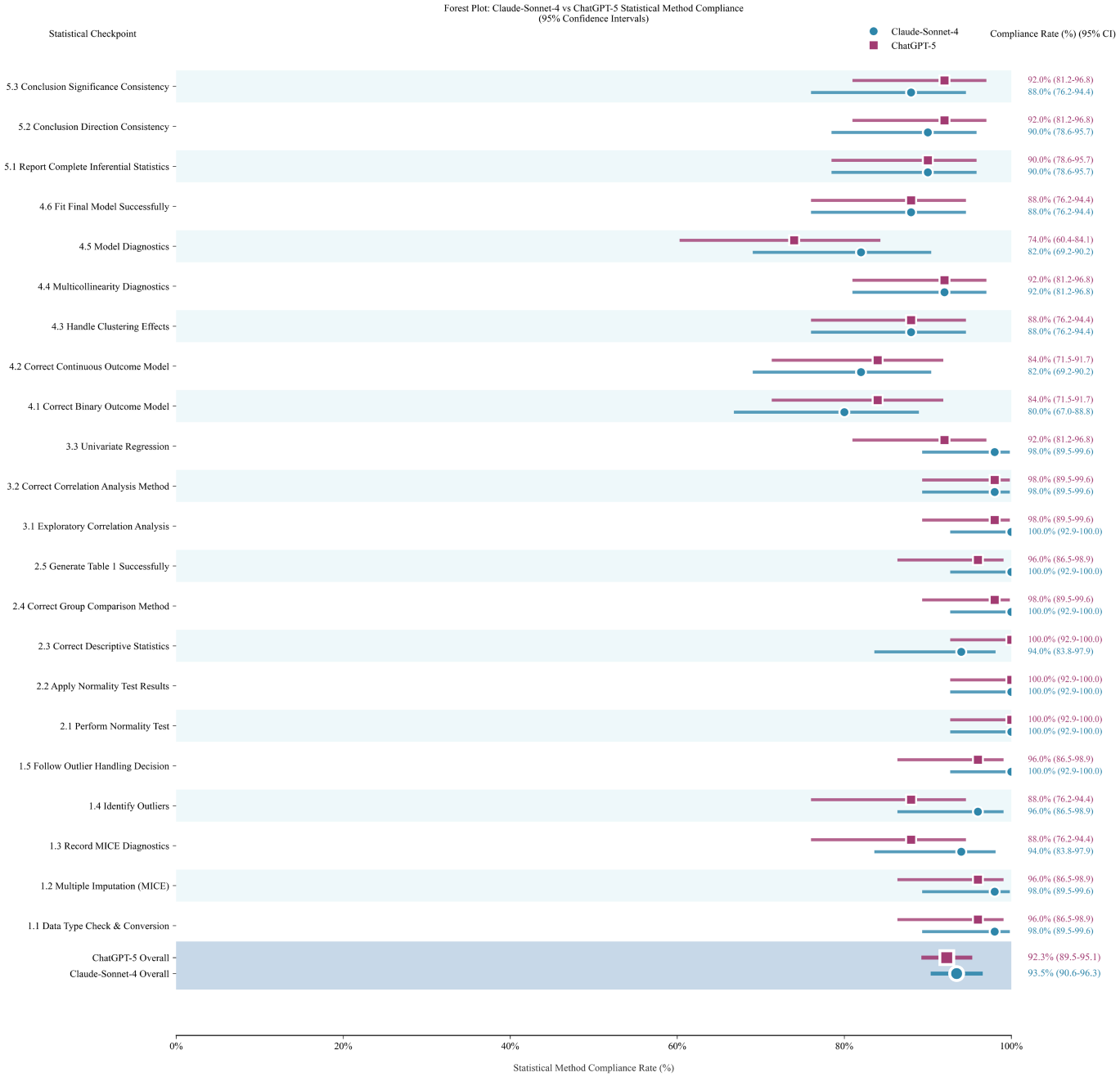


Fig. 6 Execution fidelity in translating abstract logic into concrete code. Forest plot evaluating the execution analysis agent's reliability in operationalizing statistical plans. The y-axis details 22 binary checkpoints derived from the study protocol (see Supplementary File 1, Material S7, Chapter 7), ranging from initial data type conversion to final model diagnostics. Performance is compared between Claude-Sonnet-4 (blue circles) and ChatGPT-5 (maroon squares) backends. Data points represent the mean compliance rate calculated from $n = 50$ independent execution loops for each model, with horizontal error bars denoting 95% confidence intervals. The high aggregate adherence rates (bottom shaded region) and the visual overlap between the two series demonstrate that the system consistently bridges the gap between methodological intent and computational reality, regardless of the underlying model architecture. The source data is provided in Supplementary File 2.

from suggestions into executable contracts. The system does not merely recommend best practices; it procedurally enforces them.

The comparison with 300 published articles (Fig. 3) reveals that the 'artisanal mode' of science, which relies on human experts to voluntarily adhere to guidelines, suffers from systemic, substantial, and persistent execution biases. AgenMed_CSS fundamentally resolves this by externalizing and proceduralizing tacit expert knowledge. The agent does not simulate a perfect scientist but rather faithfully executes a methodological checklist that is more comprehensive and tireless than human cognitive capacity allows, thereby raising the baseline of checklist-concordant methodological execution.

This baseline, however, refers specifically to the process-level quality of methodological execution. It does not imply that the system can judge whether a research question is clinically important, biologically plausible, or theoretically significant. That judgment remains part of the conceptual layer of scientific inquiry and requires human expertise.

https://doi.org/10.48130/stati-0026-0014

Peer review as a narrative filter: implications for process auditing

The systematic superiority of AgenMed_CSS over published articles provides a unique, large-scale empirical perspective for reconsidering the essential function of peer review. Our results indicate (Fig. 3) that successful publication in respected journals is not synonymous with methodological completeness. This observation does not seek to fault individual reviewers or editors but rather illuminates an intrinsic structural blind spot within the peer review system.

Current peer review functions primarily as a narrative filter rather than a process auditor[37–39]. It is adept at evaluating whether a manuscript's narrative is coherent, its question significant, and its conclusions novel. However, it possesses neither the capacity nor the design to audit the generative process of the research itself, verifying whether every decision from initial conception to final data analysis adhered to best practices. Reviewers evaluate static text, not the dynamic execution path replete with decision branches. By generating a complete computational lineage containing all intermediate artifacts and decision logs (see Supplementary File 1, Material S7), AgenMed_CSS addresses this gap in process auditing, offering a novel dimension for assessing research credibility.

Process as proof: a new foundation for trust

Under the artisanal paradigm, trust in scientific research relies heavily on proxy metrics: author reputation, institutional ranking, and journal impact factor. This model of trust is indirect and opaque. The emergence of AgenMed_CSS signals the possibility of a more robust foundation for trust: Process as Proof[37,40,41].

When the entire trajectory of a study—from every design decision in the protocol to every line of analysis code and its execution result—is recorded in a complete, immutable format (see Supplementary File 1, Material S7), the locus of trust shifts from faith in the person to verification of the process. The academic community no longer simply believes that an author did the right thing; they can directly review the computational records and reproduce the analysis workflow with a single click. This trust, grounded in a verifiable process, is more objective, transparent, and robust than any reputation-based trust.

Cognitive stratification: the scientist as architect of inquiry

Concerns regarding AI displacing scientists are widespread; however, our research points toward a fundamentally different future. The comparison with the unconstrained human-AI collaboration mode (Fig. 5) reveals that simply equipping humans with powerful tools does not resolve the underlying issues. True paradigmatic progress lies in a re-delineation of the boundaries of responsibility between humans and machines, a concept we term Cognitive Stratification.

In this new model, the cognitive tasks of machines and humans are clearly stratified and decoupled. The execution layer belongs to the agentic system and encompasses methodological steps that are procedural, error-intolerant, and require high fidelity and consistency; in this layer, the machine serves as the guarantor of the methodological foundation. The conceptual layer remains the domain of the human scientist and includes formulating insightful scientific questions, constructing theoretical frameworks, interpreting complex background knowledge, critically evaluating research results, and making final value judgments.

This stratification represents supervision rather than replacement. In AgenMed_CSS, it is operationalized through three researcher-facing stages: structured question framing at the outset, review and refinement of the generated protocol and SAP before execution, and post-execution inspection of code, tables, figures, and audit logs. This stratification also defines the boundary of the present evaluation. The agentic system can enforce whether a protocol specifies the required design and analysis components, but it cannot determine whether the underlying research question is worth pursuing. The formulation of meaningful hypotheses, assessment of biological plausibility, appraisal of clinical relevance, and final interpretation of results remain responsibilities of human investigators.

Limitations and generalizability

This study has limitations. First, the current system targets cross-sectional designs within observational research; its applicability to cohort studies, case-control studies, and randomized controlled trials remains to be developed and verified. Second, the STROBE_CSS-based POMP score measures checklist-based methodological completeness, not the intrinsic merit of the research question. A protocol may be complete and methodologically well-specified while still addressing a clinically unimportant or biologically implausible question. This reflects the intended boundary of Agentic T-STROBE, which targets the execution layer of research rather than the conceptual layer of scientific judgment. Finally, the cross-backend output consistency analyses did not constitute full PCS stability testing, as the underlying dataset, variable definitions, model specifications, and analytical decisions were not perturbed to test the robustness of substantive scientific conclusions.

Nevertheless, the core paradigm of this study—forward-shifted guidelines, agentic execution, and PCS evaluation—is highly modular and generalizable. The underlying architecture is domain-agnostic. By replacing the underlying STROBE rule base with CONSORT guidelines for randomized trials[42] or ARRIVE guidelines for animal research[43], this paradigm could be seamlessly migrated to other research fields. This universality suggests its potential to serve as a general-purpose infrastructure for systematic scientific inquiry across disciplines.

Conclusions

AgenMed_CSS represents a concrete implementation of a new research paradigm—one in which methodological rigor is encoded as a system property rather than left to individual discretion. It utilizes engineering principles to transform passive best practices into active, executable workflows, and to convert invisible cognitive processes into visible, auditable evidence. In doing so, it transforms methodological rigor from a variable quality dependent on individual cultivation into a systemic attribute embedded within the workflow.

The variability documented here is not intrinsic to observational research but stems from the absence of enforceable methodological constraints at the design stage. Embedding such constraints within structured, agentic workflows can make key elements of execution more transparent, auditable, and reproducible, while the scientific value of any given study will continue to rest on human judgment and domain expertise.

Ethical statements

The study was conducted in accordance with the Declaration of Helsinki, and all procedures were approved by the Ethics Committee of the School of Public Health, Shandong University, Identification number: LL20240805, Approval date: August 28, 2024. During the conduct of this study, the authors used Claude-Sonnet-4 (Anthropic, released/accessed 2025), ChatGPT-5 via ChatGPT (OpenAI, released/accessed 2025), Gemini 2.5 Pro (Google, released/accessed 2025), and DeepSeek-V3 (DeepSeek, released 2024; accessed 2025) for research protocol generation and statistical-analysis plan execution as integral components of the AgenMed_CSS system under investigation. These tools functioned as research instruments rather than writing aids. All AI-generated outputs used in the study were reviewed, verified, and interpreted by the authors. The authors take full responsibility for the integrity and accuracy of all reported findings. No AI tool is credited as an author.

Author contributions

The authors confirm their contributions to the paper as follows: Xue F and Ji X jointly conceived the original study concept and supervised the overall research project. Xue F further established the epistemological framework and constructed the methodological decision tree. Ji X provided comprehensive technical guidance for the LLM agent framework and system optimization. Bo C designed the Agentic T-STROBE architecture, developed the dual-agent software (AgenMed_CSS), performed the benchmarking experiments, and wrote the original draft. Zhao J supervised the LLM benchmarking experiments and rigorously reviewed the manuscript. Feng S supervised the on-site epidemiological survey and ensured data validity. Li R assisted in data curation. All authors reviewed the results and approved the final version of the manuscript.

Data availability

The data underlying Figs 3–6 and Table 1 are provided in the accompanying Source Data file (Suppplementary File 2). Bibliographic information for the 300 articles included in the benchmark is provided in Supplementary File 1, Material S15. The individual-level health data used for the real-world validation are not publicly available because they contain sensitive health information and are subject to ethical and data protection restrictions. Computational audit logs are provided in Supplementary File 1, Material S7. The AgenMed_CSS source code, prompt chains, and methodological decision rules are publicly available at https://github.com/boccxx/AgenMed_CSS.

Acknowledgments

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of interest

The authors declare that they have no conflict of interest.

Supplementary information accompanies this paper online at: https://doi.org/10.48130/stati-0026-0014.

Dates

Received 8 March 2026; Revised 11 May 2026; Accepted 29 May 2026; Published online 13 October 2026

References

[1] Errington TM, Mathur M, Soderberg CK, Denis A, Perfito N, et al. 2021. Investigating the replicability of preclinical cancer biology. *eLife* 10:e71601

[2] Camerer CF, Dreber A, Holzmeister F, Ho TH, Huber J, et al. 2018. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour* 2:637–644

[3] Holzmeister F, Johannesson M, Camerer CF, Chen Y, Ho TH, et al. 2025. Examining the replicability of online experiments selected by a decision market. *Nature Human Behaviour* 9:316–330

[4] Chakravorti T, Koneru S, Rajtmajer S. 2025. Reproducibility and replicability in research: What 452 professors think in Universities across the USA and India. *PLoS One* 20:e0319334

[5] Angelopoulos A, Cahoon JF, Alterovitz R. 2024. Transforming science labs into automated factories of discovery. *Science Robotics* 9:eadm6991

[6] Niederer SA, Sacks MS, Girolami M, Willcox K. 2021. Scaling digital twins from the artisanal to the industrial. *Nature Computational Science* 1:313–320

[7] Seifrid M, Pollice R, Aguilar-Granda A, Chan ZM, Hotta K, et al. 2022. Autonomous chemical experiments: challenges and perspectives on establishing a self-driving lab. *Accounts of Chemical Research* 55:2454–2466

[8] von Kortzfleisch VT, Ambrée O, Karp NA, Meyer N, Novak J, et al. 2022. Do multiple experimenters improve the reproducibility of animal studies? *PLoS Biology* 20:e3001564

[9] von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, et al. 2007. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ* 335:806–808

[10] Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, et al. 2007. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *Epidemiology* 18:805–835

[11] Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, et al. 2015. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Medicine* 12:e1001885

[12] Vandenbroucke JP. 2007. The making of STROBE. *Epidemiology* 18:797–799

[13] Pouwels KB, Widyakusuma NN, Groenwold RHH, Hak E. 2016. Quality of reporting of confounding remained suboptimal after the STROBE guideline. *Journal of Clinical Epidemiology* 69:217–224

[14] Lu MY, Chen B, Williamson DFK, Chen RJ, Zhao M, et al. 2024. A multimodal generative AI copilot for human pathology. *Nature* 634:466–473

[15] Zhang J, Liu L, Xiang P, Fang Q, Nie X, et al. 2024. AI co-pilot bronchoscope robot. *Nature Communications* 15:241

[16] Zhang Z, Ren Z, Hsu CW, Chen W, Hong ZW, et al. 2025. A multimodal robotic platform for multi-element electrocatalyst discovery. *Nature* 647:390–396

[17] Gao CA, Howard FM, Markov NS, Dyer EC, Ramesh S, et al. 2023. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *npj Digital Medicine* 6:75

[18] Piccolo SR, Denny P, Luxton-Reilly A, Payne SH, Ridge PG. 2023. Evaluating a large language model's ability to solve programming exercises from an introductory bioinformatics course. *PLoS Computational Biology* 19:e1011511

[19] Dennstädt F, Zink J, Putora PM, Hastings J, Cihoric N. 2024. Title and abstract screening for literature reviews using large language models: an exploratory study in the biomedical domain. *Systematic Reviews* 13:158

[20] Tang L, Sun Z, Idnay B, Nestor JG, Soroush A, et al. 2023. Evaluating large language models on medical evidence summarization. *npj Digital Medicine* 6:158
[21] Alber DA, Yang Z, Alyakin A, Yang E, Rai S, et al. 2025. Medical large language models are vulnerable to data-poisoning attacks. *Nature Medicine* 31:618–626
[22] Vaccaro M, Almaatouq A, Malone T. 2024. When combinations of humans and AI are useful: a systematic review and meta-analysis. *Nature Human Behaviour* 8:2293–2303
[23] Hua HU, Kaakour AH, Rachitskaya A, Srivastava S, Sharma S, et al. 2023. Evaluation and comparison of ophthalmic scientific abstracts and references by current artificial intelligence chatbots. *JAMA Ophthalmology* 141:819–824
[24] Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, et al. 2023. Large language models in medicine. *Nature Medicine* 29:1930–1940
[25] Roustan D, Bastardot F. 2025. The clinicians' guide to large language models: a general perspective with a focus on hallucinations. *Interactive Journal of Medical Research* 14:e59823
[26] Dip SA, Mallick D, Shuvo UA, Soumma SB, Rafsani F, et al. 2026. Large language model agents for biological intelligence across genomics, proteomics, spatial biology, and biomedicine. *Briefings in Bioinformatics* 27(2):bbag110
[27] Bu D, Sun J, Li K, He Z, Huang W, et al. 2026. Empowering AI data scientists using a multi-agent LLM framework with self-evolving capabilities for autonomous, tool-aware biomedical data analyses. *Nature Biomedical Engineering* Ahead of print
[28] Tzanis E, Adams LC, Akinci D'Antonoli T, Bressem KK, Cuocolo R, et al. 2026. Agentic systems in radiology: principles, opportunities, privacy risks, regulation, and sustainability concerns. *Diagnostic and Interventional Imaging* 107(1):7–16
[29] Xu Q, Soto C, Shahnawaz M, Liu X, Jiang X, et al. 2025. Multi agent large language models for biomedical hypothesis generation in drug combination discovery. *iScience* 28(12):113984
[30] Yu B, Kumbier K. 2020. Veridical data science. *Proceedings of the National Academy of Sciences USA* 117:3920–3929
[31] Rosner B. 2015. *Fundamentals of biostatistics*. 8th ed. Boston: Cengage Learning. 888 pp
[32] Wang L, Chen X, Deng XW, Wen H, You MK, et al. 2024. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *npj Digital Medicine* 7:41
[33] Chen S, Gao M, Sasse K, Hartvigsen T, Anthony B, et al. 2025. When helpfulness backfires: LLMs and the risk of false medical information due to sycophantic behavior. *npj Digital Medicine* 8:605
[34] Adams AD, Benner RS, Riggs TW, Chescheir NC. 2018. Use of the STROBE checklist to evaluate the reporting quality of observational research in obstetrics. *Obstetrics and Gynecology* 132:507–512
[35] Ziemann S, Paetzolt I, Grüßer L, Coburn M, Rossaint R, et al. 2022. Poor reporting quality of observational clinical studies comparing treatments of COVID-19: a retrospective cross-sectional study. *BMC Medical Research Methodology* 22:23
[36] Cohen P, Cohen J, Aiken LS, West SG. 1999. The problem of units and the circumstance for POMP. *Multivariate Behavioral Research* 34:315–346
[37] Chambers CD, Tzavella L. 2022. The past, present and future of registered reports. *Nature Human Behaviour* 6:29–42
[38] Proctor DM, Dada N, Serquiña A, Willett JLE. 2023. Problems with peer review shine a light on gaps in scientific training. *mBio* 14:e0318322
[39] Hamilton DG, Hong K, Fraser H, Rowhani-Farid A, Fidler F, et al. 2023. Prevalence and predictors of data and code sharing in the medical and health sciences: systematic review with meta-analysis of individual participant data. *BMJ* 382:e075767
[40] Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, et al. 2022. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nature Medicine* 28:924–933
[41] Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, et al. 2024. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 385:e078378
[42] Schulz KF, Altman DG, Moher D. 2010. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ* 340:c332
[43] Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. 2010. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLoS Biology* 8:e1000412

Copyright: © 2026 by the author(s). Published by Maximum Academic Press, Fayetteville, GA. This article is an open access article distributed under Creative Commons Attribution License (CC BY 4.0), visit https://creativecommons.org/licenses/by/4.0/.