An AI Co-Data-Scientist for Prioritizing Candidate Biomarkers from Wearable Sensor Data

Yubin Kim[†,1,3], **Salman Rahman**[1], **Samuel Schmidgall**[2], **Chunjong Park**[2], **A. Ali Heydari**[1], **Ahmed A. Metwally**[1], **Hong Yu**[1], **Xin Liu**[1], **Xuhai Xu**[1], **Yuzhe Yang**[1], **Hyeonhoon Lee**[5], **Hyewon Jeong**[3], **Kyungho Lim**[7], **MingYu Lu**[6], **Dongjae Lee**[8], **Theodora Pappa**[9], **Hanseul Cho**[10], **Maxwell A. Xu**[1], **Zhihan Zhang**[1], **Cynthia Breazeal**[3], **Tim Althoff**[1], **Petar Sirkovic**[4], **Ivor Rendulic**[4], **Annalisa Pawlosky**[1], **Nicolas Stroppa**[4], **Juraj Gottweis**[4], **Elahe Vedadi**[2], **Alan Karthikesalingam**[2], **Pushmeet Kohli**[2], **Mark Malhotra**[1], **Shwetak Patel**[1], **Samir Tulebaev**[10], **Hae Won Park**[3], **Vivek Natarajan**[†,2], **Hamid Palangi**[†,1] **and Daniel McDuff**[†,1]

[†]Corresponding Authors, [1]Google Research, [2]Google DeepMind, [3]Massachusetts Institute of Technology, [4]Google Cloud AI, [5]Seoul National University Hospital, [6]University of Washington, [7]Yonsei University College of Medicine, [8]Korea University Guro Hospital, [9]Mass General Brigham, [10]Brigham and Women's Hospital

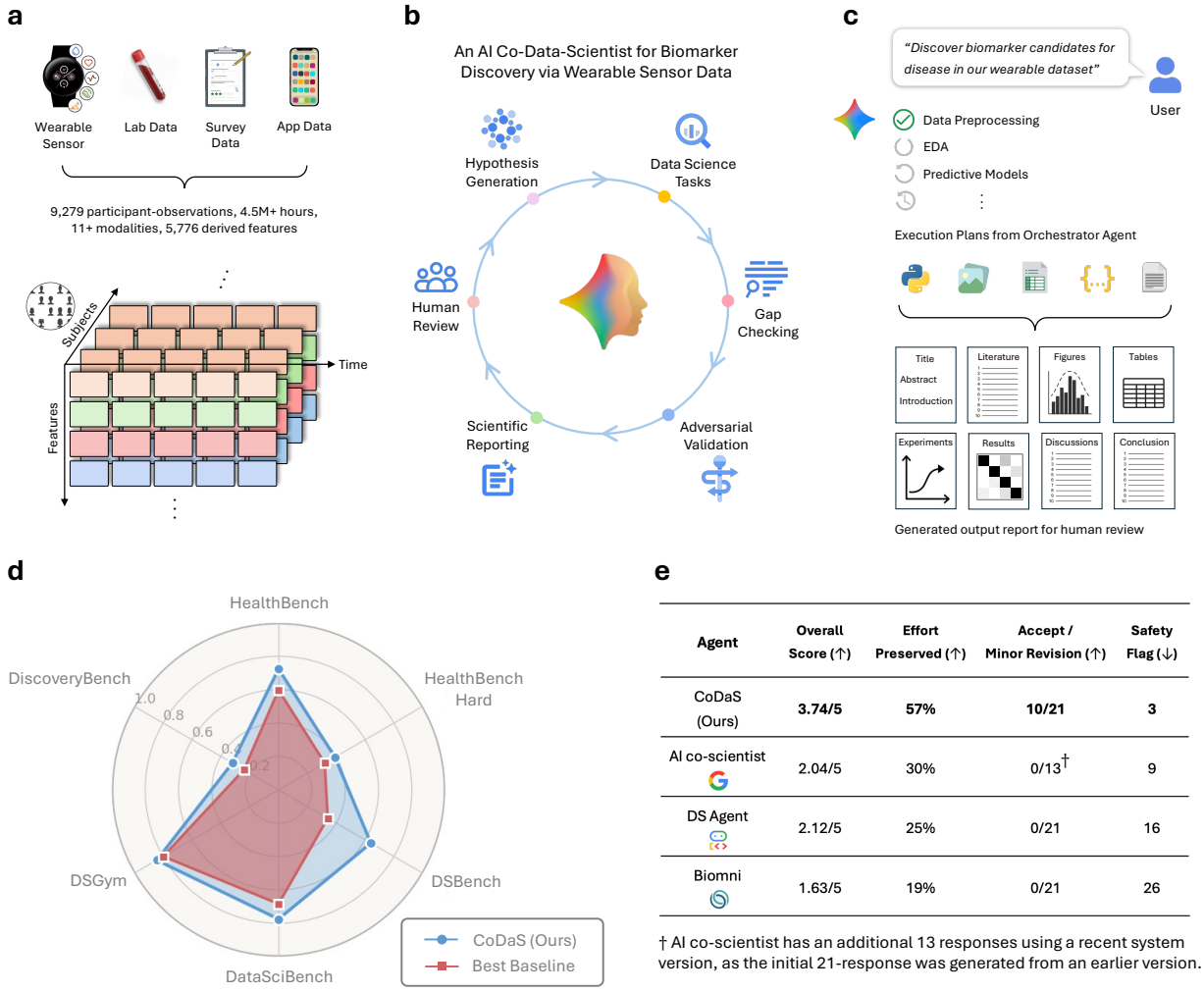
Wearable devices generate continuous physiological and behavioral data, but converting these signals into clinically reviewable biomarker hypotheses remains labor-intensive. We introduce CoDaS, an AI co-data-scientist that integrates multi-agent hypothesis generation, deterministic statistical analysis, adversarial validation and literature-grounded interpretation under human oversight. Across three wearable cohorts comprising 9,279 participant-observations, CoDaS prioritized candidate associations for mental-health and metabolic endpoints after internal checks for replication, stability, robustness and leakage. The system identified related circadian-instability signals associated with depression, including sleep-duration variability in DWB ($\rho = 0.252$, $p < 0.001$) and sleep-onset variability in GLOBEM ($\rho = 0.126$, $p < 0.001$), and derived a wearable cardiovascular-fitness index associated with insulin resistance (steps/resting heart rate; $\rho = -0.374$, $p < 0.001$). Adding these features to demographic models produced modest gains ($\Delta R^2 = 0.040$ for depression, 0.021 for insulin resistance). In a 12-clinician review totaling approximately 25 active hours, clinician validity judgments aligned with CoDaS confidence tiers ($\rho = 0.67$, $p = 0.005$), whereas added clinical value and confidence to act were rated lower. CoDaS supports traceable, hypothesis-generating prioritization of wearable candidate biomarkers.

1. Introduction

Consumer wearables, including smartwatches, continuous ECG patches, and temperature sensors provide continuous longitudinal measurements of human physiology beyond traditional clinical environments (Brasier et al., 2024; Daniore et al., 2024; Dunn et al., 2021). These devices generate high-dimensional data streams capturing heartbeat dynamics, activity patterns, sleep architecture, and thermoregulation, enabling continuous detection of early physiological deviations preceding clinical presentation (Goldsack et al., 2020; Li et al., 2026; Topol, 2019). However, translating these signals into clinically useful candidate digital measures at scale remains challenging (Alonso et al., 2024; Babrak et al., 2019; Coravos et al., 2019; Vasudevan et al., 2022). Existing approaches rely on handcrafted features within narrowly defined disease settings, limiting generalizability across populations, devices, and clinical contexts.

Digital biomarkers have shown clinical utility across several domains, from smartphone-based cognitive assessments for neurodegeneration (Dagum, 2018) and nocturnal movement signatures for

Corresponding author(s): {ybkim, natviv, hamidpalangi, dmcduff}@google.com



Agent	Overall Score (↑)	Effort Preserved (↑)	Accept / Minor Revision (↑)	Safety Flag (↓)
CoDaS (Ours)	**3.74/5**	**57%**	**10/21**	**3**
AI co-scientist	2.04/5	30%	0/13†	9
DS Agent	2.12/5	25%	0/21	16
Biomni	1.63/5	19%	0/21	26

† AI co-scientist has an additional 13 responses using a recent system version, as the initial 21-response was generated from an earlier version.

Figure 1 | **Overview. (a)** CoDaS takes continuous physiological time series data from wearable sensors, laboratory results, surveys, and third-party applications. **(b)** A closed-loop architecture orchestrates six phases mirroring the biomarker discovery lifecycle, enabling iterative refinement of candidate biomarkers for statistical robustness and physiological interpretability. **(c)** Given a natural-language research goal, CoDaS decomposes the objective into a phased execution plan, and sub-agents carry out each step through iterative analysis and deterministic code execution, producing a report draft with literature reviews, statistical summaries, mechanistic hypotheses, and assessments for human review. **(d)** CoDaS was additionally evaluated on data-science and health benchmarks, achieving competitive performance against per-benchmark strongest baselines. **(e)** In a blinded human expert evaluation (n = 15), CoDaS received the highest mean scores across quality dimensions among compared systems.

Parkinson's disease (Yang et al., 2022) to activity-derived indices for cardiometabolic risk (Chapman et al., 2022; Powell, 2024). However, these approaches remain confined to single disease contexts, lacking a scalable framework for cross-domain biomarker discovery (Fagherazzi et al., 2020; Powell, 2024).

At the same time, recent advances in large language models (LLMs) have enabled partial automation of scientific workflows, including literature synthesis (Lu et al., 2024; Zheng et al., 2025), hypothesis generation (Tong et al., 2024; Zhou et al., 2024), and experimental planning (Clusmann et al., 2023). In clinical settings, LLMs have approached expert-level performance in controlled settings such as in medical question answering (Nori et al., 2023; Singhal et al., 2025) and clinical decision support (Savage et al., 2024; Tu et al., 2024), with growing applications in safety-aware diagnostics (Kim et al., 2025d) and personalized health monitoring (Cosentino et al., 2024; Heydari et al., 2025; Kim et al., 2024). Multi-agent frameworks extend these capabilities to autonomous research workflows, with demonstrations in chemistry (Boiko et al., 2023), nanobody design (Swanson et al., 2025), single-cell analysis (Xiao et al., 2024), and end-to-end manuscript generation (Lu et al., 2024; Yamada et al., 2025). In these systems, agents such as hypothesis generators, statistical analysts, and critics collaborate through shared state to mirror the division of labor in human research teams (Gao et al., 2024; Li et al., 2023).

However, the majority of these systems operate in symbolic, textual, or algorithmic domains rather than on real world, high-dimensional physiological time series data with clinical or validated survey endpoints. This creates a practical tension between exploratory capacity and scientific rigor, particularly in high-dimensional physiological data where spurious correlations and feature leakage are prevalent. For candidate biomarkers to translate into clinical impact, it is insufficient to demonstrate statistical significance alone. They must also exhibit physiological interpretability, mechanistic plausibility, and consistency across independent cohorts (Babrak et al., 2019; Coravos et al., 2019; Goldsack et al., 2020). Single agent or monolithic systems may struggle to balance exploratory breadth with scientific rigor, risking spurious associations or uninterpretable feature composites (Jiang et al., 2024; Rotem and Zaritsky, 2024). Moreover, while autonomous AI systems excel at pattern recognition, translating discovered patterns into clinically reviewable hypotheses requires integrating domain knowledge, evaluating mechanistic coherence, and maintaining expert oversight throughout the discovery process (Daniore et al., 2024; Vasudevan et al., 2022).

To this end, we introduce **CoDaS** (AI Co-Data-Scientist), a multi-agent system for biomarker discovery from wearable time-series data. The system organizes discovery as an iterative six-phase loop spanning data profiling, hypothesis generation, statistical and machine learning analysis, adversarial validation, mechanistic reasoning, and report synthesis (Figure 1b). Specialized agents operate over shared state to enforce separation between exploration, validation, and critique, reducing the risk of uninterpretable findings (Figure 2). Unlike fully autonomous systems that operate in isolation, CoDaS is designed to preserve human oversight where the framework supports optional human feedback on intermediate findings and mechanistic interpretation at predefined checkpoints.

We evaluated CoDaS on three large-scale wearable datasets (Figure 1a) spanning mental health and metabolic disease contexts, selected to stress test the pipeline across complementary axes of analytical difficulty: high-frequency multi-modal sensing with rich behavioral signals, sparse longitudinal data with severe missingness, and cross-sectional wearable–clinical linkage requiring mechanistic integration of noninvasive and laboratory features. Specifically, we use the Digital Wellbeing (DWB) dataset (7,497 participants; hourly multimodal sensing of sleep, activity, heart rate,

and smartphone usage) (McDuff et al., 2023), the GLOBEM dataset (704 participant-wave observations from 497 unique individuals; longitudinal passive sensing from smartphones and wearables) (Xu et al., 2022), and the WEAR-ME dataset (1,078 participants from an original cohort of 1,165; physiological aggregates linked to comprehensive clinical panels) (Metwally et al., 2026). Across 9,279 total participant-observations, CoDaS generated and prioritized candidate digital biomarkers, including sleep-timing variability signatures and nocturnal digital behaviour indices associated with depression severity, and a wearable-derived cardiovascular fitness index associated with insulin resistance (details in Section 5). Principal candidates were evaluated with an internal validation battery operationalizing four complementary dimensions: replication, stability, robustness, and discriminative signal via 11 checks (including permutation testing, bootstrap confidence intervals, subgroup consistency, and methodological triangulation), consistent with emerging digital biomarker standards. Effect sizes are modest in magnitude, and all findings should be interpreted as hypothesis-generating signals requiring prospective validation.

Our primary contributions are as follows:

1. **Wearable-specific multi-agent pipeline.** We introduce CoDaS, a multi-agent system designed to support biomarker discovery from wearable sensor data. Its six-phase pipeline, spanning data profiling, hypothesis generation, parallel statistical and machine learning exploration, adversarial validation, deep literature research, and structured report generation forms an iterative loop that mirrors the human biomarker discovery lifecycle.
2. **Population-scale hypothesis generation across disease domains.** We demonstrate candidate biomarker prioritization across 9,279 participant-observations and three datasets, subjecting principal candidates to a structured internal validation battery and using an internal audit that operationalizes held-out confirmation, stability, robustness and discriminative signal via 11 checks.
3. **Human oversight and auditability.** CoDaS operates the discovery phase with human supervision through a feedback module for post-discovery review, interpretation, and optional follow-up guidance.

2. An AI Co-Data-Scientist for Candidate Biomarker Prioritization from Wearable Sensors

In this section, we introduce **CoDaS** (AI Co-Data-Scientist), a multi-agent system that implements a structured, hybrid discovery pipeline. CoDaS structures candidate-biomarker prioritization by systematically exploring, generating, and prioritizing candidate hypotheses at a scale that would be challenging to be achieved manually.

2.1. System Architecture and Agent Specialization

CoDaS operates as a set of AI agents coordinated through a shared workflow, using Gemini-3.1 Pro Preview for research-intensive reasoning and code generation and Gemini-3 Flash Preview for repeated lower-latency tasks. To support the full discovery process, CoDaS replaces monolithic reasoning with

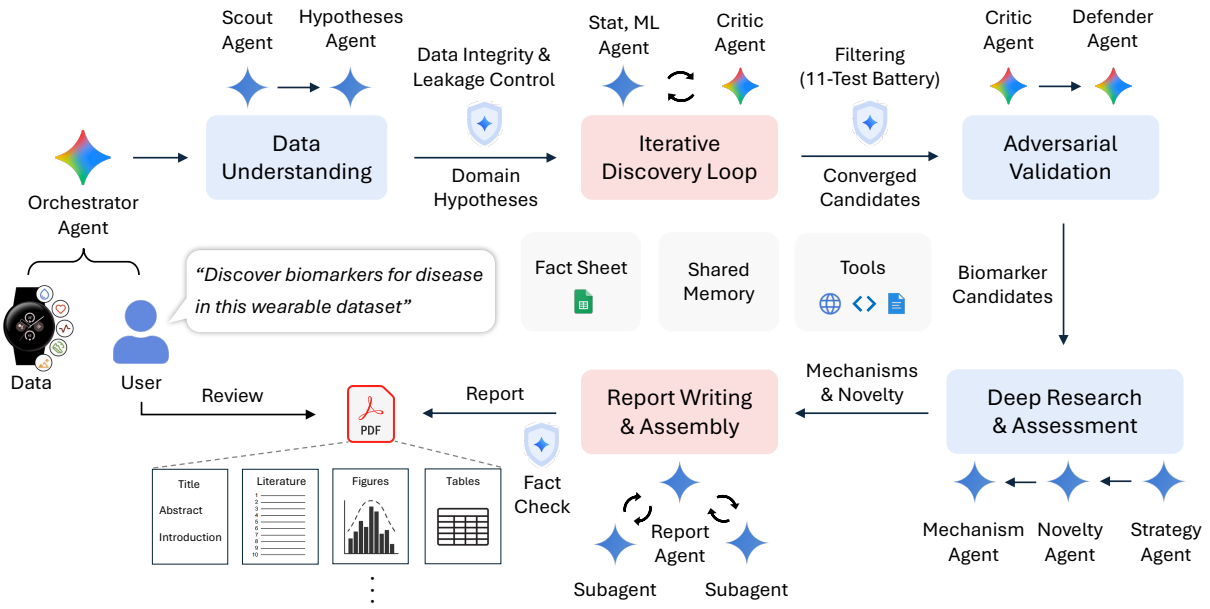


Figure 2 | **CoDaS Architecture.** Given a research goal and wearable sensor data, an Orchestrator Agent coordinates sequential stages through specialized sub-agents. **(1)** *Data Understanding*: Scout and Hypotheses Agents profile the dataset and generate domain-grounded hypotheses. **(2)** *Iterative Discovery Loop*: Statistical/ML and Critic Agents iteratively refine candidate biomarkers until convergence. **(3)** *Adversarial Validation*: Critic and Defender Agents debate each candidate to filter for statistical robustness. **(4)** *Deep Research & Assessment*: Mechanism, Novelty, and Strategy Agents evaluate biological plausibility, literature novelty, and translational potential of statistically prioritized candidates. **(5)** *Report Writing & Assembly*: a Report Agent with dedicated sub-agents compiles a draft manuscript for expert review. All agents share a memory, fact sheet and tool sets. CoDaS enforces safety mechanisms to support statistical validity and reduce the risk of spurious reporting. A *leakage-prevention* module separates feature construction from target signals, while candidate associations are filtered through statistical tests with FDR correction and assigned explicit reporting labels, including screened, conditional, exploratory, rejected or unstable. An *adversarial validation step* further audits each candidate for target leakage, overfitting, construct overlap, confounding sensitivity and physiological implausibility. A detailed architectural diagram is provided in Figure 8.

specialized personas, dedicated tool sets, and distinct mandates within a six-phase workflow. This specialization separates empirical analysis, theoretical grounding, and strategic oversight.

Researcher Ensemble. The Researcher ensemble compiles a structured summary of the relevant clinical literature to guide subsequent biomarker discovery.

- **Inputs:** A high level research query (e.g., "discover predictive signatures for depression severity") and access to scientific literature databases.
- **Process:** The ensemble utilizes a set of sub-agents. A *Literature searcher* executes targeted queries to retrieve a corpus of relevant scientific abstracts and full text articles. Simultaneously, a *BibTeX*

Validator and a *Paper Verifier* deterministically crosscheck retrieved claims to reduce hallucination. Finally, specialized *Novelty* and *Mechanism* agents summarize candidate mechanistic links reported in the literature.

- **Outputs:** A structured biological prior containing: (1) a list of established biomarkers and their reported physiological pathways; (2) assessments determining the true novelty of generated candidates; and (3) mechanistic rationales linking the data-driven findings to plausible biological pathways.

Data Science Engine. The Data Science engine represents the analytical core of CoDaS, responsible for all direct interactions with the wearable sensor data. It abandons purely generated code in favor of a hybrid deterministic and generative approach.

- **Inputs:** Raw, high-dimensional time series datasets intersecting with the structured knowledge base provided by the *Researcher* ensemble.
- **Process:** The engine comprises paired deterministic code runners and language model interpreters. A *Scout* agent first maps the dataset schema, defining the clinical target variable. A *DataLoader* and *Exploratory Data Analysis (EDA)* runner then profile the data structure. Following this, *Statistical runners* implement robust univariate testing (e.g., correlation, effect size), while *Machine Learning runners* execute multivariate predictive modeling with cross validation (e.g., Ridge regression, ensemble trees) in parallel. Paired interpreters parse the respective outputs to form coherent analytical narratives.
- **Outputs:** A comprehensive suite of empirical results including: (1) ranked lists of novel biomarker candidates (ordered by composite effect size, validation pass rate, and clinical plausibility); (2) robust statistical metrics including significance levels adjusted for multiple comparisons; (3) out-of-sample predictive performance metrics; and (4) generated correlation heatmaps and feature importance distributions.

Orchestrator and Critical Evaluators. The Orchestrator coordinates the CoDaS pipeline, providing strategic direction, managing state transitions across the six phases, and ensuring rigorous internal review.

- **Inputs:** The overarching research objective, real time outputs from specialist agents, and periodic interactive feedback from human domain experts.
- **Process:** The Orchestrator manages the pipeline trajectory. It implements a *GapChecker* to identify unresolved questions following iterative empirical analysis, deciding whether to pursue deeper feature engineering or move to validation. Crucially, it initiates an adversarial debate phase involving *Critic* and *Defender* agents, which actively stress test the proposed biomarkers for confounding variables, statistical leakage, and physiological implausibility before finalization. This hierarchical oversight structure aligns with recent paradigms emphasizing tiered agentic architectures to ensure AI safety and rigorous validation in clinical contexts (Kim et al., 2025c). A worked Critic and

Defender exchange, covering the rejection of a leakage feature and a redundant proxy and the handling of a confounded behavioral candidate, is provided in Section 5.4.

- **Outputs:** The coordinated sequence of agent invocations, automated manuscript generation via paired *Writer* and *Reviewer* agents, and structured interactive prompts requesting domain expert feedback at critical junctures.

2.2. The Hybrid Discovery Pipeline

CoDaS implements a structured six-phase pipeline that narrows the search space from thousands of raw sensor permutations to a curated set of statistically prioritized, mechanistically grounded biomarker candidates (see Figure 2 for an overview and Figure 8 for the full details).

Phase A: Automated Data Profiling and Literature Grounding The objective of the initial phase is to build both a conceptual and empirical map of the clinical task.

- **Empirical Contextualization:** Deterministic loaders and the EDA runner survey the raw data. The Scout agent synthesizes these statistical profiles to establish an analytical baseline, understanding data sparsity, longitudinal variance, and demographic distributions.
- **Biological Anchoring:** The Orchestrator tasks the Researcher ensemble to construct the theoretical foundation. By identifying established clinical predictors, CoDaS seeds the search space with literature-derived priors, encouraging subsequent exploration to remain anchored to biological plausibility rather than spurious correlations.

Phases B & C: Parallel Agentic Search and Adversarial Validation The core discovery engine actively synthesizes raw features into composite physiological parameters over iterative loops.

- **Dual Track Parallel Exploration:** To balance interpretability with maximal predictive power, the Orchestrator runs parallel statistical and machine learning iterations. Generative interpreters propose physiologically rational transformations (e.g., standard deviations of resting heart rates, or ratios between activity profiles), which deterministic runners immediately evaluate. A GapChecker module monitors marginal gains in model performance, candidate novelty, and validation yield, and uses these signals to decide whether further refinement is warranted.
- **Adversarial Stress Testing:** Candidate biomarkers prioritized for Phase C undergo an additional review stage designed to identify potential sources of spurious association. A Critic agent evaluates alternative explanations, including statistical artifacts, data leakage, confounding and inconsistencies with the literature, whereas a Defender agent assesses whether the available empirical evidence supports retention of the candidate. Biomarkers that cannot be supported following this review are removed from further consideration. A representative example illustrating how this process identifies leakage, redundant proxies and confounding-prone candidates is provided in Section 5.4.

Phases D, E, & F: Mechanistic Reasoning and Automated Reporting The concluding phases reconstruct the surviving empirical signals into a structured draft manuscript.

- **Novelty and Mechanism Extraction (Phase D):** The Researcher ensemble executes deep secondary literature sweeps focused exclusively on the surviving biomarker candidates. It formally assesses literature novelty and formulates hypothesis-generating physiological rationales linking the digital signature to plausible pathways described in the literature.
- **Drafting and Interactive Feedback (Phases E & F):** Writer and Reviewer agents format the findings, encompassing statistical summaries, machine learning benchmarks, and mechanistic rationales, into structured draft reports suitable for expert review. The system then enters the final collaborative stage, presenting the comprehensive draft to the human expert.

2.3. Coordination and Interactive Feedback

The Orchestrator agent facilitates communication through a shared state, allowing downstream agents to access and build upon preceding qualitative evaluations. Most importantly, the framework embraces collaborative discovery with human practitioners. The Orchestrator is programmed to identify complex scenarios demanding biological intuition and immediately solicit expert input. These interactive triggers include:

- **Plausibility Gaps:** A generated biomarker achieves higher-than-expected predictive performance but lacks a clear analogue in the retrieved literature. The Orchestrator pauses progression, querying the medical professional for interpretative guidance.
- **Conflicting Evidence Paradigm:** Substantial data-driven findings fundamentally contradict established paradigms identified by the Researcher ensemble.
- **Search Stagnation:** The generative exploration fails to surpass predefined performance baselines over sequential iterations, prompting the expert to explicitly suggest customized feature transformations.

This mechanism is intended to preserve expert interpretive authority while allowing the system to surface candidate findings more efficiently.

2.4. Multi-axis candidate assessment

Optimizing solely for machine learning accuracy frequently yields noninterpretable blackbox algorithms unsuited for medical application. To mitigate this, the Orchestrator evaluates candidate features across a multidimensional framework spanning both quantitative validity and qualitative clinical assessment.

1. **Statistical Validity:** Evaluates signal strength across robust metrics including correlation effect sizes, multiple hypothesis adjusted significance levels, and out-of-sample predictive performance using extensive cross validation procedures.

2. **Clinical Plausibility:** Ensures alignment with biological plausibility. Clinical plausibility is assessed by asking the mechanism agents to link each candidate feature to a plausible physiological pathway supported by retrieved primary literature.
3. **Originality:** Quantifies the conceptual distance between the proposed biomarker and established indicators curated during the initial literature seeding phase, surfacing potentially novel candidate metrics.
4. **Generalizability:** Assesses robustness by running available subgroup analyses, examining biomarker performance consistency across differing demographic cohorts, sensor platforms, or distinct disease severity stratifications.
5. **Interpretability:** Penalizes extreme mathematical complexity. The system preferentially weights intuitive, physiologically meaningful composites (e.g., activity recovery gradients or nocturnal variance indices) over highly abstract, untethered neural embeddings.
6. **Clinical Utility:** Separates candidate discovery from clinical actionability. Candidates are later assessed by clinicians for practical measurability, added value over existing biomarkers, likelihood of influencing advice and confidence to act. Candidates with low actionability are retained as hypotheses for prospective evaluation.

These axes deliberately separate statistical validity from clinical value, so that a candidate can be statistically significant, biologically plausible and novel yet still rank low when its incremental clinical utility is weak. Measured confounding is partially assessed within the validity axis through a causal-robustness check, which residualizes each candidate against demographic covariates and previously validated biomarkers and requires the association to survive partial-correlation control (Section 2.6). This guards against candidates whose association is carried by an obvious common cause, for example physical activity that influences both step count and mood. Because the data are observational, this control cannot exclude unmeasured confounding, so a high multi-axis score marks a candidate as worth confounder-controlled prospective study rather than as an established causal biomarker.

2.5. Data Integrity and Leakage Guardrail

Information leakage and confounding are critical threats to automated biomarker discovery, where the combinatorial feature space can inadvertently include transformations of the outcome variable. CoDaS enforces the following procedural safeguards throughout the pipeline, designed to ensure that reviewers can audit the boundary between discovery inputs and evaluation targets.

1. **Raw variable exclusion.** The target variable (e.g., PHQ-8, HOMA-IR) and its direct clinical proxies (e.g., fasting glucose and fasting insulin for HOMA-IR, BDI-II for PHQ-4) are excluded from the candidate feature pool at data loading time, before any feature engineering begins.
2. **Transformation prohibition.** Monotonic transformations of excluded variables (squared, log, rank-transformed) are detected by the Critic agent's construct overlap analysis, which computes the Spearman correlation of every candidate against all excluded variables. Features exceeding $|\rho| > 0.85$ with any excluded variable are flagged and removed from the candidate pool. We adopt

0.85 as the operational threshold for "very strong" monotonic association in biomedical research; in practice, tautological transformations (e.g., `glucose_sq`) typically exhibit $|\rho| > 0.95$ with the target, and no candidate in any cohort fell in the 0.80–0.90 range with an excluded variable, so moderate variations in this threshold would not alter the reported results.

3. **Discovery-evaluation separation.** All predictive performance is reported exclusively via 5-fold cross-validation with stratified participant-level splits. No participant appears in both training and validation folds within any round. Hyperparameter selection occurs within each training fold only.
4. **Label isolation.** Target labels are never exposed to the hypothesis generation, feature engineering, or literature grounding agents. Labels are loaded only by the deterministic statistical and ML runners, which execute in isolated subprocesses. The LLM-based agents observe only summary statistics (e.g., correlation direction, p-value ranges) returned by these runners, not the underlying label data.
5. **Human review boundary.** Human feedback during the interactive phase is restricted to mechanistic interpretation, plausibility assessment, and high-level analytical guidance (e.g., suggesting domain-informed feature transformations when the pipeline stagnates). Crucially, domain experts do not have access to raw data, model performance metrics, or fold-level predictions during feedback sessions, preventing target-informed feature selection.
6. **Construct overlap gating.** Every candidate surviving statistical screening undergoes a construct overlap test measuring its independence from existing validated clinical instruments and from other candidates. Features with high intra-cluster correlation ($|\rho| > 0.85$) are reported but not double-counted in the final validated set.

2.6. Statistical Internal Validation Battery and Reporting Integrity

Autonomous AI systems operating in biomedical research face inherent risks of hallucination, spurious discovery, and unverified reporting (Cornelio et al., 2025; Luo et al., 2025; Tang et al., 2025; Zhu et al., 2025). CoDaS addresses these through two complementary mechanisms: a multi-stage statistical internal validation battery that every candidate must pass, and a deterministic reporting framework that reduces LLM-generated prose from diverging from empirical results.

Internal Validation battery: four dimensions, eleven checks. Every candidate biomarker that passes the statistical filtering stage must survive an internal validation battery organized around four complementary dimensions: replication, stability, robustness, and discriminative power, operationalized via 11 checks executed in a deterministic subprocess isolated from the language model agents. None of these checks were preregistered; they are components of the pipeline design, not a prespecified analysis plan. The checks are not independent tests; several share underlying data (e.g., replication and bootstrap both operate on the same sample; subgroup consistency and robustness both use demographic variables), and should be interpreted as a structured post-hoc audit rather than 11 orthogonal significance tests. All results from this battery are hypothesis-generating and do not replace prospective external validation.

1. **Held-out confirmation.** Spearman correlation on a held-out confirmation set (distinct participant-level split, $N \geq 20$), verifying that the effect shows directionally consistent association in a held-out split. For cohorts with repeated measures (e.g., GLOBEM), the confirmation set retains only one randomly selected observation per participant so that the test statistic is computed on fully independent rows.
2. **Permutation test.** Empirical null distribution from 1,000 label-permuted resamples; guards against inflated significance due to distributional properties of the feature.
3. **Bootstrap stability.** 1,000 bootstrap resamples the validation set with replacement providing 95% confidence intervals for the association estimate; rejects candidates whose CI straddles zero, indicating directional instability.
4. **Leave-one-out influence.** Rejects any candidate whose association sign flips when any single participant is excluded, indicating sensitivity to outliers.
5. **Subgroup consistency.** Requires the association to hold within each half of the cohort (class split for classification; median split for regression), protecting against Simpson's paradox.
6. **Method triangulation.** Recomputes the association using Pearson and Kendall's τ; the candidate must remain significant across all applicable methods, guarding against method-specific artifacts.
7. **Construct validity hard gate.** Rejects candidates with implausibly strong correlations ($|\rho| > 0.85$ for $N > 30$; adaptive thresholds for smaller samples), which typically indicate undetected tautological transformations.
8. **Confounding-sensitivity check.** Residualizes the candidate against demographic confounders and previously validated biomarkers; the association must survive partial-correlation control.
9. **Construct independence hard gate.** Detects derived features whose components correlate strongly with the target, classifying each candidate as independent, proxy, or compositional. Proxy and compositional candidates are rejected or flagged for disclosure.
10. **CI consistency hard gate.** Verifies that the point estimate and bootstrap CI midpoint agree in sign; directional inconsistency indicates numerical instability.
11. **Discriminative signal check.** For classification analyses, we report AUC as a descriptive measure of above-chance discrimination. We use AUC ≥ 0.55 only as a permissive internal screen to identify whether a candidate carries non-random information; it is not a threshold for clinical utility, diagnostic use or deployment readiness.

If all three of Tests 1 through 3 fail simultaneously, the candidate is immediately rejected. Candidates passing at least 70% of applicable checks, with all core tests (replication, permutation, bootstrap, and CI consistency) passed, are labeled SCREENED; candidates passing 40–70% of checks, or downgraded from screened status due to marginal effect sizes, are labeled CONDITIONAL; the remainder are REJECTED. Three hard gates (construct validity, construct independence, CI consistency) trigger automatic rejection regardless of overall pass rate.

All verdicts, evidence summaries, and per-test results are persisted in pipeline state and injected into the manuscript so that reported validation statistics derive from deterministic computation rather than LLM-generated prose.

Fact Sheet. A known vulnerability of LLM-based scientific writing is hallucination of numerical values, including sample sizes, effect sizes, and validation counts (Young, 2025). CoDaS mitigates this through a *Fact Sheet*: a flat key-value dictionary of every reportable number computed deterministically from pipeline state before any section writer is invoked. The Fact Sheet captures sample sizes, demographic distributions, model performance (R^2, AUC), validation counts, feature counts, discovery round tallies, and construct exclusion summaries. All section-writing agents receive the Fact Sheet as a structured context attachment and are instructed to copy values verbatim rather than infer them from narrative context.

Numeric verification and consistency enforcement. After each section is drafted, a dedicated numeric verification pass applies pattern-based correction to detect and fix common hallucination targets: sample size claims, prioritized candidate counts, validation test counts, method counts, and feature counts. Corrections are applied only when the LLM-written value falls within a tolerance of the known ground truth, limiting false positives. All corrections are logged to a per-run audit file for transparency. A final consistency check cross-references every section against the Fact Sheet before LaTeX compilation.

Quality gates for output suppression. CoDaS applies five deterministic quality gates before report assembly: (i) a multicollinearity gate suppresses OLS tables when the variance inflation factor exceeds 50; (ii) a performance gate suppresses ML result tables when the best cross-validated AUC falls below 0.55 or R^2 falls below 0; (iii) an overfitting gate suppresses results when the train-to-CV ratio exceeds 5; (iv) an ablation gate suppresses feature importance tables when all models perform at chance; and (v) a forest plot deduplication gate limits any single feature family to two representatives, preventing visual dominance by correlated variants.

Together, these mechanisms align with emerging best practices for long-running autonomous scientific agents (Lu et al., 2026; Wu et al., 2026; Young, 2025) and address the integrity risks identified for AI scientist systems operating in high-stakes biomedical domains (Tang et al., 2025; Zhu et al., 2025).

3. Data

We assembled three complementary clinical datasets spanning mental health and metabolic disease domains, collectively representing 9,279 participant-observations (9,072 unique individuals) with continuous physiological monitoring. Each dataset provides distinct advantages for wearable-based biomarker discovery: comprehensive high-frequency longitudinal sensing for depression severity, multiwave passive behavioral phenotyping, and precisely synchronized biometric monitoring linked to robust clinical blood panels. Together, these datasets enable comprehensive evaluation of our autonomous data science framework across diverse disease mechanisms, demographic distributions, and temporal scales. Cohort demographics, modality coverage, feature counts, missingness, and endpoint definitions are summarized in Table 1.

Table 1 | **Cohort characteristics.** Participant demographics, data modalities, and endpoint definitions across the three evaluation cohorts. Abbreviations: DWB, Digital Wellbeing; GLOBEM, Generalization of Longitudinal Behavior Modeling; WEAR-ME, Wearables for Metabolic Health; SD, standard deviation; BMI, body mass index; RHR, resting heart rate; HRV, heart rate variability; GPS, Global Positioning System; PHQ, Patient Health Questionnaire; GAD-7, Generalized Anxiety Disorder-7; PSS, Perceived Stress Scale; PROMIS, Patient-Reported Outcomes Measurement Information System; BDI-II, Beck Depression Inventory-II; BFI-10, Big Five Inventory-10; ERQ, Emotion Regulation Questionnaire; PANAS, Positive and Negative Affect Schedule; CMP, comprehensive metabolic panel; CBC, complete blood count; CRP, C-reactive protein; GGT, gamma-glutamyl transferase; HbA1c, hemoglobin A1c; HOMA-IR, homeostasis model assessment of insulin resistance; RAPIDS, Reproducible Analysis Pipeline for Data Streams.

Variable	**DWB (N = 7,497)**	**GLOBEM (N = 704[c])**	**WEAR-ME (N = 1,078[d])**
Age (mean ± SD)	43.9 ± 12.7	19.2 ± 1.4	46.9 ± 12.5
Sex (%)			
Female	70.0	58.8	54.4
Male	26.5	40.2	43.6
Other/Not reported	3.5	1.0	2.0
Race/Ethnicity (%)[a]			
White/Caucasian	84.9	36.4	—[e]
Asian	2.8	57.4	—
Black/African American	4.0	3.3	—
Hispanic	7.6	7.5	—
Other/Multiracial	5.3	11.2	—
Not reported	0.6	—	—
BMI (mean ± SD)	—	—	29.2 ± 6.7
Device type	Fitbit + smartphone	Fitbit + smartphone	Fitbit / Pixel Watch
Available modalities			
Wearable signals	Steps, RHR, sleep architecture	Fitbit steps, Fitbit sleep	RHR, HRV, steps, sleep duration, active zone minutes
Smartphone logs	Screen time, unlocks, app usage, activity recognition, GPS mobility	Bluetooth proximity, GPS location, phone calls, screen events, WiFi connectivity	—
Clinical screeners	PHQ-8, GAD-7, PSS, PROMIS Sleep	PHQ-4, BDI-II	—
Fasting lab panels	—	—	Lipid panel, CMP, CBC with differential, CRP, insulin, GGT, testosterone, HbA1c
Survey instruments	Big Five personality, sleep quality	BFI-10, BDI-II, ERQ, PHQ-4, PSS-4, PANAS	Demographics, health history
Feature count	197	5,508	71
Monitoring duration	26.5 ± 4.7 days	77.5 ± 8.9 days	Cross-sectional
Missingness (%)	3.0	54.6[b]	0.1
Target endpoint	PHQ-8 (0–24)	PHQ-4 (0–12)	HOMA-IR (continuous); HbA1c-based diabetes status

[a]Hispanic/Latino ethnicity was collected separately in DWB and GLOBEM; race/ethnicity percentages may therefore sum to >100%. [b]GLOBEM missingness reflects sparse passive sensing features; features with >70% missingness were removed and remaining values were median-imputed within sensing wave. PHQ-4 coverage was 65.0%. [c]GLOBEM includes 704 participant-wave observations from 497 individuals across four annual cohorts (Xu et al., 2022). [d]WEAR-ME includes 1,078 analytic participants after excluding 87 with incomplete wearable feature coverage from the original cohort of 1,165 (Metwally et al., 2026). [e]WEAR-ME race/ethnicity variables were not available in the participant-level analytic dataset; source-study distributions are reported in the original cohort publication. — indicates not collected or not available.

3.1. Digital Wellbeing (DWB) Cohort

To investigate continuous physiological indicators of depression and anxiety, we analyzed data from the Digital Wellbeing study, a prospective observational cohort of 7,500 individuals from the United States (McDuff et al., 2024). Three participants were excluded owing to incomplete baseline assessments, yielding a final analytic sample of 7,497. The study was designed to investigate patterns and relationships between digital device use, continuous physiological signals, and self reported measures of mental health over a four week tracking period. Review and approval for participant enrollment were granted by the Institutional Review Board of the University of Oregon (MOD00000379), and all participants provided informed consent prior to data collection.

The recruitment protocol intentionally targeted demographic diversity. The enrollment stratified participants comprehensively across race, ethnicity (Caucasian, African American, Asian, Latino, Indigenous Populations), sex, and age distributions (18 to 40, and over 40). Data completeness was incentivized via a raffle structure requiring a minimum threshold of seven days of daily status assessments alongside baseline and post study questionnaires.

Participants completed a comprehensive battery of validated psychiatric questionnaires. Baseline assessments included the Patient Health Questionnaire 8 (Kroenke et al., 2009) for depression screening, the Generalized Anxiety Disorder Scale (Löwe et al., 2008), the Patient Reported Outcomes Measurement Information System Sleep Disturbance subscale (Cella et al., 2007; Yu et al., 2012), the Smartphone Addiction Scale (Kwon et al., 2013), and the Perceived Stress Scale (Cohen et al., 1983).

The final analyzed cohort for our pipeline comprised 7,497 participants contributing 4.55 million hourly records of continuous physiological sensing. The recorded modalities included sleep architecture, step counts, resting heart rate, and smartphone application usage statistics. This temporal resolution supports the extraction of circadian variability indices and nocturnal behavioral markers relevant to affective disorders.

3.2. Generalization of Longitudinal Behavior Modeling (GLOBEM) Cohort

To evaluate early behavioral modeling for depression detection across independent temporal cohorts, we utilized the Generalization of Longitudinal Behavior Modeling dataset (Xu et al., 2022). This dataset aggregates multiwave passive sensing data collected from undergraduate students at the University of Washington via smartphone interactions and wrist worn commercial fitness trackers (Fitbit Flex2 and Inspire2), spanning four consecutive annual cohorts (2018–2021). The study was approved by the University of Washington Institutional Review Board.

The analyzed cohort comprised 704 participant-wave observations from 497 unique individuals (mean age 19.2 ± 1.4 years; 58.8% female), with some participants contributing data across multiple annual waves. Rather than isolated cross-sectional measurements, this dataset documents transitions in mental health status over extended academic periods. Continuous modalities include Bluetooth proximity logs, location mobility metrics, sleep duration variance, and daily aggregated activity counts.

Ground truth psychiatric status was established through periodic clinical surveys administered throughout the sensing waves, including weekly PHQ-4 ecological momentary assessments and end-of-term BDI-II administrations. By testing our autonomous discovery pipeline on this cohort, we evaluated reproducible behavioral signatures, such as diminished geospatial mobility and highly variable sleep architecture, that persist across multiple annual cohorts and sensor hardware transitions.

3.3. Wearables for Metabolic Health (WEAR-ME) Cohort

To assess the capacity of consumer wearable devices to capture subclinical metabolic dysregulation, we analyzed data from the Wearables for Metabolic Health study, a prospective observational cohort of 1,165 participants recruited remotely across the United States via the Google Health Studies application (Metwally et al., 2026). The study was approved by Advarra IRB (Protocol Pro00074093). The primary objective evaluates the feasibility of translating continuous data from standard fitness trackers and smartwatches into composite algorithms correlating strongly with rigorous metabolic assays.

After excluding 87 participants with incomplete wearable feature coverage during preprocessing, the final analytic sample comprised 1,078 participants (mean age 46.9 ± 12.5 years; 54.4% female; mean body mass index 29.2 ± 6.7). Adherence was defined as a minimum of 14 days of continuous wearable data paired with a comprehensive clinical blood draw and completed demographic surveys. Participants wore Fitbit or Google Pixel Watch devices capturing high resolution heart rate, heart rate variability, step counts, and sleep stages. To reduce circadian variability in laboratory measurements, participants underwent fasting laboratory tests (minimum of 8 fasting hours) early in the morning (7–10 am) at Quest Diagnostics centers to stabilize circadian blood markers.

The resulting clinical ground truth includes a panel of metabolic readouts: homeostasis model assessment of insulin resistance, fasting insulin, HbA1c, comprehensive metabolic panels (fasting glucose, creatinine, electrolytes), lipid profiles (total cholesterol, high density lipoproteins, triglycerides), high sensitivity C reactive protein, and advanced hematological indices. Demographic endpoints encompassing waist circumference, blood pressure, undiagnosed metabolic syndrome classifications, and body mass index distributions were verified via structured clinical reporting.

By anchoring high-frequency, continuously sampled physiological streams such as resting heart rate and derived cardiovascular fitness indices against these fasting laboratory measurements, this dataset provides a useful testbed for algorithmic biomarker discovery. This linkage permits our framework to explicitly target prediabetic transitions and insulin resistance using noninvasive, passively collected wearable-derived signatures evaluated against fasting laboratory endpoints.

3.4. Endpoint Specification and Statistical Power

Endpoint prespecification. PHQ-8 total score was pre-specified as the primary endpoint for the DWB cohort before CoDaS pipeline execution. HOMA-IR (continuous) was pre-specified for WEAR-ME. For GLOBEM, PHQ-4 was selected by the Scout agent from available clinical instruments as the endpoint with the greatest target coverage (65.0%; the only alternative, BDI-II, had substantially lower

coverage as it was administered only at end-of-term rather than weekly); this selection was therefore data-driven rather than pre-specified. This constitutes a form of data-driven endpoint selection and introduces potential optimism bias; results from this cohort should accordingly be interpreted with additional caution beyond that noted for the other two cohorts.

Sample size rationale. The DWB cohort (N = 7,497) provides >99% power to detect correlations of $|\rho| \geq 0.10$ at $\alpha = 0.05$ (post-hoc power analysis). The minimum detectable effect at 80% power is $|\rho| = 0.036$, providing adequate sensitivity for the modest effect sizes typical in passive-sensing digital phenotyping. The GLOBEM cohort comprises 704 participant-wave observations from $N_{\text{unique}} = 497$ unique individuals; because some participants contributed multiple annual waves, power is computed conservatively on the number of unique participants rather than total observations. At $N = 497$, the cohort achieves 80% power for $|\rho| \geq 0.13$ at $\alpha = 0.05$, and all validation tests that assume independent rows (e.g., Test 1, Held-out confirmation) are computed on one randomly selected wave per participant to eliminate within-subject correlation. The WEAR-ME cohort (N = 1,078) achieves 80% power for $|\rho| \geq 0.09$.

Analysis status. All analyses were exploratory; endpoints and analysis strategies were not registered in a public trial registry prior to pipeline execution. The CoDaS pipeline autonomously selects feature engineering strategies, statistical tests, and ML methods based on data characteristics. No subgroup analyses or biomarker candidates were pre-specified.

4. Candidate Biomarker Prioritization Tasks

We designed biomarker discovery tasks spanning mental health and metabolic disease domains to evaluate the ability of CoDaS to autonomously generate, test, and interpret physiological biomarkers from wearable data. Each task was formulated as a hypothesis generation problem in which the system iteratively proposes candidate features, executes statistical tests, and refines its search guided by mechanistic priors from the medical literature. This design parallels recent frameworks for open ended discovery in artificial intelligence while remaining grounded in real world biomedical data consistent with extensive cohort studies of behavioral tracking and cardiometabolic risk.

4.1. Task 1: Mental Health and Circadian Resilience Signatures

Using the Digital Wellbeing Hourly dataset, CoDaS aimed to discover wearable correlates of depression severity as measured by the Patient Health Questionnaire 8. The system integrated millions of hourly observations spanning resting heart rate, step execution, and smartphone application usage to identify behavioral states associated with depression symptom severity. Iterative hypothesis refinement by generative interpreters, guided by clinical literature on sleep disturbance, directed the search toward composite indices capturing circadian resilience. Quantitative findings are reported in Section 5.

4.2. Task 2: Longitudinal Behavior Modeling for Depression

To assess autonomous pipeline generalization across diverse sensor modalities and temporal cohorts, CoDaS evaluated the Generalization of Longitudinal Behavior Modeling dataset. This task required identifying robust behavioral signatures of shifting depression severity tracked over four consecutive annual cohorts (2018–2021). Rather than relying strictly on continuous physiological output, the system adapted to passive smartphone sensing and aggregated wearable metrics. The parallel exploration runners surfaced diminished geospatial mobility and increased day-to-day sleep variability as candidate indicators of depressive transitions. Despite the inherent noise of longitudinal passive tracking, CoDaS maintained its rigorous evaluation framework, utilizing adversarial validation protocols to discard spurious correlations resulting from academic calendar artifacts. This task provided an opportunity to stress-test the pipeline under noisier passive-sensing conditions outside more controlled clinical settings.

4.3. Task 3: Metabolic Risk Stratification and Insulin Resistance

Using the Wearables for Metabolic Health cohort comprising 1,078 comprehensively characterized participants, CoDaS investigated physiological features predictive of metabolic dysfunction, verified against comprehensive fasting laboratory assays. Through extensive temporal modeling over multimodal continuous sensing streams, the system sought to identify composite features mapping directly to insulin resistance measured via the homeostasis model assessment of insulin resistance. This task uniquely challenged CoDaS to distinguish wearable derived (noninvasive) biomarkers from clinical laboratory features, with both categories subjected to the full internal screening battery. Quantitative findings are reported in Section 5.

4.4. Cross Domain Validation and Interpretability

In the DWB and WEAR-ME cohorts, CoDaS produced biomarker candidates that were internally robust and physiologically coherent; the GLOBEM cohort yielded fewer battery-passing candidates, consistent with its analytical constraints (54.6% feature-level missingness, coarse PHQ-4 endpoint; see Section 5). Cross domain evaluation revealed strong alignment between statistical significance and mechanistic plausibility. In DWB and WEAR-ME, the principal candidates reported in Table 3 passed the structured internal validation battery or were explicitly downgraded for construct overlap. In GLOBEM, low-signal candidates are reported separately as exploratory or unstable because several passed only a subset of the checks under substantial missingness and near-chance predictive performance. Ablation evidence throughout development indicates that operating without literature-grounded reasoning substantially reduces physiological interpretability, underscoring the importance of integrating autonomous search with domain priors. Together, these results position CoDaS as a capable framework for interpretable biomarker discovery across heterogeneous disease processes.

Comparison with existing AI discovery systems. Table 2 positions CoDaS relative to four contemporary AI-assisted discovery systems. Google's **AI co-scientist** (Gottweis et al., 2025) is a general-purpose

Table 2 | **Comparison of agent-based scientific discovery frameworks.** CoDaS integrates hypothesis-generation, literature reviews, and statistical validation specifically for wearable biomarker prioritization. While existing systems emphasize general scientific reasoning or algorithmic discovery, CoDaS implements physiologically interpretable, population-level candidate-prioritization and internal validation procedures for wearable healthcare. · indicates not applicable or not designed for this domain

Method	**CoDaS (Ours)**	**AI co-scientist**	**AlphaEvolve**	**Biomni**	**Data Science Agent**
Primary Focus	Biomarker discovery	General science discovery	Algorithm design	Biomedical AI	Data analysis
1. Research					
Hypothesis generation	✓	✓	·	✓	·
Literature exploration	✓	✓	·	✓	·
Experimental design	✓	✓	·	✓	·
Iterative refinement	✓	✓	✓	✓	·
2. Data Science					
Time-series analysis	✓	·	·	·	·
Wearable data processing	✓	·	·	·	·
Code execution	✓	✓	·	✓	✓
Statistical validation	✓	✓	·	·	·
Population-level calibration	✓	·	·	·	·
3. Collaboration & Oversight					
Human-in-the-loop	✓	✓	·	✓	·
Multi-agent debate	✓	✓	·	·	·
Transparent audit trail	✓	✓	·	✓	·
4. Domain					
Healthcare/Medical	✓	·	·	✓	·
Wearable	✓	·	·	·	·

scientific reasoning system that excels at hypothesis generation, literature synthesis, and multi-agent debate (tournament-based idea refinement). Its data science research module can execute code in a sandboxed environment (data loading, aggregation, statistical analysis, and ML modeling) to support hypothesis evaluation; however, code execution is orchestrated by the platform as a knowledge-building step rather than integrated into the iterative hypothesis generation loop. Google DeepMind's **AlphaEvolve** (Novikov et al., 2025) targets algorithmic and mathematical discovery through evolutionary search with iterative refinement, operating in a fundamentally different domain from clinical biomarker identification; it does not perform literature search, hypothesis generation, or statistical validation in the biomedical sense. **Biomni** (Huang et al., 2025) is a general-purpose biomedical AI agent that executes LLM-generated Python code within a ReAct loop, performs PubMed literature search, and can reason about experimental design; however, it does not implement structured statistical validation pipelines, cross-validated ML with FDR correction, or population-level calibration against clinical endpoints. Google ADK's **Data Science (DS) Agent** executes a deterministic Python pipeline that can load data, compute correlations, and train ML models, but operates without literature grounding, hypothesis generation, adversarial validation, or iterative refinement; it runs a fixed linear sequence (load → clean → EDA → ML) without LLM-guided analytical decision-making. Among the evaluated configurations and based on public system descriptions, CoDaS is the only system in this comparison that integrates these four capability dimensions within a wearable biomarker discovery

workflow.

5. Results

To evaluate the discovery capabilities of CoDaS, we deployed the framework independently on three clinically distinct datasets: high-frequency mental health tracking (Digital Wellbeing Hourly, N = 7,497), multiwave longitudinal behavioral phenotyping (GLOBEM, N = 704), and cross-sectional cardiometabolic risk stratification anchored by fasting blood panels (Wearables for Metabolic Health cohort, N = 1,078). Each dataset presents qualitatively different analytical challenges; dense temporal streams, sparse longitudinal sensing with severe missingness, and wearable laboratory feature integration, allowing us to assess the pipeline's versatility across diverse data modalities and clinical endpoints without implying transfer or generalization between cohorts.

All reported candidate biomarkers are subjected to a structured internal validation battery organized around four complementary dimensions: replication, stability, robustness, and discriminative power, operationalized via 11 checks executed in a deterministic subprocess (detailed in Section 2.6). For each dataset and discovery round, we applied Benjamini–Hochberg FDR correction ($\alpha = 0.05$) across the full family of univariate feature tests evaluated in that round. Candidates surviving round-level correction were then tracked across rounds, and cumulative reporting was restricted to features that remained significant under the final round-level screened set. Predictive performance is reported using 5-fold cross-validation with stratified participant-level splits to prevent information leakage between discovery and evaluation partitions. Subgroup consistency was assessed across biological sex (female vs. male) and age decade; features were required to show a consistent direction of effect across all subgroups to pass the pipeline. Statistical power for the age-decade test is concentrated in the young-to-middle-aged range that dominates all three cohorts. In DWB about 11% of participants were younger than 30 and 32% were 50 or older, with only 3.5% aged 70 or older. In WEAR-ME 39% were 50 or older and 5.1% (55 participants) were 70 or older. GLOBEM was an undergraduate cohort with no participant over 25. The age-decade check should therefore be read as evidence of directional stability across young-to-middle-aged strata rather than across the full adult age range, and the candidate biomarkers require dedicated validation in older, multimorbid populations (see Limitations). We note that the 11 checks are not fully independent (e.g., features with strong replication correlations will typically also pass bootstrap stability); the battery is best understood as a structured audit across four complementary validation dimensions rather than 11 orthogonal significance tests. No blinding of the analytical pipeline was performed, as CoDaS supports optional expert feedback; human review occurred in the post-discovery feedback phase and was restricted to mechanistic interpretation (see Section 2.5). A summary of comparative performance across methods and datasets is provided in Table 4.

5.1. Mental Health Monitoring via Digital Phenotyping

In the Digital Wellbeing Hourly study, CoDaS navigated over 4.5 million physiological and device interaction records to identify candidate predictors of depression severity (Patient Health Questionnaire

8). The system identified structural variance in sleep, quantified as main sleep duration variability (ρ = 0.252, 95% CI [0.23, 0.27], $p < 0.001$), as the top-ranked candidate predictor of depression severity. This finding is consistent with an established body of literature linking sleep variability to depression severity (Fang et al., 2021), and its autonomous recovery by CoDaS serves as a sanity check demonstrating the pipeline's ability to recapitulate known clinical signals without prior instruction. Additionally, CoDaS identified elevated nocturnal social application usage (ρ = 0.246, 95% CI [0.22, 0.27], $p < 0.001$) as a candidate behavioral correlate, and generated the hypothesis that nocturnal digital engagement may index nocturnal hyperarousal or sleep-disruptive behavior.

To quantify the incremental value of these candidates beyond established sociodemographic predictors, we compared nested Ridge regression models under 5-fold cross-validation: a demographics-only baseline (8 covariates: age, gender, education, marital status, disability status, Hispanic ethnicity, financial status, living arrangement) achieved CV R^2 = 0.188 (SD 0.010), while adding the top five CoDaS-selected biomarkers yielded CV R^2 = 0.228 (SD 0.010), yielding a ΔR^2 of 0.040. Although modest, this increment is statistically stable across folds and demonstrates that wearable-derived features capture variance in depression severity not explained by sociodemographic factors alone. Effect sizes are modest in absolute magnitude (ρ = 0.15 to 0.25), consistent with the well-documented difficulty of predicting PHQ scores from passive sensing alone, and should be interpreted as prioritized hypothesis-generating signals warranting prospective replication. Since PHQ-8 includes a sleep-related symptom item, sleep-variability candidates may partly capture criterion overlap with the depression endpoint rather than an entirely independent behavioral signal. We therefore interpret these sleep-related findings as candidate monitoring signals requiring endpoint designs that separate sleep symptoms from non-sleep affective burden.

Notably, several of CoDaS's highest-ranked candidates are not raw sensor features but *autonomously constructed composite indices*: the night-to-day social media ratio (ρ = 0.222), which captures the displacement of social engagement into nocturnal hours; the hedonic-to-productivity app ratio (ρ = 0.152), which captures relative enrichment of hedonic compared with productivity-oriented phone use rather than anhedonia itself; and polyphasic sleep percentage (ρ = 0.193), the proportion of nights exhibiting multiple distinct sleep episodes. These composites were generated by the pipeline's feature engineering phase guided by mechanistic priors from the literature grounding phase. The ability to construct and prioritize clinically interpretable composite features, rather than merely rank existing variables, distinguishes CoDaS from conventional automated feature-selection pipelines. Complete lists of all battery-passing (screened and conditionally prioritized) candidates for each cohort are provided in Tables 8–10 in the Appendix. These sleep and nocturnal-use candidates should also be interpreted within the age and health profile of the discovery cohort. Sleep fragmentation, polyphasic sleep and drifting sleep timing can arise from aging, medical comorbidity or medication exposure, so specificity for depression should not be assumed in older or multimorbid populations without age-stratified prospective validation. Throughout this paper, *screened* refers exclusively to survival of the internal validation battery and does not imply prospective clinical validation, external replication, or regulatory endorsement (see Limitations).

Imputation sensitivity analysis confirmed that all prioritized candidates were stable across three

imputation strategies (median, KNN, iterative; maximum $\Delta\rho < 0.001$), and threshold sensitivity analysis showed that the same 13 features survived at both default ($p < 0.05$, $|\rho| \geq 0.20$) and lenient ($p < 0.10$, $|\rho| \geq 0.10$) thresholds, indicating robustness to analytical choices.

When evaluated on the GLOBEM dataset as a stress test of pipeline robustness under data-poor conditions, CoDaS extracted subtle environmental phenotypes across multiwave sensing periods, achieving a primary classification CV AUC of 0.535 using the all-screened-feature evaluation protocol. A separate top-five logistic-regression ablation produced a higher AUC and is reported only as an exploratory component analysis, not as the primary GLOBEM performance estimate. This near-chance discriminative performance reflects the substantial analytical challenges inherent to this cohort: 54.6% feature level missingness, the coarseness of the PHQ-4 outcome instrument (4-item, 0–12 scale), and the heterogeneity of longitudinal passive sensing across annual cohorts. Although discriminative performance was near chance, this result is consistent with a conservative pipeline that did not surface stronger predictive claims from a sparse and noisy cohort. Despite this performance ceiling, the pipeline identified the sequential diversity of unique WiFi access points encountered over a seven day trend ($\rho = 0.128$, 95% CI [0.05, 0.20], $p < 0.001$) as a candidate predictor of depressive shifts. While previous digital phenotyping studies have correlated static location entropy with depression (Saeb et al., 2015), the CoDaS-generated mechanistic hypothesis advanced this by framing dynamic network scanning as a proxy for what the system termed “environmental instability” and “agitated restlessness” (CoDaS-generated interpretive labels, not established clinical constructs). The Google Data Science Agent baseline achieved a nearly identical predictive variance (CV AUC = 0.523), suggesting that the discriminative ceiling is largely data-driven rather than method-driven.

5.2. Cardiometabolic Risk Stratification

Within the Wearables for Metabolic Health cohort, CoDaS mapped continuous physiological readouts against clinical fasting assays to identify candidate predictors of the homeostasis model assessment of insulin resistance. The pipeline first recovered established clinical laboratory markers as a sanity check supporting the analytical validity of the pipeline. It derived TG/HDL ratio ($\rho = 0.542$, $p < 0.001$; subsequently rejected by the construct independence gate as non-independent of the target construct), C reactive protein ($\rho = 0.393$, 95% CI [0.34, 0.44], $p < 0.001$), and HDL cholesterol ($\rho = -0.380$, $p < 0.001$). These laboratory features are included solely to demonstrate that the pipeline recovers known metabolic relationships; the primary translational claim of this cohort rests on the noninvasive wearable-derived candidates described below.

Transitioning to noninvasive wearable sensors, CoDaS identified a derived cardiovascular fitness index (the ratio of step counts to resting heart rate) as a robust metabolic predictor ($\rho = -0.374$, 95% CI [-0.42, -0.32], $p < 0.001$). The system’s mechanistic engine anchored this finding in established metabolic literature (Petersen and Shulman, 2018), hypothesizing that the index reflects peripheral glucose disposal efficiency, a process primarily driven by skeletal muscle mitochondrial capacity. Beyond the cardiovascular fitness index, CoDaS autonomously constructed two additional composite features with strong associations: a derived AST/ALT ratio (the De Ritis ratio; $\rho = -0.375$, $p < 0.001$), a hepatic function index used in liver disease staging, and a known correlate of insulin resistance, here

recovered as a strong signal in a general population cohort; and an HRV-to-RHR ratio ($\rho = -0.203$, $p < 0.001$), capturing the balance between parasympathetic tone and sympathetic activation (not shown in Table 3, which reports a curated subset of candidates per cohort). The pipeline also identified red cell distribution width (RDW; $\rho = 0.281$, $p < 0.001$) as an emerging metabolic inflammation marker (Barbieri et al., 2001; Salvagno et al., 2015), consistent with a small but growing body of evidence linking erythropoietic stress to insulin resistance. Because AST/ALT ratio and RDW can also reflect frailty, sarcopenia, anemia or chronic inflammatory disease, these laboratory candidates should be interpreted as metabolic-risk correlates in this cohort rather than disease-specific insulin-resistance biomarkers. To isolate the wearable-only contribution, a Ridge model using only wearable-derived features achieved CV $R^2 = 0.281$, while the full model including clinical laboratory features achieved CV $R^2 = 0.389$; relative to the demographics-only baseline (CV $R^2 = 0.260$), wearable features alone contributed $\Delta R^2 = 0.021$, a modest increment that highlights the difficulty of extracting metabolic signal from consumer-grade wearables when clinical laboratory features are available. The primary translational value of the wearable-derived cardiovascular fitness index lies not in predictive superiority over blood panels but in its noninvasive, continuously measurable nature, which could enable longitudinal monitoring where repeated phlebotomy is impractical. This interpretation assumes that mobility limitations and rate-modifying medications are observed or can be adjusted for. In older or multimorbid populations, step counts can reflect osteoarthritis, gait limitation or disability, and resting heart rate can be driven by beta-blockers or other cardiovascular medications, so the index requires validation of medication and mobility prior to general clinical use. The Google Data Science Agent baseline, which framed the task as binary insulin-resistance classification, struggled with the inherent collinearity of continuous physiological data, yielding an overfit random forest classifier that failed to generalize beyond chance (CV AUC = 0.429).

5.3. Hypothesis-generating convergence across depression cohorts

Although no identical candidate feature was replicated in an independent cohort with an aligned endpoint, the two depression-focused cohorts provided hypothesis-generating evidence for related circadian-instability features. In the DWB cohort (N = 7,497, PHQ-8), CoDaS ranked sleep duration variability as the top candidate correlate of depression severity ($\rho = 0.252$). In the GLOBEM cohort (N = 704 participant-wave observations from 497 unique individuals, PHQ-4), sleep onset time variability also emerged as an exploratory circadian-instability candidate ($\rho = 0.126$). The pipeline also initially prioritized evening incoming call duration ($\rho = -0.145$), but we treat this association as unstable rather than as evidence for cross-cohort consistency because its held-out estimate reversed direction and the clinician panel rated it lowest for validity. The DWB and GLOBEM sleep-related candidates differ in operationalization, endpoint instrument, and study population (US adults versus US college students), so this pattern should be interpreted as suggestive construct-level convergence around circadian instability rather than direct replication. Given the near-chance discriminative performance in GLOBEM (CV AUC = 0.535), this observation remains hypothesis-generating.

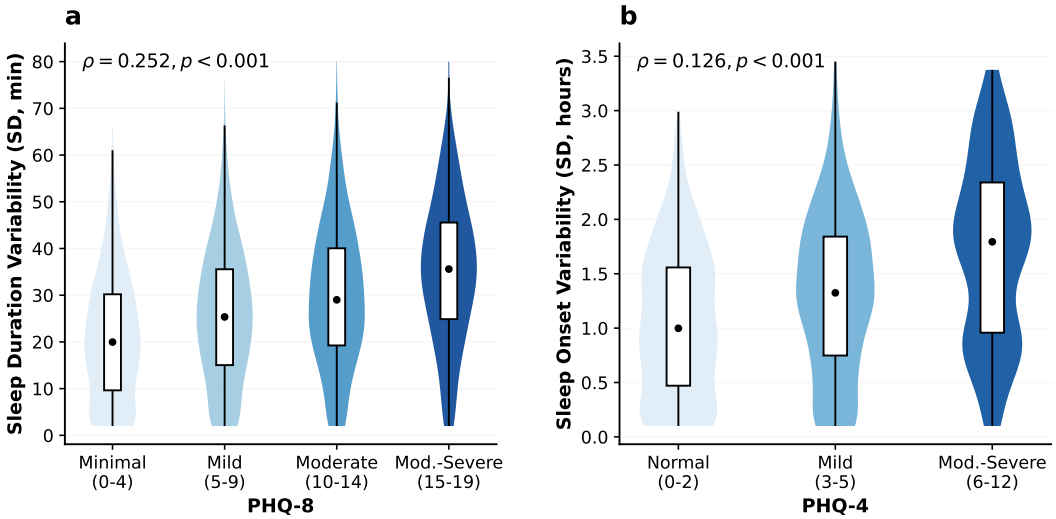


Figure 3 | **Related circadian-instability candidates in two depression cohorts. (a)** The DWB cohort identifies sleep duration variability as a top-ranked candidate correlate of depression severity, and **(b)** The GLOBEM cohort surfaces sleep onset variability as an exploratory circadian-instability-related candidate. Although the specific operationalizations differ, both cohorts point to circadian instability as a hypothesis-generating construct. This pattern is suggestive rather than confirmatory, especially given the near-chance discriminative performance in GLOBEM. Violins show the distribution of the candidate biomarker across depression severity bands. Internal boxplots show medians and interquartile ranges. The increasing trend across severity strata is consistent with a circadian-instability hypothesis, but should not be interpreted as direct replication because the cohorts, instruments and feature operationalizations differ.

5.4. Ablation of Adversarial Evaluation and Failure Cases

To quantify the contribution of each CoDaS component, we conducted systematic ablation experiments across all three datasets, removing one module at a time from the full pipeline (Table 4).

The integration of the adversarial Critic and Defender debate framework differs from standard automated feature-selection pipelines. Without adversarial oversight, tautological features (such as monotonic transformations of the target variable, e.g., squared proxies of blood glucose) can pass statistical screening and manifest as severe data leakage. In a preliminary ablation configuration that additionally disabled construct overlap analysis, such features inflated the CV R^2 to 0.963 on WEAR-ME. In the full pipeline, the Critic agent successfully identified and rejected these tautologies (e.g., the feature `glucose_sq` was explicitly rejected despite passing 10/11 statistical tests) by recognizing their lack of genuine construct independence, preventing the leakage feature from being reported as an independent candidate.

Adversarial audit. The WEAR-ME leakage case illustrates the specific role of the Critic–Defender step. The Defender initially argued to retain *glucose_sq* because the feature passed 10 of 11 statistical checks and produced an apparently large gain in predictive performance. The audit flagged this

interpretation as invalid because the feature was a monotonic transformation of an excluded glycaemic variable and therefore represented target-proximal leakage rather than an independent candidate biomarker. The Orchestrator consequently rejected the feature despite its statistical strength. A second audit rejected TG/HDL despite its strong association with HOMA-IR because triglycerides and HDL are already clinically measured metabolic components; the candidate was retained only as a positive control for construct-independence gating. In contrast, the steps/resting heart-rate index was retained as a wearable-derived candidate because it was noninvasive, passed the internal audit, and was not a direct component or transformation of the target, while being flagged for medication- and mobility-aware prospective validation. Beyond leakage and redundancy, the same step addresses confounding and the correlation-to-causation gap. For behavioral candidates such as nocturnal application use, the Critic flags that the association may be driven by a shared cause rather than a direct effect, and because observational data cannot settle this question, such candidates are retained as hypothesis-generating signals subject to the causal-robustness control rather than asserted as causal. These four functions, namely detecting leakage, detecting redundant or proxy features, testing confounding sensitivity, and marking the correlation-to-causation limit, define what the adversarial step can and cannot establish.

We note that removing adversarial debate left the cross-validated R^2 essentially unchanged on DWB (0.228 in both conditions) and marginally higher on WEAR-ME (0.389 to 0.407). Because the debate stage filters candidates after model fitting rather than re-training the model, its contribution appears in the reliability and verdict quality of reported candidates, as in the `glucose_sq` rejection above, rather than in the cross-validated point estimate. The small WEAR-ME difference also reflects single-run variation in the LLM-guided feature proposals. Removing the Scout agent (responsible for initial data profiling and clinical target identification) produced the largest performance degradation on DWB (CV R^2 dropping from 0.228 to 0.120), demonstrating that informed analytical framing is essential for efficient biomarker search in large feature spaces. In the post-hoc GLOBEM ablation protocol, removing the iterative discovery loop reduced the number of intermediate candidates from 48 to 4. Because this protocol differs from the primary all-screened-feature GLOBEM evaluation, we interpret this result as evidence about search behavior.

The system's failure cases also suggest that the prioritization scheme was conservative, often rejecting statistically significant but low-value features. Across the Digital Wellbeing cohort, the data science agents initially generated a vast pool of 145 discrete statistical metrics. However, through the 11 component validation pipeline, the orchestrator autonomously filtered this funnel down to just 34 screened candidates, explicitly rejecting dozens of statistically significant but functionally trivial features (e.g., the standard deviation of hourly phone unlocks, $\rho = 0.059$, $p < 0.01$). This stringent self-regulation helps ensure that the final reported candidate features exhibit effect sizes relevant for hypothesis generation, rather than merely exploiting population-scale statistical power.

Autonomous feature construction. A distinctive property of CoDaS is its capacity to propose interpretable composite features that were not present as raw input columns. To avoid mixing primary and low-signal exploratory findings, we summarize this analysis only for the DWB and WEAR-ME

candidate lists. Across these two primary cohorts, 9 of 59 non-rejected candidates (15.3%) were autonomously constructed composite indices: four DWB digital-behavior ratios (night-to-day social media ratio, night-to-day unlock ratio, hedonic-to-productivity app ratio, and night-to-day screen ratio) and five WEAR-ME derived indices (cardiovascular fitness index, AST/ALT ratio, cholesterol/HDL ratio, HRV/RHR ratio, and albumin/globulin ratio). The rejected TG/HDL ratio is excluded from this count because it was retained only as a positive control for clinical redundancy and construct-overlap gating.

These composites were generated by feature-engineering module in response to literature-grounded physiological rationales. For instance, the cardiovascular fitness index was proposed after the system retrieved literature linking cardiorespiratory fitness, habitual activity, autonomic tone, and insulin sensitivity. In a descriptive comparison, the retained composite indices showed larger absolute associations with their endpoints than their listed constituent raw features (mean $|\rho| = 0.28$ for composites versus 0.21 for constituent raw features). This comparison was not used as an inferential test or as an independent validation criterion.

The value of explicit composite construction is interpretability rather than access to otherwise unavailable information. For the additive linear models used in the candidate summaries, a pre-computed ratio provides a compact nonlinear transformation that is not represented by including only the two raw variables as separate additive terms. More flexible models with interactions or transformations could, in principle, recover related information. Therefore, we interpret autonomous feature construction as a mechanism for producing compact, auditable candidate definitions for follow-up testing, not as evidence that the ratios are causal biomarkers or that their constituent variables contain inaccessible signal.

5.5. Benchmark Evaluation

To assess whether the CoDaS architecture possesses the analytical capabilities required for biomarker discovery, spanning (i) data profiling, (ii) statistical analysis, (iii) causal inference, (iv) code driven computation, (v) iterative hypothesis refinement, and (vi) clinical reasoning, we evaluated the system across six benchmarks that collectively probe every stage of the translational data science and research pipeline (see Fig. 9). Unlike narrow code completion benchmarks such as HumanEval (Chen et al., 2021) or static medical knowledge examinations like MedQA (Jin et al., 2021), the selected benchmarks require end-to-end analytical reasoning over real-world tabular datasets, multi-step code generation and execution, and integration of domain expertise, the same skills that underpin biomarker identification from clinical and omics cohorts.

Datasets. Each benchmark was chosen to stress test a specific facet of the biomarker discovery workflow. For instance, **DiscoveryBench** (Majumder et al., 2024) evaluates the complete hypothesis generation pipeline: given raw scientific datasets and a natural language research question, the system must autonomously perform exploratory data analysis, variable selection, statistical testing, and articulate a structured hypothesis, precisely the intellectual workflow a translational researcher follows

when searching for candidate biomarkers in a new cohort. **HealthBench** (Arora et al., 2025) and its **Hard** subset assess clinical reasoning fidelity across 5,000 physician designed, multiturn medical conversations spanning 674 diseases and 26 specialties, ensuring that the system's medical knowledge and safety aware communication meet the standards required for biomarker interpretation in clinical contexts. **DataSciBench** (Zhang et al., 2025) measures end to end data science code generation across 222 tasks involving data cleaning, statistical computation, machine learning, and visualization, the programmatic building blocks of any computational biomarker pipeline. **DSBench** (Jing et al., 2024) tests analytical reasoning over large, heterogeneous datasets sourced from real data competitions, requiring multitable joins, domain specific statistical reasoning, and quantitative answer extraction, challenges that directly parallel working with multimodal clinical datasets. Finally, **DSGym** (Nie et al., 2026), which integrates QRData (Liu et al., 2024) (statistical and causal reasoning) and DAEval (Hu et al., 2024) (data analysis tasks), specifically probes causal inference capabilities and quantitative reasoning grounded in real data, skills that are essential for distinguishing genuine biomarker associations from confounded correlations.

Results. We used external data-science and clinical-reasoning benchmarks as supplementary stress tests of the system components. We emphasize that this comparison is inherently asymmetric: CoDaS is a domain-specialized multi-agent *system* with iterative execution, domain-specific tooling, and multi-agent orchestration, whereas baselines are individual models or generic agent frameworks; reported margins should be interpreted as system-level comparisons under this evaluation setup rather than model-to-model differences (Fig. 9). These benchmarks serve as supplementary validation of the analytical capabilities underlying the biomarker discovery pipeline, not as the primary contribution of this work. On **HealthBench**, CoDaS attained an overall score of 0.724, surpassing the previous best result of o3 (0.598) by 12.6 percentage points. On the **HealthBench Hard** subset, which comprises cases empirically identified as resistant to all frontier models, CoDaS scored above the o3 baseline in this evaluation. The remaining baselines scored substantially lower: GPT-4.1 at 0.230, Gemini-2.5 Pro at 0.190, and o1 at 0.160. On **DiscoveryBench**, CoDaS achieved a Hypothesis Matching Score (HMS) of 0.32 on real scientific datasets, exceeding the oracle assisted Reflexion agent with GPT-4o backbone (0.24) by 8.0 percentage points and the Reflexion agent with Llama-3 backbone (0.23) by 9.0 percentage points. Notably, the Reflexion+Oracle baseline receives the gold HMS score as iterative feedback, an advantage CoDaS does not require; without oracle access, CodeGen and ReAct agents achieve only 0.15 HMS each. On **DataSciBench**, CoDaS reached a completion rate of 77.5%, outperforming GPT-4o (68.4%) by 9.1 percentage points, DeepSeek-Coder-33B (61.2%) by 16.3 pp, GPT-4 Turbo (58.9%) by 18.6 pp, and Claude-3.5 Sonnet (58.1%) by 19.4 pp. On **DSBench**, CoDaS achieved 64.0% task accuracy, matching the reported human expert baseline for this benchmark (64.0%) and exceeding the best prior agent system AutoGen+GPT-4o (34.1%) by 29.9 pp, Gemini-1.5 Pro (31.6%) by 32.4 pp, GPT-4o (28.1%) by 35.9 pp, and GPT-4 (26.0%) by 38.0 pp. On **DSGym**, CoDaS scored 84.4% overall accuracy, surpassing Kimi K2 (79.9%) by 4.5 pp, Claude Sonnet 4.5 (78.2%) by 6.2 pp, GPT-4o (78.0%) by 6.4 pp, and DeepSeek v3.1 (71.2%) by 13.2 pp.

Scientific Hypothesis Generation. DiscoveryBench (Majumder et al., 2024) requires autonomous

hypothesis generation from raw scientific datasets. CoDaS achieved HMS of 0.32, exceeding the oracle-assisted Reflexion agent (0.24) by 8.0 pp, attributable to domain-aware data profiling, iterative gap-based discovery, and question-type-adaptive analysis strategies.

Clinical Reasoning. On HealthBench (Arora et al., 2025) (5,000 physician-designed medical conversations), CoDaS scored 0.724 versus o3's 0.598 (+12.6 pp). On the Hard subset (cases resistant to all frontier models), CoDaS scored 0.391 versus 0.320 (+7.1 pp), with particularly strong performance on hedging under uncertainty (0.565), suggesting that the adversarial critic–defender architecture produces calibrated uncertainty.

Data Science Code Generation. On DataSciBench (Zhang et al., 2025) (222 end-to-end data science tasks), CoDaS achieved 77.5% completion rate versus 68.4% for GPT-4o (+9.1 pp), with uniform performance across task types (deep learning: 77.8%, CSV/tabular: 77.6%, human-authored: 77.4%).

Real-World Data Analysis. On DSBench (Jing et al., 2024) (466 tasks from professional data competitions), CoDaS achieved 64.0% task accuracy, matching the human expert baseline and exceeding the best prior agent system AutoGen+GPT-4o (34.1%) by 29.9 pp.

Quantitative and Causal Reasoning. On DSGym (Nie et al., 2026) (QRData + DAEval; statistical, causal, and data analysis tasks), CoDaS scored 84.4% overall accuracy, surpassing Kimi K2 (79.9%), Claude Sonnet 4.5 (78.2%), and GPT-4o (78.0%).

Cross-benchmark patterns. Three patterns are relevant to the biomarker discovery setting. First, iterative refinement reliably outperforms single-pass generation, with margins exceeding 8 pp on every benchmark employing feedback-driven iteration. Second, the hybrid deterministic–agentic architecture avoids hallucination-prone numerical reasoning by executing computations in deterministic subprocesses while reserving LLM reasoning for interpretation and gap analysis. Third, domain-aware data profiling (automatic detection of replication structures, temporal layouts, categorical encodings) enables appropriate analytical strategy selection without manual specification.

6. User Study

To rigorously assess the quality of CoDaS research outputs against frontier AI-driven scientific discovery systems, we conducted a blinded expert evaluation. Fifteen domain experts independently evaluated manuscripts produced by four systems: (i) CoDaS, (ii) Google's AI co-scientist (Gottweis et al., 2025), (iii) Biomni (Huang et al., 2025), and (iv) Google ADK's Data Science Agent, across three wearable datasets, yielding 34 evaluation sessions and 76 individual manuscript assessments. For CoDaS, Data Science Agent, and Biomni, each reviewer scored all three systems in a single session ($n = 21$ sessions); for AI co-scientist, reviewers scored the system in a separate set of sessions ($n = 13$) using the same instrument and datasets.

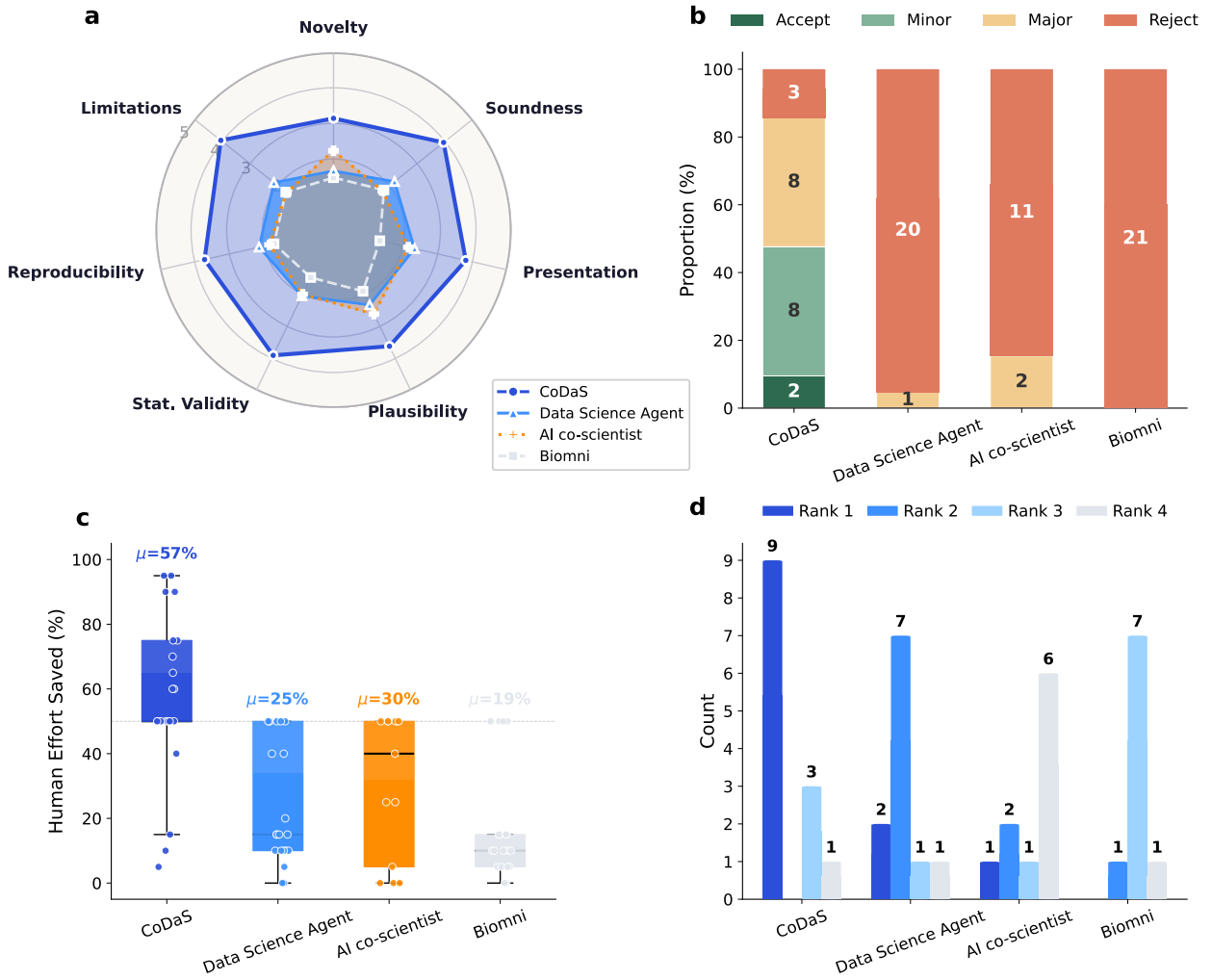


Figure 4 | **Blinded Human Evaluation. (a)** Radar chart of mean expert scores (1 to 5 Likert scale) across seven quality dimensions. CoDaS achieves scores of 3.1 to 4.1, while all baselines cluster at 1.3 to 2.6. AI co-scientist is shown separately as an auxiliary unpaired comparison (n = 13), whereas CoDaS, Biomni, and Data Science Agent are compared within the 21-session balanced subset. **(b)** Simulated reviewer recommendation distribution. CoDaS received 2 Accept, 8 Minor Revision, 8 Major Revision, and 3 Reject (86% non-rejection rate). AI co-scientist received 11 Reject and 2 Major Revision (n = 13). No baseline received Accept or Minor Revision. **(c)** Effort preservation scores (percentage of content reviewers would retain). CoDaS: μ = 56.9% vs. 18.8 to 30.4% for baselines. **(d)** Forced ranking distribution (n = 13 sessions across all four systems). CoDaS was ranked #1 in 9 of 13 sessions (69%).

6.1. Study Design

Blinded evaluation protocol. Each evaluator was presented with de-identified manuscripts (labeled Model A through Model D or Model E) generated from the same input dataset and biomarker discovery task. The mapping between system identities and blind labels was randomized independently for each session, ensuring that reviewers could not infer which system produced which output. Evaluators

were not informed of the number or identity of the systems under comparison. The evaluation was blinded to system identity, randomized model-label mapping, study hypotheses, and other reviewers' assessments. Reviewers were told that they would evaluate blinded reports, but not which system generated each report.

Interface design. Evaluators interacted with a custom web-based review interface designed to emulate the workflow of a biomedical manuscript peer-review process. Each manuscript was rendered within a standardized reading interface that preserved the original structure of the generated report, including title, abstract, introduction, figures, tables, and methodological descriptions. To avoid presentation bias, all manuscripts were formatted using the same layout template and figure rendering pipeline. Reviewers examined one manuscript at a time through a scrollable document viewer and were allowed to navigate freely between sections before completing the evaluation form. The evaluation form was displayed in a structured panel adjacent to the manuscript viewer and included Likert-scale scoring fields, editorial decision options (Accept / Minor Revision / Major Revision / Reject), and free-text feedback fields for qualitative assessment. Representative screenshots of the evaluation interface are provided in Figure 10 through 17 in the Appendix.

Assessment instrument. Our evaluation instrument was designed to mirror established peer review practices at top biomedical venues, comprising four components as follows.

1. **Multi-axis quality assessment** (1 to 5 Likert scale) across seven dimensions: *Novelty* (originality of biomarker hypotheses), *Soundness* (methodological rigor), *Presentation* (writing clarity and structure), *Plausibility* (biological credibility of findings), *Statistical Validity* (correctness of statistical analyses), *Reproducibility* (sufficiency of methodological detail for replication), and *Limitations* (acknowledgment of methodological constraints).
2. **Reviewer recommendation**: Accept, Minor Revision, Major Revision, or Reject, using generic peer-review recommendation categories rather than categories specific to any one journal.
3. **Effort preservation score** (0 to 100%): the proportion of the manuscript a domain expert would retain in a revision, quantifying practical utility.
4. **Safety and reliability audit**: reviewers flagged instances of hallucinated content, statistical errors, logical flaws, or biological contradictions, with mandatory free-text justification for each flag.

Additionally, reviewers provided: (i) a forced ranking of all systems from best to worst with free-text rationale; (ii) estimates of the person-days each research phase would require if conducted manually (data preprocessing, feature engineering, ML modeling, literature review, result interpretation, and paper writing); and (iii) the validation steps they would require before considering findings publishable.

Expert panel. The evaluation panel consisted of 15 domain experts with backgrounds in medicine, biomedical data science, machine learning, bioinformatics, and digital health research. Panelists

had between 0 and 19 years of research experience (mean: 6.7 years) and had prior experience conducting peer-reviewed biomedical research or reviewing scientific manuscripts. Experts were recruited from both academic and industrial research environments. All reviewers were blinded to the identities of the models and the study hypotheses throughout the evaluation process.

6.2. Results

CoDaS was the only system to receive Accept/Minor Revision recommendations in blinded evaluation. Figure 4b presents the editorial decision distribution. CoDaS received 2 Accept, 8 Minor Revision, 8 Major Revision, and 3 Reject decisions ($n = 21$), corresponding to an 86% non-rejection rate (18 of 21 assessments). In contrast, Biomni received 21/21 Reject decisions, AI co-scientist received 11 Reject and 2 Major Revision ($n = 13$), and Data Science Agent received 20 Reject and 1 Major Revision. No baseline system received an Accept or Minor Revision decision from any reviewer. Fisher's exact tests confirmed the significance of these differences: CoDaS versus each baseline for the non-rejection rate (18/21 vs. 0/21 for Biomni, OR = ∞, $p_{\text{adj}} = 2.3 \times 10^{-8}$; 18/21 vs. 2/13 for AI co-scientist, OR = 33.0, $p = 7.7 \times 10^{-5}$; 18/21 vs. 1/21 for Data Science Agent, OR = 120.0, $p_{\text{adj}} = 3.8 \times 10^{-7}$).

On the WEAR-ME dataset, where CoDaS's hold-out validation pipeline was most mature, the non-rejection rate reached **100%** (2 Accept, 3 Minor, 3 Major; 0 Reject), underscoring the relationship between pipeline completeness and perceived manuscript quality.

CoDaS scored higher than all baselines across quality dimensions. Figure 4a presents the multi-axis radar comparison. CoDaS achieved a mean overall score of 3.74 (95% CI: 3.42 to 4.06) across all seven assessment dimensions ($n = 21$), compared to 2.12 (1.85 to 2.39) for Data Science Agent ($n = 21$), 2.04 for AI co-scientist ($n = 13$), and 1.63 (1.45 to 1.81) for Biomni ($n = 21$). Paired Wilcoxon signed-rank tests (matched within each evaluation session) confirmed statistically significant advantages over Data Science Agent and Biomni on the composite score: $\Delta = +2.11$ vs. Biomni (Cohen's $d = 2.40$, $p_{\text{adj}} = 0.001$) and $\Delta = +1.62$ vs. Data Science Agent ($d = 1.64$, $p_{\text{adj}} = 0.001$); all p-values Bonferroni-corrected; $n = 21$ paired assessments per system. A Mann-Whitney U test (unpaired, given the different sample sizes) confirmed that CoDaS also scored significantly higher than AI co-scientist ($U = 265.5$, $p = 5.0 \times 10^{-6}$, Cohen's $d = 2.49$), with all seven individual axes reaching significance at $p < 0.003$. The strongest separations were on Limitations ($d = 3.16$) and Soundness ($d = 2.81$).

CoDaS's advantage was most pronounced on Limitations acknowledgment (4.05 vs. 1.69 to 2.14 for baselines), Soundness (3.95 vs. 1.77 to 2.19), and Statistical Validity (3.90 vs. 1.48 to 2.05), reflecting its multi-stage validation pipeline. This pattern held consistently across all three datasets. On DWB Hourly ($n = 9$), CoDaS scored a mean of 3.76 versus 1.71 to 2.02 for baselines; on GLOBEM ($n = 4$), 3.61 versus 1.64 to 2.21; and on WEAR-ME ($n = 8$), 3.79 versus 1.54 to 2.20. Baseline systems rarely exceeded a mean score of 2.5 on any individual axis, with isolated exceptions on Novelty and Presentation for specific dataset and system combinations.

Reviewers' free-text feedback on AI co-scientist highlighted persistent methodological gaps: "*mul-*

tiple testing correction was not performed; text and figure has AI feel" (R3), "*just suggestions of ideas*" (R4), "*heavy on new hypothesis, but very weak on analysis to test and prove them*" (R15), and "*some of the figures look hallucinated*" (R4).

Effort preservation quantifies practical utility. The effort preservation score measures how much of a generated manuscript a domain expert would retain in a revision (Figure 4c). CoDaS achieved a mean effort score of 56.9% (95% CI: 45.2 to 68.6%; range: 5 to 95%), indicating that reviewers would, on average, keep over half of the generated content as a starting point for their own work. Baseline systems scored markedly lower: 18.8% (Biomni), 24.5% (Data Science Agent), and 30.4% (AI co-scientist, $n = 13$), meaning fewer than one-third of baseline outputs were judged salvageable. Paired Wilcoxon signed-rank tests confirmed significant effort advantages over the paired baselines: $\Delta = +38.1\%$ vs. Biomni ($d = 1.11$, $p_{\text{adj}} = 0.003$) and $\Delta = +32.4\%$ vs. Data Science Agent ($d = 0.85$, $p_{\text{adj}} = 0.007$). Effort preservation for CoDaS was also significantly higher than for AI co-scientist (Mann-Whitney $p = 0.003$, $d = 1.08$). Multiple reviewers assigned CoDaS effort scores of 90 to 95%, suggesting review-ready quality in some instances. One reviewer (R13) characterized CoDaS output as comparable to "*a first-year PhD student, if I was able to have discussions with them, I think we could iterate on this original research.*"

Accordingly, the user study should be interpreted primarily as evidence of strong relative preference among systems rather than high absolute agreement in the assigned scores.

6.3. Safety and Reliability Analysis

A critical dimension for clinical and biomedical applications is minimizing hallucinated or erroneous content that could propagate to downstream research (Kim et al., 2025b). Reviewers flagged a total of 51 safety concerns across the three baseline systems, compared to 3 for CoDaS (Table 5). The three CoDaS flags were: one statistical error (a confidence interval whose bounds did not contain the reported point estimate), one hallucination flag (incorrect citations in references), and one other issue (a LaTeX truncation artifact). While reviewers categorized the first two under StatError and Hallucination, respectively, manual inspection confirmed these were isolated reporting errors rather than systematic data fabrication or flawed analytical reasoning. AI co-scientist received 9 safety flags ($n = 13$), with hallucination as the primary concern (5 flags; 0.69 flags per session). Among the other baseline systems, the most prevalent safety concerns were:

- **Hallucination** (8 flags for Data Science Agent and Biomni combined): fabricated images, invented data, and false claims of successful execution. Biomni "*hallucinated a successful literature review outcome from failed API calls*" (R2) and produced manuscripts that "*claim strong biomarker candidates while simultaneously printing that those candidates are non-existent and the tables are empty*" (R16).
- **Logic errors** (12 flags): contradictory conclusions, incoherent reasoning, and broken analytical workflows. Biomni was "*congratulating itself for work it explicitly failed to produce*" (R2). Data Science Agent exhibited "*fatal statistical errors: running 159 tests with $p < 0.05$ without multiple comparison correction fundamentally invalidates the primary claims*" (R7).

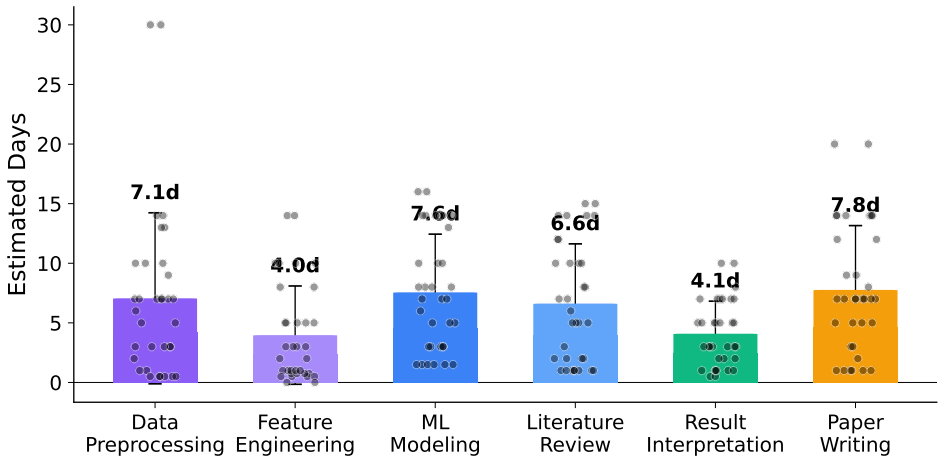


Figure 5 | **Estimated human effort to manually reproduce the equivalent research workflow.** Expert estimates of person-days per research phase (n = 34 responses). Mean total: 37 days (median: 40; range: 5 to 90 days). Paper writing (7.8d), ML modeling (7.6d), and data preprocessing (7.1d) were the most time-intensive phases. CoDaS reduces end-to-end wall-clock time for an automated discovery run to 6–8.5 hours on a single machine. This should be interpreted in contrast to reviewers' estimates of manual human effort (mean: 37 person-days), not as a direct productivity ratio. Error bars: standard deviation; grey markers: individual estimates.

- **Statistical errors** (11 flags): data leakage, missing corrections, and sub-random performance. Data Science Agent committed "*a textbook example of definitional label leakage, invalidating the primary claim*" (R2) and reported "*AUC-ROC below 0.5, implying the model is predicting the wrong class*" (R2). Multiple reviewers noted that Biomni produced "*an AUC-ROC of 0.517, statistically indistinguishable from random noise*" (R2) and failed to apply "*imputation before train/test split, introducing data leakage*" (R7).
- **Biological contradictions** (6 flags): claims inconsistent with established domain knowledge. Data Science Agent "*ranked sodium and chloride as top biomarkers for insulin resistance, primary electrolytes with no established role as direct IR biomarkers*" (R8), while "*discovering triglycerides as the top predictor for metabolic syndrome is a textbook example of definitional label leakage*" (R2).

The lower number of safety flags for CoDaS may reflect the contribution of its multi-stage validation architecture, although the causal contribution of individual safeguards was not isolated in this study.

6.4. Human Effort Estimation

To contextualize the practical impact of end-to-end automation, we asked each reviewer to estimate the person-days required to manually conduct the equivalent research workflow. Across all 34 responses, the mean estimated total effort was 37 ± 23 person-days (median: 40; range: 5 to 90 days), broken down by research phase in Figure 5. The most time-intensive phases were paper writing (mean: 7.8 days), ML modeling (7.6 days), and data preprocessing (7.1 days). Literature review (6.6 days), result interpretation (4.1 days), and feature engineering (4.0 days) constituted the remaining effort. CoDaS

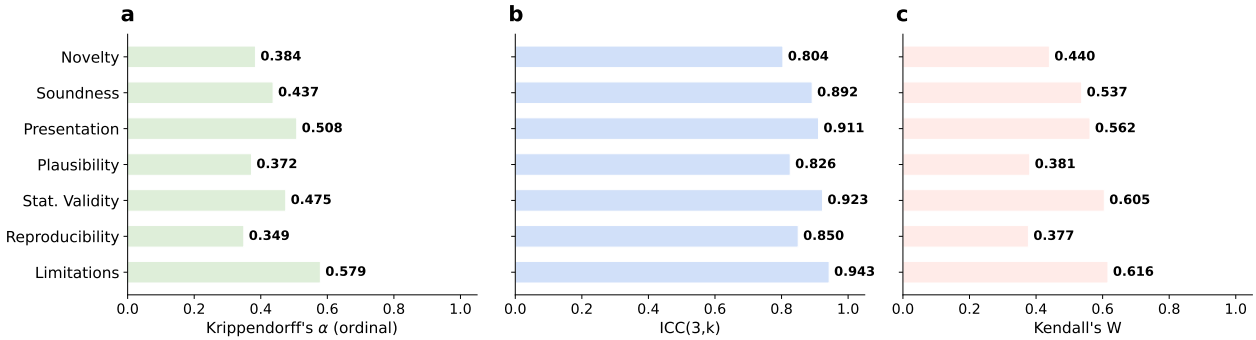


Figure 6 | **Inter-rater reliability across seven assessment axes. (a)** Krippendorff's alpha (ordinal) measures absolute agreement accounting for chance and missing data; overall $\alpha = 0.443$ (low-to-moderate agreement). **(b)** Intraclass correlation coefficient ICC(3,k) measures consistency of relative system rankings; overall ICC = 0.888 (excellent reliability), with all axes significant at $p < 0.0001$. **(c)** Kendall's coefficient of concordance W measures ordinal agreement among raters within datasets; overall $W = 0.503$ (moderate concordance). Dashed lines indicate conventional thresholds. The divergence between α (moderate) and ICC (excellent) indicates that reviewers differed in absolute calibration but strongly agreed on relative system quality.

automates all six phases end-to-end, with a typical wall-clock runtime of 6 to 8.5 hours per dataset on a single machine (a detailed phase-wise runtime comparison across all four systems is provided in Table 7 in the Appendix). This comparison contrasts estimated human labor with automated wall-clock runtime and is intended as an order-of-magnitude illustration rather than a direct productivity ratio. Reviewers judged the resulting outputs substantively usable (57% effort preservation). Reviewers also indicated the validation steps they would require before considering AI-generated findings publishable: external validation on independent cohorts (85%), held-out confirmation of key findings (62%), code review of the analytical pipeline (47%), and wet-lab experimental confirmation (26%). Notably, 15% of respondents selected "never publishable without human involvement," reflecting healthy skepticism about fully autonomous scientific pipelines. These responses position CoDaS outputs not as finished publications, but as high-fidelity first drafts that substantially accelerate the hypothesis-to-validation cycle.

Qualitative feedback and identified limitations. Beyond numerical scores, reviewers provided constructive feedback identifying areas for improvement. Several reviewers noted that CoDaS "*still lacks a modern ML/DL pipeline*" and relies primarily on "*classical ML methods*" for feature selection (R1), which, while appropriate for the current sample sizes, may limit scalability to larger cohorts. Another reviewer observed that "*writing is still not satisfying*" despite being the best among all systems (R1), and one expressed skepticism: "*very suspicious of it and still would never accept it*" (R6), noting concerns about the volume of automated validation. Multiple reviewers suggested incorporating "*human evaluation at intermediate stages to enable collaboration*" rather than fully autonomous end-to-end generation (R4). These critiques, combined with the 14% rejection rate and 3 safety flags, underscore that while CoDaS substantially outperforms existing baselines, its outputs require expert

review and refinement before publication, a positioning we explicitly adopt in our system design.

6.5. Statistical Analysis

Inter-rater reliability. To quantify the degree to which independent reviewers produced consistent assessments, we computed four complementary inter-rater reliability (IRR) metrics across all seven evaluation axes using a balanced subset of sessions where all reviewers scored the same set of systems ($n = 21$ sessions, 3 systems; Figure 6), which provides the multi-rater × multi-item design required for valid IRR computation. Using *Krippendorff's alpha* with ordinal weighting, the most conservative metric appropriate for Likert-scale data with missing raters, we obtained an overall $\alpha = 0.443$. Agreement was strongest for Limitations ($\alpha = 0.579$), Presentation ($\alpha = 0.508$), and Statistical Validity ($\alpha = 0.475$), and weakest for Reproducibility ($\alpha = 0.349$) and Plausibility ($\alpha = 0.372$), consistent with the expectation that subjective dimensions elicit greater rater heterogeneity. Across datasets, agreement was comparable: DWB Hourly ($\alpha = 0.462$), GLOBEM ($\alpha = 0.473$), and WEAR-ME ($\alpha = 0.431$), indicating no dataset-specific calibration bias.

Intraclass correlation coefficients: ICC(3,k), two-way mixed, consistency, average measures yielded substantially higher reliability estimates: overall ICC = 0.888 (excellent; > 0.75 threshold), with all seven axes reaching ICC ≥ 0.804 and six of seven exceeding 0.85 (all $p < 0.0001$). The highest ICCs were observed for Limitations (0.943), Statistical Validity (0.923), and Presentation (0.911). The divergence between Krippendorff's α and ICC is expected: α penalizes systematic rater bias (some reviewers consistently score higher or lower), while ICC(3,k) measures consistency of relative rankings, indicating that although reviewers differed in absolute calibration, they agreed strongly on *which systems were better or worse*.

Kendall's coefficient of concordance (W) confirmed moderate-to-strong concordance among raters in their ordinal rankings of systems within each dataset: overall $W = 0.503$, with Statistical Validity ($W = 0.605$) and Limitations ($W = 0.616$) showing the strongest agreement.

Pairwise Spearman correlations across the 60 unique reviewer pairs with ≥3 common items yielded a mean $\rho = 0.646$ (median: 0.632; range: 0.200 to 1.000), with 13% of pairs reaching statistical significance ($p < 0.05$) despite the limited number of shared items per pair (typically 4).

Taken together, these metrics indicate that despite heterogeneous backgrounds (0 to 19 years experience, 8 institutions, occasional-to-heavy review frequency), our panel exhibited *low-to-moderate absolute agreement* (below the $\alpha = 0.667$ threshold recommended by Krippendorff for tentative conclusions) and *excellent relative consistency*.

7. Clinician Assessment of Translational Relevance

Internal statistical screening can identify associations that are stable within the available data, but it does not establish clinical utility or readiness for patient management. To assess these properties for the candidates in Table 3, we convened a panel of twelve physicians and physicians-in-training involved in clinical care (median 5 years of clinical experience after medical qualification, range 0 to 27) and asked

them to evaluate every candidate against a structured rubric interrogating the translational relevance. The panel spanned psychiatry, internal medicine, family medicine, endocrinology, geriatrics, pain medicine, radiology, and general practice, and was drawn from academic medical centers and clinical research institutions in the United States and the Republic of Korea. Each physician independently reviewed all 16 candidate associations together with the pipeline-rejected TG/HDL ratio, which served as a calibration control, giving 17 candidates reviewed per rater and 204 reviews in total.

7.1. Evaluation

Protocol. Each candidate was presented on a single standardized interface which stated the input feature, the clinical endpoint, the reported effect size in the discovery cohort, and a translation of that effect size into real-world units, expressed as the expected change in the endpoint across the interquartile range and per standard deviation. For example, for main sleep duration variability a discovery-cohort correlation with the PHQ-8 depression score (Spearman $\rho = 0.252$) was accompanied by an empirical distributional translation showing the observed PHQ-8 difference across the feature interquartile range. The demographic covariates available in the cohort model were listed for context. The interface reported derived population-level statistics only and contained no participant-level data.

Rating instrument. For every task, reviewers rated seven dimensions on a five-point scale. The dimensions were validity (confidence that the association is genuine rather than an artifact), effect-size meaningfulness, novelty relative to current knowledge, feasibility of measurement, added value over biomarkers already in use, likelihood of influencing patient advice, and confidence to act on the result for an individual patient. These seven dimensions were selected to span the translational properties most relevant to whether a biomarker would be adopted in practice, rather than adapted from a single existing scale. Reviewers also classified the supporting literature into one of several ordered categories ranging from strongly supportive to contradicted, and provided written justification for the validity, effect-size, and added-value ratings. The task order was randomized for each reviewer to limit order effects.

Calibration probe. The rejected TG/HDL ratio was presented alongside the genuine candidates as a sanity check. Because this ratio is an established lipid-derived metabolic-risk index already available from routine laboratory testing, a well-calibrated panel should rate its validity high and its added value low. A failure to separate these two judgments would indicate that reviewers were responding to surface plausibility rather than to incremental clinical value.

Reviewer effort. The panel completed the tasks asynchronously through a web interface that recorded a timestamp each time each task was opened and submitted. Each reviewer actively rated for a median of just over two hours (range 1.0 to 3.3 hours), for a panel total of approximately 25 hours.

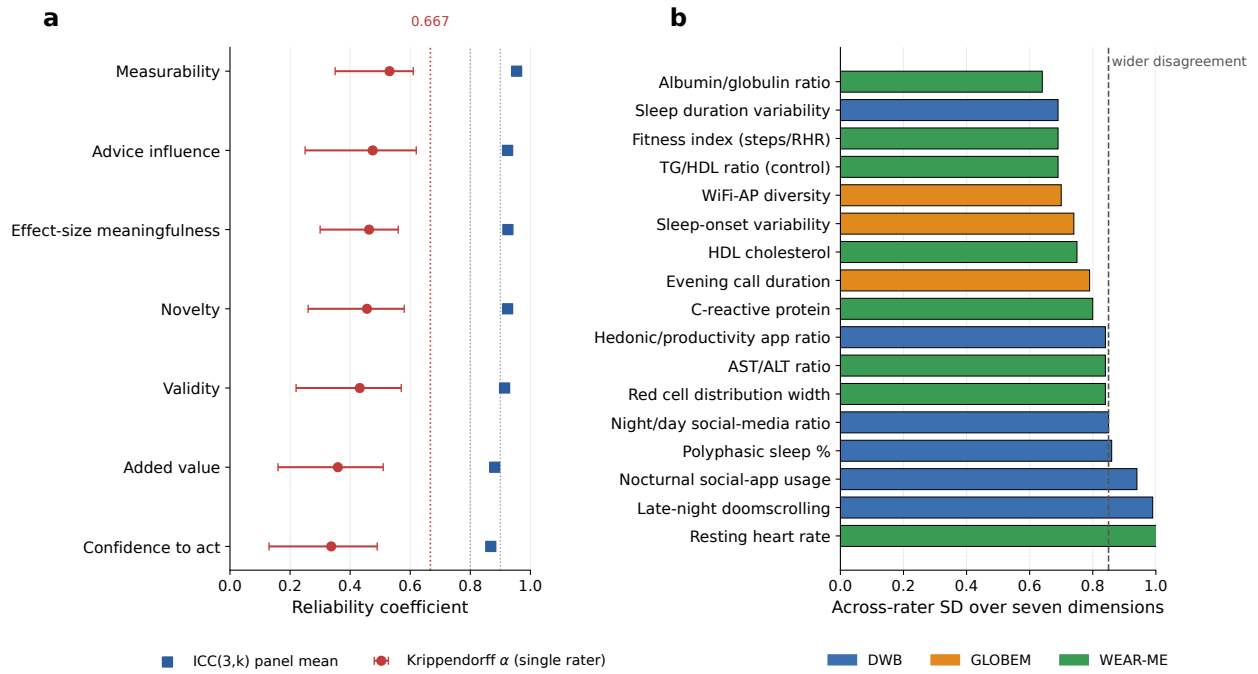


Figure 7 | **Clinician review distinguishes validity from clinical actionability. (a)** Reliability of the seven rated dimensions. Red markers show Krippendorff's ordinal alpha for a single rater with 95% bootstrap confidence intervals from 3,000 resamples, which remain below the 0.667 threshold for tentative agreement on every dimension. Blue markers show the two-way mixed-effects ICC(3,k) for consistency, the same form used in Section 6.5, for the twelve-clinician mean, which reaches the good to excellent range (0.87 to 0.95). The separation between the two reflects differences in absolute scale use alongside agreement on relative ranking. **(b)** Across-rater standard deviation over the seven dimensions for each candidate, colored by cohort. Lower values indicate closer agreement. Candidates to the right of the dashed line fall in the upper tercile of disagreement.

7.2. Results

Physician review aligned with CoDaS prioritization and identified clinical utility gaps. Averaged over all candidates and reviewers, validity received the highest rating (mean 3.69 on the five-point scale), indicating that the panel considered the associations more likely genuine than artifactual. Practical measurability was rated moderate (3.28). The three dimensions tied most directly to clinical action were rated lower, with added value over existing biomarkers at 2.37, confidence to act for an individual patient at 2.34, and likelihood of influencing advice at 2.40. These ratings indicate that the panel distinguished evidence for genuine association from evidence for added clinical value, supporting the use of CoDaS as a candidate-prioritization system whose outputs require follow-up clinical validation before use in patient management. This reading is consistent with the evidence tiers in Table 3, where most candidates are labeled Supported or Emerging rather than Established.

Panel judgment converged with the pipeline's confidence ranking and independently identified the candidate whose statistics did not replicate. Validity ratings were highest for candidates

in the Established tier (mean 4.2) and lower for the newer Supported and Emerging tiers (means about 3.8 and 3.3), and the rating correlated with both the tier CoDaS had assigned (Spearman $\rho = 0.67$, $p = 0.005$) and the reported effect size ($\rho = 0.77$, $p = 0.001$). Part of this agreement reflects representations in routine clinical documentation, since the Established tier consists of quantities that clinicians already encounter in the electronic medical record (EMR) system, such as laboratory markers (HDL and C-reactive protein) and vital signs (resting heart rate). The most informative case was the candidate the pipeline had itself flagged as unstable. Evening incoming call duration, flagged for a reversal of the effect direction between the discovery and held-out estimates, was rated the least valid candidate of all (1.75), and reviewers across specialties independently identified the same reversal. One reviewer observed that *a stable biological signal cannot reverse direction across partitions of the same cohort*. Without access to the tier labels, the panel corroborated the candidates the pipeline ranked highest and isolated the one whose statistical support did not replicate.

Agreement was strongest on the clearest accept and reject decisions. Table 6 and Figure 7b report agreement for each candidate, summarized as the across-rater standard deviation over the seven dimensions. The highest validity ratings went to the established metabolic markers in WEAR-ME, with HDL cholesterol and C-reactive protein rated most clearly genuine (validity 4.75 and 4.33). Agreement was closest on candidates with an unambiguous standing, whether positive, such as the albumin/globulin ratio and the cardiovascular fitness index, or negative, such as evening incoming call duration in GLOBEM, which received the lowest validity (1.75) rating in the set. Disagreement was widest on resting heart rate and on the behavioral DWB candidates such as late-night doomscrolling and nocturnal social-app usage. The highest-rated DWB candidate, main sleep duration variability, drew a high but divided validity rating (4.17 with a standard deviation of 1.11). Eleven of the twelve reviewers rated its validity at 4 or 5, while one rated it 1 on the grounds that the sleep feature overlaps the sleep item of the PHQ-8 outcome, a form of criterion contamination that the effect size alone would not reveal.

Inter-rater reliability matched the pattern seen in the system-level evaluation. We quantified agreement with the same battery used in Section 6.5 and report it in Figure 7a. Single-rater absolute agreement, measured by Krippendorff's ordinal alpha, ranged from 0.34 to 0.53 across the seven dimensions and remained below the 0.667 threshold for tentative conclusions on every dimension. Agreement was highest for measurability (alpha 0.53, 95% CI 0.35 to 0.61) and lowest for added value (0.36, 95% CI 0.16 to 0.51) and confidence to act (0.34, 95% CI 0.13 to 0.49). The reliability of the twelve-clinician mean, measured by a two-way mixed-effects intraclass correlation coefficient for consistency, ICC(3,k), the same form reported for the system-level panel in Section 6.5, ranged from 0.87 to 0.95 and met the conventional standard for good to excellent reliability. As in the system-level study, the gap between low single-rater alpha and high panel-mean ICC indicates that clinicians differed in their absolute use of the scale while agreeing on the relative standing of the candidates. We therefore treat the panel mean, rather than any individual rating, as the unit of analysis. Agreement on the categorical literature item was slight (Fleiss' kappa 0.16), consistent with reviewers drawing on different bodies of evidence. Per-dimension reliability statistics, including the

single-rater ICC(3,1), are reported in the Appendix (Table 12).

The calibration probe behaved as designed. The rejected TG/HDL control received a validity rating of 4.83, above the mean of 3.95 for the genuine WEAR-ME candidates, and an added-value rating of 1.50, below the genuine WEAR-ME candidate mean of 2.46. Reviewers recognized TG/HDL as a quantity that is real but clinically redundant for incremental biomarker discovery, which supports the interpretation that the panel separated statistical validity from incremental clinical value rather than rating candidates favorably by default.

Taken together, these results position the clinician panel as an external check that both corroborated the pipeline and constrained its claims. Expert validity judgments aligned with the pipeline's own confidence ranking, the metabolic candidates were judged closest to clinical use, and the panel rated lowest the candidate whose effect did not replicate. The low ratings on added value and confidence to act indicate that most candidates are best regarded as hypotheses for prospective evaluation rather than biomarkers ready for clinical use. These conclusions are bounded by the composition of the panel, in particular the single psychiatric specialist among the reviewers of the behavioral cohorts, and by the use of a rubric built for this study rather than a previously validated instrument.

Table 3 | **Candidate biomarkers prioritized by CoDaS across three cohorts.** DWB and WEAR-ME candidates passed or were explicitly downgraded by the structured internal validation battery. GLOBEM candidates are included as low-signal exploratory or unstable associations and should not be interpreted as battery-passing validated candidates. One rejected candidate (TG/HDL ratio) is included for transparency as a sanity check demonstrating the pipeline's construct independence gate. Effect sizes are Spearman correlations (ρ). Adjusted p values reflect Benjamini–Hochberg FDR correction ($\alpha = 0.05$). Cross-validation metrics are reported for models trained on the candidate biomarker sets. 95% confidence intervals are obtained by bootstrap resampling (1,000 replicates) for the DWB and WEAR-ME cohorts. For GLOBEM, where repeated within-participant observations preclude simple row resampling, intervals are analytic (Fisher transformation).

Cohort	Biomarker	Effect Size (ρ)	95% CI	Adj. p	Mechanistic Hypothesis	Prior Evidence
DWB Hourly (N = 7,497; target: PHQ-8; Ridge CV R^2 = 0.228[‡])	Main sleep duration variability	0.252	[0.23, 0.27]	< 0.001	Circadian instability impairs sleep homeostatic drive and emotional regulation	Established
	Nocturnal social app usage	0.246	[0.22, 0.27]	< 0.001	Blue-light exposure and hyperarousal suppress melatonin secretion	Established
	Late-night doom-scrolling	0.177	[0.16, 0.20]	< 0.001	Nocturnal news/social scrolling sustains rumination and cortisol release	Supported*
	Night-to-day social media ratio	0.222	[0.20, 0.24]	< 0.001	Displaced nocturnal social engagement reflects rumination and insomnia	Supported*
	Hedonic-to-productivity app ratio	0.152	[0.13, 0.17]	< 0.001	Behavioral narrowing may increase hedonic relative to productivity-oriented app use in participants with higher depressive symptom scores	Emerging**
	Polyphasic sleep percentage	0.193	[0.17, 0.21]	< 0.001	Fragmented nocturnal architecture reflects HPA axis dysregulation	Emerging**
GLOBEM (N = 704; target: PHQ-4; CV AUC = 0.535)	Sleep onset time variability (circadian acrophase)	0.126	[0.05, 0.20]	< 0.001	Irregular sleep scheduling disrupts circadian entrainment	Established
	Evening incoming call duration (circadian acrophase)	−0.145	[−0.22, −0.07]	< 0.001	Reduced evening social communication reflects social withdrawal	Flagged unstable
	WiFi AP sequential diversity (7-day)	0.128	[0.05, 0.20]	< 0.001	Dynamic network scanning as proxy for environmental instability	Emerging**
WEAR-ME (N = 1,078; target: HOMA-IR; Ridge CV R^2 = 0.389[§])	Derived TG/HDL ratio[R]	0.542	[0.50, 0.58]	< 0.001	Atherogenic dyslipidemia directly indexes hepatic insulin resistance	Established
	HDL cholesterol	−0.380	[−0.43, −0.33]	< 0.001	Low HDL reflects impaired reverse cholesterol transport in metabolic syndrome	Established
	C-reactive protein (CRP)	0.393	[0.34, 0.44]	< 0.001	Systemic inflammation mediates adipose-derived insulin resistance via TNF-α/IL-6	Established
	Derived AST/ALT ratio (De Ritis)	−0.375	[−0.42, −0.32]	< 0.001	Hepatic gluconeogenic stress and subclinical steatosis marker	Emerging**
	Resting heart rate (mean)[†]	0.348	[0.30, 0.40]	< 0.001	Autonomic imbalance: sympathetic overdrive increases hepatic glucose output	Established
	Cardiovascular fitness index (steps/resting HR)[†]	−0.374	[−0.42, −0.32]	< 0.001	May proxy cardiorespiratory fitness, habitual activity and autonomic tone, pathways linked to insulin sensitivity	Supported*
	Red cell distribution width (RDW)	0.281	[0.22, 0.33]	< 0.001	Erythropoietic stress marker of chronic low-grade metabolic inflammation	Emerging**
	Albumin/globulin ratio	−0.220	[−0.29, −0.17]	< 0.001	Hepatic synthetic dysfunction and subclinical inflammatory protein shift	Emerging**

Established = consistent with known clinical associations with substantial literature support; **Supported*** = the underlying physiological axis is established but this specific operationalization from wearable or digital phenotyping data is new; **Emerging**** = limited prior evidence for this specific feature–endpoint association within the searched literature corpus; requires held-out confirmation. [R]Rejected by the construct-overlap and clinical-redundancy gate: triglycerides and HDL are routine lipid measurements already used in metabolic risk assessment, so the ratio was treated as a positive control for distinguishing statistical validity from incremental clinical value. TG/HDL is not an algebraic component of HOMA-IR and was not counted among non-rejected candidates. [†]Wearable-derived feature. All other WEAR-ME features are derived from fasting clinical laboratory panels. [‡]Ridge regression CV R^2 using demographics plus top CoDaS-selected biomarkers; 5-fold nested cross-validation. [§]Ridge regression CV R^2 using demographics plus top CoDaS-selected biomarkers; 5-fold nested cross-validation. [¶]Discovery and held-out estimates differed in sign for this feature.

Table 4 | **Exploratory component ablation under fixed evaluation settings.** CV R^2 (Demographics + Biomarkers model) for DWB and WEAR-ME, and CV AUC (Logistic Regression, Demographics + Biomarkers) for GLOBEM (binary depression classification, PHQ-4 > 2), across all three datasets for each ablation condition. Higher is better except where leakage is indicated (†). N Val. = number of candidates passing the validation battery in each ablation run; these counts reflect each run's own discovery output and may differ from the curated candidates reported in the appendix tables (Tables 8–10), which apply the stricter 11-test pipeline to the canonical full-pipeline run.

	DWB		GLOBEM		WEAR-ME	
Configuration	**CV R^2**	**N Val.**	**CV AUC**	**N Val.**	**CV R^2**	**N Val.**
Full CoDaS	0.228	35	0.694	48	0.389	49
– Adversarial debate	0.228	46	0.694	48	0.407	42
– Iterative loop	0.217	35	0.653	4	0.387	50
– Literature grounding	0.207	49	0.694	50	0.407	43
– Scout agent	0.120	48	0.707	48	0.365	50
– Reinvestigation	0.228	48	0.694	48	0.407	45
– Validation procedure	0.228†	—	0.694	—	0.407†	—
Demographics only	0.188	—	0.588	—	0.260	—

†Without the validation pipeline, leakage-inflated features may be present (see text); the CV R^2 is not directly comparable. — indicates the metric is not applicable. The GLOBEM AUCs in this table are post-hoc component-ablation metrics using a top-five logistic-regression protocol. They should not be compared directly with the primary GLOBEM performance estimate in Section 5.1, which used the all-screened-feature evaluation protocol. Ablation counts reflect pre-dedup / intermediate candidate set under the ablation evaluation protocol.

Table 5 | **Safety and reliability concerns identified by expert reviewers.** Count of safety flags per system across five categories. CoDaS received 3 flags (1 StatError, 1 Hallucination, 1 Other); baseline systems accumulated 51 flags total (42 from Data Science Agent and Biomni, 9 from AI co-scientist). The most severe failure modes, including hallucinated results, sub-random classification performance presented as findings, and label leakage invalidating primary claims, were observed exclusively in baseline systems.

	Hallucination (↓)	StatError (↓)	Logic (↓)	BioContradiction (↓)	Other (↓)
CoDaS (**Ours**)	1	1	0	0	1
Data Science Agent	1	5	5	4	1
AI Co-scientist	5	2	1	0	1
Biomni	7	6	7	2	4

Table 6 | **Clinician panel ratings for each candidate biomarker.** Twelve physicians rated every candidate on a five-point scale. Validity is the mean across reviewers with the standard deviation in parentheses. Overall is the mean across the seven rated dimensions. Across-rater standard deviation is the standard deviation among reviewers averaged over the seven dimensions, where a lower value indicates closer agreement. The rejected TG/HDL ratio is included as a positive control.

Cohort	Candidate	Validity	Overall	Across-rater s.d.
DWB (PHQ-8)	Main sleep duration variability	4.17 (1.11)	3.15	0.69
	Nocturnal social app usage	3.92 (1.08)	3.06	0.94
	Late-night doomscrolling	3.50 (1.00)	2.73	0.99
	Night-to-day social media ratio	3.67 (0.89)	2.70	0.85
	Hedonic-to-productivity app ratio	3.17 (1.03)	2.31	0.84
	Polyphasic sleep percentage	3.67 (0.78)	2.69	0.86
GLOBEM (PHQ-4)	Sleep onset time variability	3.83 (0.58)	2.55	0.74
	Evening incoming call duration	1.75 (1.14)	1.80	0.79
	WiFi AP sequential diversity	2.58 (0.90)	2.14	0.70
WEAR-ME (HOMA-IR)	HDL cholesterol	4.75 (0.45)	2.96	0.75
	C-reactive protein (CRP)	4.33 (0.49)	3.06	0.80
	Derived AST/ALT ratio (De Ritis)	4.00 (0.60)	3.26	0.84
	Resting heart rate (mean)	4.08 (1.08)	3.20	1.01
	Cardiovascular fitness index (steps/resting HR)	4.25 (0.45)	3.55	0.69
	Red cell distribution width (RDW)	2.92 (1.24)	2.52	0.84
	Albumin/globulin ratio	3.33 (0.78)	2.69	0.64
	Derived TG/HDL ratio (rejected control)	4.83 (0.39)	2.88	0.69

8. Limitations

Exploratory, non-preregistered design. All analyses reported in this study are exploratory. No endpoints, biomarker candidates, subgroup analyses, or analysis strategies were registered in a public trial registry prior to pipeline execution. The GLOBEM endpoint (PHQ-4) was selected by the pipeline itself based on target coverage rather than pre-specified by investigators. Consequently, all reported candidates should be interpreted as statistically prioritized hypotheses, not validated biomarkers in the regulatory sense. The term "screened" as used throughout refers exclusively to survival of the internal validation battery and does not imply prospective clinical validation, external replication, or regulatory endorsement.

GLOBEM repeated-measures structure and missingness. The GLOBEM cohort comprises 704 participant-wave observations from 497 unique individuals; some participants contributed data across multiple annual waves. Although the holdout confirmation split and validation tests operate on one randomly selected observation per participant to ensure statistical independence (see Section 2.6), the main analytical pipeline processes all 704 observations. Feature-level missingness in GLOBEM is substantial (54.6%), reflecting the inherent sparsity of passively collected mobile sensing data. While features with >70% missingness were dropped and remaining values were median-imputed, the impact of this imputation strategy on discovered associations has not been exhaustively characterized. Results from this cohort should be interpreted with these caveats.

Static labels and narrow disease scope. The DWB and WEAR-ME analyses rely on fixed or cross-sectional endpoints, while GLOBEM provides repeated survey measurements but with sparse sensing and coarse PHQ-4 labels. The mental-health analyses rely on self-report symptom scales, whereas the metabolic analysis is anchored to fasting laboratory measures. Cohort demographics further limit generalizability: participants skew White, female, and young to middle-aged in DWB and WEAR-ME, and GLOBEM is an undergraduate cohort (DWB: 84.9% White, 70.0% female; GLOBEM: 57.4% Asian, undergraduate only; WEAR-ME: 77.7% White/Caucasian in the source study). Prospective studies with repeated clinical labels, ecological momentary assessment, broader disease coverage, and more representative recruitment will be required before these candidates can be assessed for clinical utility across populations.

Associative framing and causal inference. CoDaS surfaces statistically robust associations and generates post-hoc mechanistic hypotheses grounded in retrieved literature, but does not establish causality. Statistical significance and robustness across subgroups or independent cohorts are insufficient to distinguish causal biomarkers from correlated proxies especially because only measured confounders could be adjusted for. Furthermore, the short retrospective monitoring windows are not sufficient for causal discovery of biomarkers. Consequently, candidate biomarkers identified by CoDaS should be regarded as hypothesis-generating until supported by physiological evidence, prospective validation, and rigorous causal inference frameworks. Moving forward, these capabilities

could be further integrated into the adversarial validation process to systematically identify potential confounding factors, spurious correlations, and limitations in study design. In addition, incorporating target-trial emulation, structural causal models, or quasi-experimental designs into the validation gauntlet is a necessary step toward the evidentiary standards required for clinical decision support.

External validation and replication. No single biomarker was replicated in an independent cohort with an aligned outcome instrument. The two depression cohorts use different data structures (hourly wearable metrics vs. weekly passive-sensing aggregates) and different outcome instruments (PHQ-8 vs. PHQ-4), limiting comparability beyond construct-level convergence (see Section 5.3). The WEAR-ME cohort addresses a separate disease domain (metabolic risk), demonstrating pipeline breadth but not direct cross-cohort replication. This evaluation design does not fulfill the external-validation requirements of TRIPOD or STARD reporting guidelines. Prospective replication of the top candidate biomarkers in an independent depression cohort using PHQ-8 as the primary endpoint, ideally with a temporally held-out validation set, is the clear next step for translational readiness.

Fixed agent topology and extensibility. The CoDaS agent graph, including the number of specialist agents, their functional mandates, inter agent communication protocols, and phase transition predicates, is defined a priori by system designers rather than inferred from the analytical task. Adapting the framework to new sensor modalities, data sparse disease domains, or datasets with different temporal granularities currently requires nontrivial manual re engineering of agent specifications and prompts, introducing overhead that attenuates the throughput benefits of full automation. Future architectures should explore meta-agent strategies in which a higher-order controller dynamically instantiates and wires sub-agents from a declarative task specification, enabling principled scaling to heterogeneous discovery settings, as explored in recent work on scaling multi-agent systems (Kim et al., 2025a).

Mechanistic hypothesis generation. The mechanistic hypotheses reported in Table 3 (e.g., "circadian instability impairs sleep homeostatic drive") are generated by the foundation model and grounded in retrieved literature via the BibTeX verification layer. However, LLM-generated mechanistic narratives may appear authoritative while constituting post-hoc rationalizations rather than experimentally validated causal pathways. All mechanistic claims should be treated as hypothesis-generating and require independent experimental confirmation.

Translational readiness and model dependence. The internal 11-step validation framework imposes a stricter internal screening procedure than the baselines evaluated in this study, yet it does not fulfill the sequential analytical validation, clinical validation, and randomized interventional requirements mandated by regulatory frameworks such as the FDA Biomarkers, Endpoints, and other Tools (BEST) framework. Discovered effect sizes, while statistically consistent across subgroups, are modest in absolute magnitude (e.g., $\rho = 0.252$ for sleep variability and depression severity), and incremental value over established screening instruments has not been established through head to

head comparison. CoDaS currently depends on a fixed foundation-model stack (Gemini-3.1 Pro for research-intensive tasks and Gemini-3 Flash for repeated tasks); although the deterministic BibTeX verification layer reduces hallucination risk, all retrieved references require independent confirmation, and reasoning quality is bounded by the model's pretraining distribution. These considerations position CoDaS as a high throughput hypothesis generation and prioritization platform rather than a regulatory grade diagnostic system, and define a clear translational research program to accompany further development.

Age, multimorbidity and medication generalizability. The candidate biomarkers were discovered primarily in young-to-middle-aged cohorts rather than in the older, multimorbid populations most likely to benefit from continuous monitoring between visits. Sleep fragmentation, polyphasic sleep, reduced mobility, resting heart rate and laboratory markers such as AST/ALT ratio and RDW can reflect aging, frailty, sarcopenia, medication use, anemia, osteoarthritis or gait limitation rather than the target phenotype alone. For instance, sleep-related depression candidates may partly overlap with somatic PHQ items, and resting heart rate may be strongly affected by beta-blockers. The wearable-derived fitness index should therefore be interpreted as a cohort-specific candidate requiring medication-aware, mobility-aware and age-stratified prospective validation before generalization to older clinical populations.

9. Conclusion

We present CoDaS, a multi-agent system that coordinates data profiling, literature-grounded hypothesis generation, parallel empirical exploration, adversarial validation, and mechanistic reasoning with human supervision to generate and prioritize digital biomarker candidates from population-scale wearable sensor data. Across 9,279 participant-observations spanning mental health and metabolic phenotypes, CoDaS identified composite physiological signatures, each surviving a structured internal validation battery spanning four dimensions and 11 checks. These remain candidate associations that are hypothesis-generating and require prospective clinical validation before use in patient care. Three architectural principles were important: separating deterministic computation from generative reasoning to improve reproducibility, using adversarial critic-defender evaluation to reduce tautological leakage, and maintaining human oversight at critical stages. Our findings suggest that principled integration of exploratory breadth with scientific rigor is a central bottleneck in digital biomarker science, and that agent systems with human supervision can structure and accelerate candidate prioritization as wearable health data continues to scale across disease domains and populations.

Data availability. GLOBEM data are available from the original GLOBEM release under the terms specified by the dataset providers. WEAR-ME data are available to approved researchers through the access process described in the original WEAR-ME study. DWB participant-level data are not publicly available and cannot be redistributed by the authors because they contain sensitive health, wearable, smartphone and survey data subject to participant-consent and institutional data-use restrictions. De-identified source data underlying the main figures and tables, the clinician-review instrument,

anonymized clinician ratings and benchmark output summaries will be made available to editors and reviewers on request upon publication where permitted by consent and data-use agreements.

Code availability. The code for pre-processing, feature construction, validation checks, model evaluation and figure generation will be available from the corresponding authors on reasonable request, subject to institutional, third-party and company restrictions. Proprietary model-API integrations and internal infrastructure code cannot be publicly redistributed and are described in the Methods, executable pseudocode, audit logs and benchmark-output summaries.

AI-use disclosure. CoDaS is the subject of this study. CoDaS-generated reports were treated as experimental outputs and were not accepted as scientific claims without author review, verification and additional analysis. The final manuscript was written and approved by the human authors, who take responsibility for its content. No AI system is listed as an author.

References

A. K. M. Alonso, J. Hirt, T. Woelfle, P. Janiaud, and L. G. Hemkens. Definitions of digital biomarkers: a systematic mapping of the biomedical literature. *BMJ health & care informatics*, 31(1):e100914, 2024.

R. K. Arora, J. Wei, R. S. Hicks, P. Bowman, J. Quiñonero-Candela, F. Tsimpourlas, M. Sharman, M. Shah, A. Vallone, A. Beutel, et al. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

L. M. Babrak, J. Menetski, M. Rebhan, G. Nisato, M. Zinggeler, N. Brasier, K. Baerenfaller, T. Brenzikofer, L. Baltzer, C. Vogler, et al. Traditional and digital biomarkers: two worlds apart? *Digital biomarkers*, 3(2):92–102, 2019.

M. Barbieri, E. Ragno, E. Benvenuti, G. Zito, A. Corsi, L. Ferrucci, and G. Paolisso. New aspects of the insulin resistance syndrome: impact on haematological parameters. *Diabetologia*, 44(10): 1232–1237, 2001.

D. A. Boiko, R. MacKnight, B. Kline, and G. Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

N. Brasier, J. Wang, W. Gao, J. R. Sempionatto, C. Dincer, H. C. Ates, F. Güder, S. Olenik, I. Schauwecker, D. Schaffarczyk, et al. Applied body-fluid analysis by wearable devices. *Nature*, 636(8041):57–68, 2024.

D. Cella, S. Yount, N. Rothrock, R. Gershon, K. Cook, B. Reeve, D. Ader, J. F. Fries, B. Bruce, M. Rose, et al. The patient-reported outcomes measurement information system (promis): progress of an nih roadmap cooperative group during its first two years. *Medical care*, 45(5):S3–S11, 2007.

B. P. Chapman, B. T. Gullapalli, T. Rahman, D. Smelson, E. W. Boyer, and S. Carreiro. Impact of individual and treatment characteristics on wearable sensor-based digital biomarkers of opioid use. *npj Digital Medicine*, 5(1):123, 2022.

M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. D. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

J. Clusmann, F. R. Kolbinger, H. S. Muti, Z. I. Carrero, J.-N. Eckardt, N. G. Laleh, C. M. L. Löffler, S.-C. Schwarzkopf, M. Unger, G. P. Veldhuizen, et al. The future landscape of large language models in medicine. *Communications medicine*, 3(1):141, 2023.

S. Cohen, T. Kamarck, and R. Mermelstein. A global measure of perceived stress. *Journal of Health and Social Behavior*, 24(4):385–396, 1983.

A. Coravos, S. Khozin, and K. D. Mandl. Developing and adopting safe and effective digital biomarkers to improve patient outcomes. *npj Digital Medicine*, 2(14), 2019.

C. Cornelio, T. Ito, R. Cory-Wright, S. Dash, and L. Horesh. The need for verification in ai-driven scientific discovery. *arXiv preprint arXiv:2509.01398*, 2025.

J. Cosentino, A. Belyaeva, X. Liu, N. A. Furlotte, Z. Yang, C. Lee, E. Schenck, Y. Patel, J. Cui, L. D. Schneider, et al. Towards a personal health large language model. *arXiv preprint arXiv:2406.06474*, 2024.

P. Dagum. Digital biomarkers of cognitive function. *NPJ digital medicine*, 1(1):10, 2018.

P. Daniore, V. Nittas, C. Haag, J. Bernard, R. Gonzenbach, and V. von Wyl. From wearable sensor data to digital biomarker development: ten lessons learned and a framework proposal. *NPJ digital medicine*, 7(1):161, 2024.

J. Dunn, L. Kidzinski, R. Runge, D. Witt, J. L. Hicks, S. M. Schüssler-Fiorenza Rose, X. Li, A. Bahmani, S. L. Delp, T. Hastie, et al. Wearable sensors enable personalized predictions of clinical laboratory measurements. *Nature medicine*, 27(6):1105–1112, 2021.

G. Fagherazzi, C. Goetzinger, M. A. Rashid, G. A. Aguayo, and L. Huiart. Digital health strategies to fight covid-19 worldwide: challenges, recommendations, and a call for papers. *Journal of medical internet research*, 22(6):e19284, 2020.

Y. Fang, D. B. Forger, E. Frank, S. Sen, and C. Goldstein. Day-to-day variability in sleep parameters and depression risk: a prospective cohort study of training physicians. *NPJ digital medicine*, 4(1): 28, 2021.

S. Gao, A. Fang, Y. Huang, V. Giunchiglia, A. Noori, J. R. Schwarz, Y. Ektefaie, J. Kondic, and M. Zitnik. Empowering biomedical discovery with ai agents. *Cell*, 187(22):6125–6151, 2024.

J. C. Goldsack, A. Coravos, J. P. Bakker, B. Bent, A. V. Dowling, C. Fitzer-Attas, A. Godfrey, J. G. Godino, N. Gujar, E. Izmailova, et al. Verification, analytical validation, and clinical validation (v3): the foundation of determining fit-for-purpose for biometric monitoring technologies (biomets). *npj digital Medicine*, 3(1):55, 2020.

J. Gottweis, W.-H. Weng, A. Daryin, T. Tu, A. Palepu, P. Sirkovic, A. Myaskovsky, F. Weissenberger, K. Rong, R. Tanno, et al. Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.

A. A. Heydari, K. Gu, V. Srinivas, H. Yu, Z. Zhang, Y. Zhang, A. Paruchuri, Q. He, H. Palangi, N. Hammerquist, et al. The anatomy of a personal health agent. *arXiv preprint arXiv:2508.20148*, 2025.

X. Hu, Z. Zhao, S. Wei, Z. Chai, Q. Ma, G. Wang, X. Wang, J. Su, J. Xu, M. Zhu, et al. Infiagent-dabench: Evaluating agents on data analysis tasks. *arXiv preprint arXiv:2401.05507*, 2024.

K. Huang, S. Zhang, H. Wang, Y. Qu, Y. Lu, Y. Roohani, R. Li, L. Qiu, G. Li, J. Zhang, et al. Biomni: A general-purpose biomedical ai agent. *biorxiv*, 2025.

S. Jiang, L.-b. Sweet, G. Blougouras, A. Brenning, W. Li, M. Reichstein, J. Denzler, W. Shangguan, G. Yu, F. Huang, et al. How interpretable machine learning can benefit process understanding in the geosciences. *Earth's Future*, 12(7):e2024EF004540, 2024.

D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

L. Jing, Z. Huang, X. Wang, W. Yao, W. Yu, K. Ma, H. Zhang, X. Du, and D. Yu. Dsbench: How far are data science agents from becoming data science experts? *arXiv preprint arXiv:2409.07703*, 2024.

Y. Kim, X. Xu, D. McDuff, C. Breazeal, and H. W. Park. Health-llm: Large language models for health prediction via wearable sensor data. *arXiv preprint arXiv:2401.06866*, 2024.

Y. Kim, K. Gu, C. Park, C. Park, S. Schmidgall, A. A. Heydari, Y. Yan, Z. Zhang, Y. Zhuang, M. Malhotra, et al. Towards a science of scaling agent systems. *arXiv preprint arXiv:2512.08296*, 2025a.

Y. Kim, H. Jeong, S. Chen, S. S. Li, C. Park, M. Lu, K. Alhamoud, J. Mun, C. Grau, M. Jung, et al. Medical hallucinations in foundation models and their impact on healthcare. *arXiv preprint arXiv:2503.05777*, 2025b.

Y. Kim, H. Jeong, C. Park, E. Park, H. Zhang, X. Liu, H. Lee, D. McDuff, M. Ghassemi, C. Breazeal, et al. Tiered agentic oversight: A hierarchical multi-agent system for ai safety in healthcare. *arXiv e-prints*, pages arXiv–2506, 2025c.

Y. Kim, T. Kim, W. Kang, E. Park, J. Yoon, D. Lee, X. Liu, D. McDuff, H. Lee, C. Breazeal, et al. Vocalagent: Large language models for vocal health diagnostics with safety-aware evaluation. *arXiv preprint arXiv:2505.13577*, 2025d.

K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. W. Williams, K. Berry, and A. H. Mokdad. The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, 114: 163–173, 2009. doi: 10.1016/j.jad.2008.06.026.

M. Kwon, J.-Y. Lee, W.-Y. Won, J.-W. Park, J.-A. Min, C. Hahn, X. Gu, J.-H. Choi, and D.-J. Kim. Development and validation of a smartphone addiction scale (sas). *PloS one*, 8(2):e56936, 2013.

G. Li, H. Hammoud, H. Itani, D. Khizbullin, and B. Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36: 51991–52008, 2023.

S. Li, S. Xiao, M. Joshi, A. Metwally, D. McDuff, W. Wang, and Y. Yang. Hearts: Benchmarking llm reasoning on health time series. *arXiv preprint arXiv:2603.06638*, 2026.

X. Liu, Z. Wu, X. Wu, P. Lu, K.-W. Chang, and Y. Feng. Are llms capable of data-based statistical and causal reasoning? benchmarking advanced quantitative reasoning with data. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9215–9235, 2024.

B. Löwe, O. Decker, S. Müller, E. Bräehler, D. Schellberg, W. Herzog, and P. Y. Herzberg. Validation and standardization of the generalized anxiety disorder screener (GAD-7) in the general population. *Medical Care*, 46(3):266–274, 2008. doi: 10.1097/MLR.0b013e318160d093.

C. Lu, C. Lu, R. T. Lange, J. Foerster, J. Clune, and D. Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

C. Lu, C. Lu, R. T. Lange, Y. Yamada, S. Hu, J. Foerster, D. Ha, and J. Clune. Towards end-to-end automation of ai research. *Nature*, 651(8107):914–919, 2026.

Z. Luo, A. Kasirzadeh, and N. B. Shah. The more you automate, the less you see: Hidden pitfalls of ai scientist systems. *arXiv preprint arXiv:2509.08713*, 2025.

B. P. Majumder, H. Surana, D. Agarwal, B. D. Mishra, A. Meena, A. Prakhar, T. Vora, T. Khot, A. Sabharwal, and P. Clark. Discoverybench: Towards data-driven discovery with large language models. *arXiv preprint arXiv:2407.01725*, 2024.

D. McDuff, A. Barakat, A. Winbush, A. Jiang, F. Cordeiro, R. Crowley, L. E. Kahn, J. Hernandez, and N. B. Allen. Research protocol for the google health digital well-being study. *arXiv preprint arXiv:2307.05795*, 2023.

D. McDuff, A. Barakat, A. Winbush, A. Jiang, F. Cordeiro, R. Crowley, L. E. Kahn, J. Hernandez, and N. B. Allen. The google health digital well-being study: Protocol for a digital device use and well-being study. *JMIR Research Protocols*, 13(1):e49189, 2024.

A. A. Metwally, A. A. Heydari, D. McDuff, A. Solot, Z. Esmaeilpour, A. Z. Faranesh, M. Zhou, G. Narayanswamy, M. A. Xu, X. Liu, et al. Insulin resistance prediction from wearables and routine blood biomarkers. *Nature*, 652(8109):451–461, 2026.

F. Nie, J. Wang, H. Hua, F. Bianchi, Y. Kwon, Z. Qi, O. Queen, S. Zhu, and J. Zou. Dsgym: A holistic framework for evaluating and training data science agents. *arXiv preprint arXiv:2601.16344*, 2026.

H. Nori, Y. T. Lee, S. Zhang, D. Carignan, R. Edgar, N. Fusi, N. King, J. Larson, Y. Li, W. Liu, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023.

A. Novikov, N. Vũ, M. Eisenberger, E. Dupont, P.-S. Huang, A. Z. Wagner, S. Shirobokov, B. Kozlovskii, F. J. Ruiz, A. Mehrabian, et al. Alphaevolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

M. C. Petersen and G. I. Shulman. Mechanisms of insulin action and insulin resistance. *Physiological reviews*, 98(4):2133–2223, 2018.

D. Powell. Walk, talk, think, see and feel: harnessing the power of digital biomarkers in healthcare. *NPJ Digital Medicine*, 7(1):45, 2024.

O. Rotem and A. Zaritsky. Visual interpretability of bioimaging deep learning models. *Nature Methods*, 21(8):1394–1397, 2024.

S. Saeb, M. Zhang, C. J. Karr, S. M. Schueller, M. E. Corden, K. P. Kording, and D. C. Mohr. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *Journal of medical Internet research*, 17(7):e4273, 2015.

G. L. Salvagno, F. Sanchis-Gomar, A. Picanza, and G. Lippi. Red blood cell distribution width: a simple parameter with multiple clinical applications. *Critical reviews in clinical laboratory sciences*, 52(2): 86–105, 2015.

T. Savage, A. Nayak, R. Gallo, E. Rangan, and J. H. Chen. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digital Medicine*, 7(1):20, 2024.

K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, M. Amin, L. Hou, K. Clark, S. R. Pfohl, H. Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, 2025.

K. Swanson, W. Wu, N. L. Bulaong, J. E. Pak, and J. Zou. The virtual lab of ai agents designs new sars-cov-2 nanobodies. *Nature*, 646(8085):716–723, 2025.

X. Tang, Q. Jin, K. Zhu, T. Yuan, Y. Zhang, W. Zhou, M. Qu, Y. Zhao, J. Tang, Z. Zhang, et al. Risks of ai scientists: prioritizing safeguarding over autonomy. *Nature Communications*, 16(1):8317, 2025.

S. Tong, K. Mao, Z. Huang, Y. Zhao, and K. Peng. Automating psychological hypothesis generation with ai: when large language models meet causal graph. *Humanities and Social Sciences Communications*, 11(1):1–14, 2024.

E. J. Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.

T. Tu, S. Azizi, D. Driess, M. Schaekermann, M. Amin, P.-C. Chang, A. Carroll, C. Lau, R. Tanno, I. Ktena, et al. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138, 2024.

S. Vasudevan, A. Saha, M. E. Tarver, and B. Patel. Digital biomarkers: convergence of digital health technologies and biomarkers. *NPJ digital medicine*, 5(1):36, 2022.

H. Wu, B. Zheng, D. Song, Y. Jiang, J. Gao, L. Xing, L. Sun, and Y. Yuan. Towards a medical ai scientist. *arXiv preprint arXiv:2603.28589*, 2026.

Y. Xiao, J. Liu, Y. Zheng, X. Xie, J. Hao, M. Li, R. Wang, F. Ni, Y. Li, J. Luo, et al. Cellagent: An llm-driven multi-agent framework for automated single-cell data analysis. *arXiv preprint arXiv:2407.09811*, 2024.

X. Xu, H. Zhang, Y. Sefidgar, Y. Ren, X. Liu, W. Seo, J. Brown, K. Kuehn, M. Merrill, P. Nurius, et al. Globem dataset: multi-year datasets for longitudinal human behavior modeling generalization. *Advances in neural information processing systems*, 35:24655–24692, 2022.

Y. Yamada, R. T. Lange, C. Lu, S. Hu, C. Lu, J. Foerster, J. Clune, and D. Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.

Y. Yang, Y. Yuan, G. Zhang, H. Wang, Y.-C. Chen, Y. Liu, C. G. Tarolli, D. Crepeau, J. Bukartyk, M. R. Junna, et al. Artificial intelligence-enabled detection and assessment of parkinson's disease using nocturnal breathing signals. *Nature medicine*, 28(10):2207–2215, 2022.

J. Young. Effective harnesses for long-running agents. Anthropic Engineering Blog, Nov. 2025. URL https://www.anthropic.com/engineering/effective-harnesses-for-long-running-agents.

L. Yu, D. J. Buysse, A. Germain, D. E. Moul, A. Stover, N. E. Dodds, K. L. Johnston, and P. A. Pilkonis. Development of short forms from the promis® sleep disturbance and sleep-related impairment item banks. *Behavioral Sleep Medicine*, 10(1):6–24, 2012. doi: 10.1080/15402002.2012.650148.

D. Zhang, S. Zhoubian, M. Cai, F. Li, L. Yang, W. Wang, T. Dong, Z. Hu, J. Tang, and Y. Yue. Datascibench: An llm agent benchmark for data science. *arXiv preprint arXiv:2502.13897*, 2025.

T. Zheng, Z. Deng, H. T. Tsang, W. Wang, J. Bai, Z. Wang, and Y. Song. From automation to autonomy: A survey on large language models in scientific discovery. *arXiv preprint arXiv:2505.13259*, 2025.

Y. Zhou, H. Liu, T. Srivastava, H. Mei, and C. Tan. Hypothesis generation with large language models. *arXiv preprint arXiv:2404.04326*, 2024.

K. Zhu, J. Zhang, Z. Qi, N. Shang, Z. Liu, P. Han, Y. Su, H. Yu, and J. You. Safescientist: Toward risk-aware scientific discoveries by llm agents. *arXiv preprint arXiv:2505.23559*, 2025.

Appendix

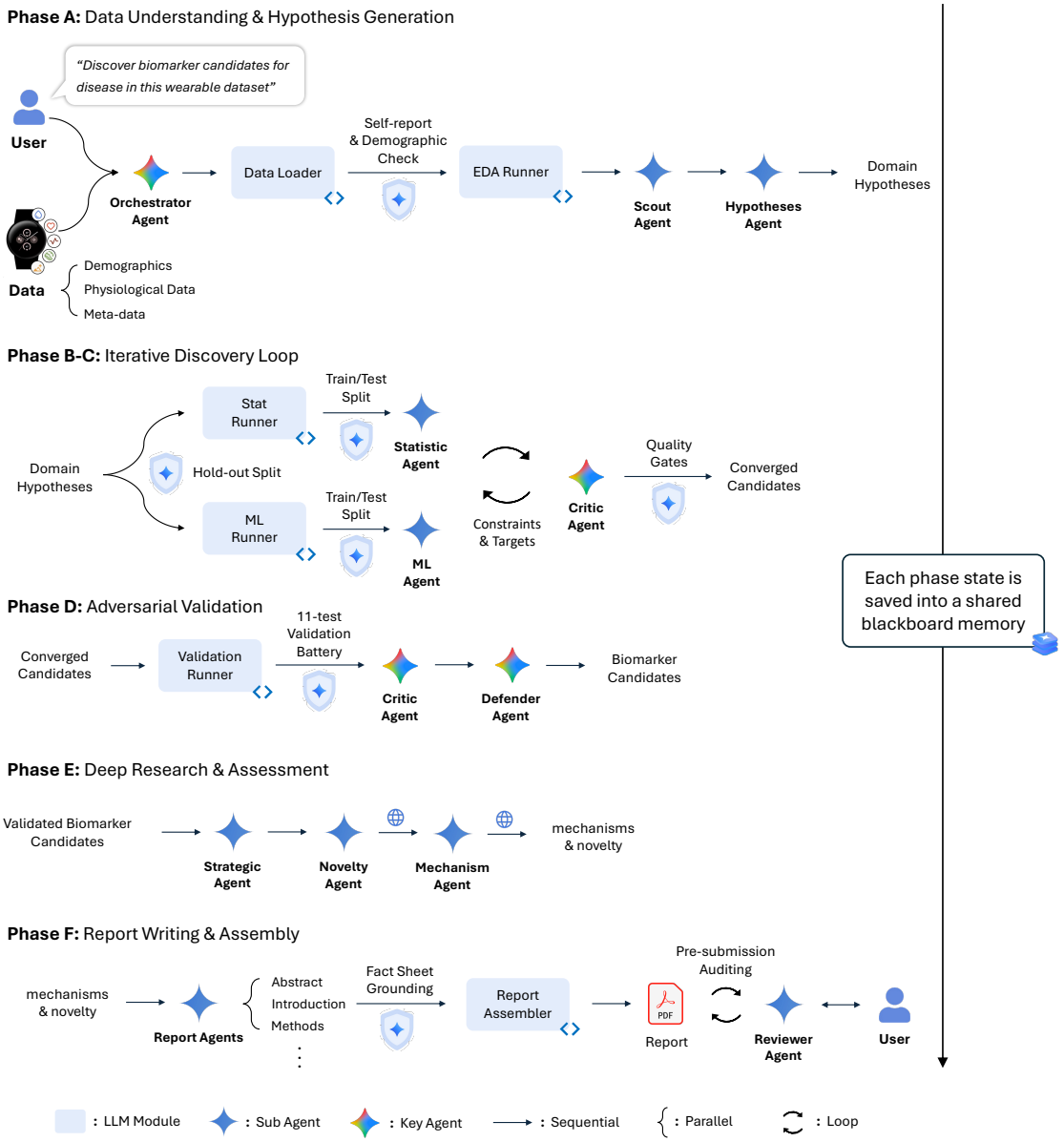


Figure 8 | **CoDaS Architecture. Phase A**: The Orchestrator receives a research question and raw data from the user, dispatching a Data Loader and EDA Runner to profile schema and statistics. A Scout Agent forms an analytical baseline, and a Hypotheses Agent generates literature-grounded domain hypotheses. **Phases B & C**: Hypotheses advance in parallel statistical and ML tracks. A Stat Runner and Agent performs univariate testing, while an ML Runner and Agent conducts cross-validated multivariate modeling. A Critic Agent enforces convergence through iteration. **Phase D**: Candidates are stress-tested by a Validation Runner; a Critic Interpreter flags artifacts and confounders, and a Defender Agent argues for retention. **Phase E**: Candidate biomarkers are evaluated by a Strategic Assessor, Novelty Classifier, and Mechanism Hypothesizer. **Phase F**: Report Agents draft sections in parallel, a Report Assembler integrates them, and a Reviewer Agent performs final quality control.

A. Phase-wise Runtime Comparison

To provide insight into how each system allocates computational effort across the biomarker discovery workflow, Table 7 compares phase-level runtimes for CoDaS and three baseline systems on the DWB dataset. Phases that a system does not perform are marked with "—".

Table 7 | **Phase-wise runtime comparison on DWB (N = 7,497).** Wall-clock time allocated to each phase of the biomarker discovery workflow. CoDaS is the only system that executes all six phases end-to-end. Phases not performed by a system are marked "—". All timings extracted from pipeline execution logs.

Pipeline Phase	CoDaS	AI co-scientist	Biomni	ADK DS Agent
1. Data Profiling & EDA	8.8 min	61.6 min*	~1 min	9 s
2. Literature Search & Synthesis	307.5 min†	34.6 min	Partial§	—
3. Hypothesis Generation		248.2 min‡	—	—
4. Statistical & ML Execution	89.5 min	—	~9 min	143 s
5. Adversarial Validation	17.4 min	—	—	—
6. Deep Research & Novelty	73.3 min	—	—	—
7. Report Writing & Review		75.0 min	~2 min	<1 s
Iterative Discovery Rounds	4	N/A	1	1
Total Wall-Clock Time	**8.28 h**	**7.00 h**	**11.6 min**	**2.76 min**
LLM-Guided Code Generation	Yes	No	Yes	No
LLM API Cost (est.)	$3.91	≥$2.79	<$0.05	<$0.01
Tokens (in / out)	7.2M / 196K	≥1.0M / 194K	99K / 9.2K	1.2K / 1.4K

*LLM-based analysis of data summaries, not deterministic code execution.

†CoDaS interleaves literature search with discovery: 307.5 min covers 3 literature-interpretation cycles (~65–121 min each) across 4 rounds; 89.5 min covers cumulative statistical/ML code execution.

‡AI co-scientist cover: data science learning (61.6 min) → topic exploration & content extraction (11.3 min) → knowledge base summarization (23.3 min) → idea generation & scoring (14.5 min) → tournament refinement (42.2 min) → deep verification (191.4 min; 339 ideas, 276 eligible, 57 verified) → finalization (56.1 min) → post-processing (18.9 min).

§Biomni performs a PubMed search (5 results) but does not use retrieved literature for hypothesis generation.

CoDaS: 4 discovery rounds with Jaccard-based convergence (converged at Round 3). AI co-scientist: continuous tournament, not discrete rounds. Biomni: single-pass agent conversation (28 LLM turns). ADK: linear 4-step pipeline without iteration. AI co-scientist token usage (≥1.0M/194K, ≥$2.79) is from its data science research module only; the Idea Forge tournament runs separately and its costs are not available. Biomni tokens (99K/9.2K) captured via LangChain callback instrumentation. ADK tokens (1.2K/1.4K) are used exclusively for report generation; the pipeline itself is deterministic Python. All timings verified against pipeline logs.

B. Complete Biomarker Candidate Lists

Tables 8–10 report the complete lists of battery-passing biomarker candidates discovered by CoDaS across all three cohorts. *Screened* candidates passed ≥70% of applicable tests including all core tests (replication, permutation, bootstrap, and CI consistency); *Conditionally prioritized* candidates passed ≥40% of applicable tests or were downgraded from screened status due to marginal effect sizes or borderline subgroup consistency (see Section 2.6 for threshold definitions). Effect sizes are Spearman correlations (ρ) with the clinical endpoint. Features are ordered by $|\rho|$ within each validation tier.

C. Held-Out Validation of Robustness Checks

Among CoDaS's validation tests, two target demographic robustness specifically: confounder control (partial Spearman correlation residualized on available demographics, retaining features with $p<0.05$) and subgroup consistency (gender-stratified Spearman, removing features with opposite-sign effects). Here, we evaluate whether applying these two checks on a training split produces candidate sets that score higher on robustness metrics computed independently on a held-out test split.

C.1. Experimental Design

Setup. For each of 10 random seeds, we split each dataset 70/30 (stratified by outcome quartile). The two checks are applied exclusively on the training split. Features failing either check are removed; the surviving set is evaluated on the held-out test split using four robustness metrics with 100-iteration bootstrap confidence intervals. No test-set information enters any analysis step.

To rule out the possibility that robustness gains arise simply from having fewer features, we include a *matched random-pruning control*: for each seed, we randomly drop features to the same count as the checked set using an independent random stream (seed+1000).

An ablation separates the contributions of each check: *confounder-only* (partial correlation without subgroup pruning) and *subgroup-only* (sign-flip pruning without confounder control). Ablation results are reported as exploratory (not FDR-corrected).

Datasets.

- **DWB** (N=7,497): depression severity (PHQ-8), 197 passive smartphone sensor features, 8 demographics.
- **WearMe** (N=1,078): insulin resistance (HOMA-IR), 71 wearable biometric features, 3 demographics.

We also tested GLOBEM (N=704, PHQ-4, non-sparse 28 features, 3 demographics); confounder survival and subgroup consistency were at floor (0.000) at baseline due to insufficient statistical power, yielding no change across conditions. GLOBEM is therefore omitted from the table and figure.

Metrics. Four held-out metrics, computed on the test split only: *confounder survival* (fraction of features with $p<0.05$ partial Spearman after residualizing demographics), *subgroup consistency* (fraction with same-sign ρ across gender subgroups), *replication rate* (fraction with $p<0.05$ Spearman), and *holdout* R^2 (Ridge regression, α=1.0, trained on training split, scored on test split). Two additional metrics (clinical AUC and effect generalization) showed no significant changes and are reported in the supplementary data files.

Note that confounder survival on the test set uses the same statistical procedure (partial Spearman) as the confounder check applied on the training set. The train/test split ensures no data leakage, but the metric and the check share a definitional basis: features that pass partial correlation on training data are expected to pass it more often on test data as well. Replication rate and subgroup consistency

provide partially independent evidence, as they are not direct targets of either check.

C.2. Results

Checked features are more robust on held-out data; randomly pruned features are not. Table 11 and Figure 18A show that applying both checks on the training split improves three of four metrics on DWB held-out data. Matched random pruning to the same feature count fails to improve any metric and reduces holdout R^2 (Δ=−0.112 ± 0.061). Direct paired comparison shows that checked features outperform randomly pruned features on confounder survival (Δ=+0.153, p<0.0001, d=3.3), subgroup consistency (Δ=+0.059, p=0.0002, d=1.9), and replication rate (Δ=+0.031, p=0.028, d=0.8). Benjamini–Hochberg FDR correction was applied across the 16 primary tests (4 metrics × 2 conditions × 2 datasets). All DWB and WearMe results with p<0.001 survive (q<0.001); DWB replication rate (q=0.001) and DWB holdout R^2 (q=0.004) also survive. All random-pruning tests remain non-significant after correction.

Results are directionally consistent on a second dataset (Figure 18B). WearMe, which differs in disease target (insulin resistance vs. depression), cohort, sample size, and feature modality, shows a consistent pattern: applying the same two checks on the training split yields candidate sets that exhibit improved robustness on held-out evaluation, including higher confounder survival +0.198 (q<0.001), subgroup consistency (+0.089, q<0.001), and replication rate +0.117 (q<0.001), all computed independently on the test split. Predictive performance remains unchanged (ΔR^2=−0.007, p=0.19), suggesting that robustness filtering primarily removes unstable associations without materially affecting the underlying signal, particularly in a relatively low-dimensional setting (15 features).

Ablation (exploratory, not FDR-corrected). Confounder control alone improves non-target metrics on DWB: subgroup consistency (+0.043, p=0.003), replication rate (+0.029, p=0.001), and holdout R^2 (+0.001, p=0.006). Adding subgroup pruning provides incremental gains on subgroup consistency (+0.019, p=0.001) and holdout R^2 (+0.001, p=0.022). Both checks contribute; confounder control accounts for the majority of the improvement.

Feature removal is consistent across splits. On DWB, 24 of ~50 pruned features are removed in all 10/10 random splits, including presleep phone-usage metrics, app-composition ratios, sleep-physiology indices, and mobility features.

C.3. Limitations

1. **Definitional overlap between check and metric.** Confounder survival on the test set measures the same statistical quantity (partial Spearman significance) that the confounder check enforces on the training set. The train/test split prevents data leakage, but an improvement on this metric is expected by the design of the check. Replication rate and subgroup consistency provide less circular evidence, as they are not direct targets of either check.

2. **Not iterative.** These checks are applied once. This experiment validates their effectiveness on held-out data, not an iterative improvement process.
3. **Two of many checks.** CoDaS's full validation pipeline includes permutation testing, bootstrap stability, CI consistency, and defender-critic debate. Only confounder control and subgroup consistency are evaluated here.
4. **WearMe R^2.** The small feature space (15 features) limits the room for pruning on WearMe, and holdout R^2 does not improve.
5. **GLOBEM uninformative.** With N=704 and 3 demographics, baseline robustness metrics are at floor, preventing any measurable effect.

D. Clinician Panel Reliability Detail

Table 12 reports the per-dimension reliability of the clinician panel evaluation in Section 7. The single-rater Krippendorff alpha and ICC(3,1) quantify the agreement of an individual clinician, while ICC(3,k) quantifies the reliability of the twelve-clinician mean.

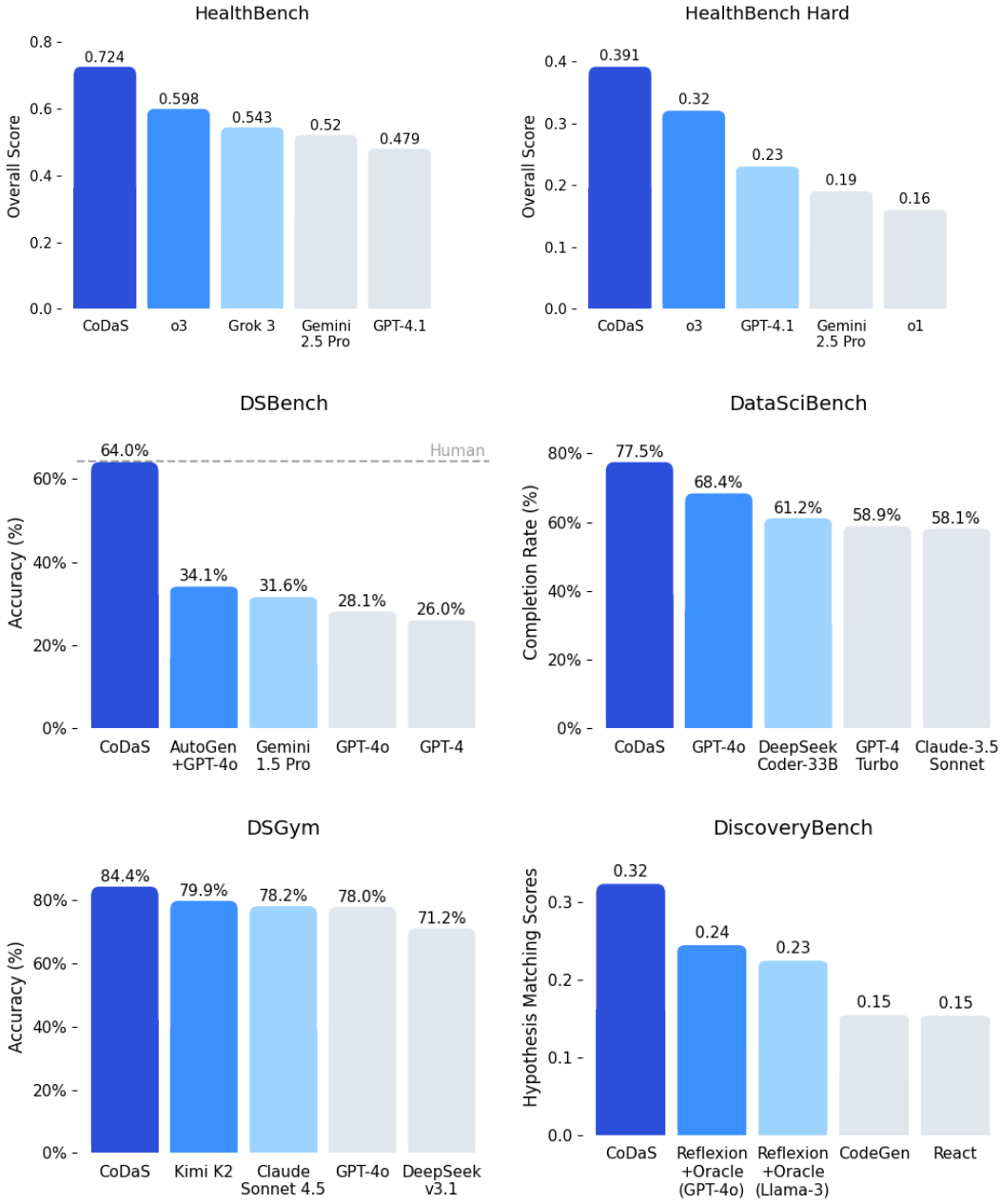


Figure 9 | **Benchmark evaluation of CoDaS and baselines.** CoDaS is compared against frontier LLMs and agent based frameworks on benchmarks that collectively evaluate the core capabilities required for autonomous biomarker discovery. Evaluated tasks include clinical reasoning over multiturn physician designed medical conversations (HealthBench, HealthBench Hard), real world data analysis over heterogeneous competition datasets (DSBench), end to end data science code generation spanning data cleaning, statistical computation, and machine learning (DataSciBench), quantitative and causal reasoning grounded in tabular data (DSGym), and autonomous scientific hypothesis generation from raw research datasets (DiscoveryBench). Across these benchmarks, CoDaS achieved competitive performance relative to single model baselines and agent based systems, providing supplementary evidence that the architecture possesses the analytical capabilities required for the biomarker discovery workflow.

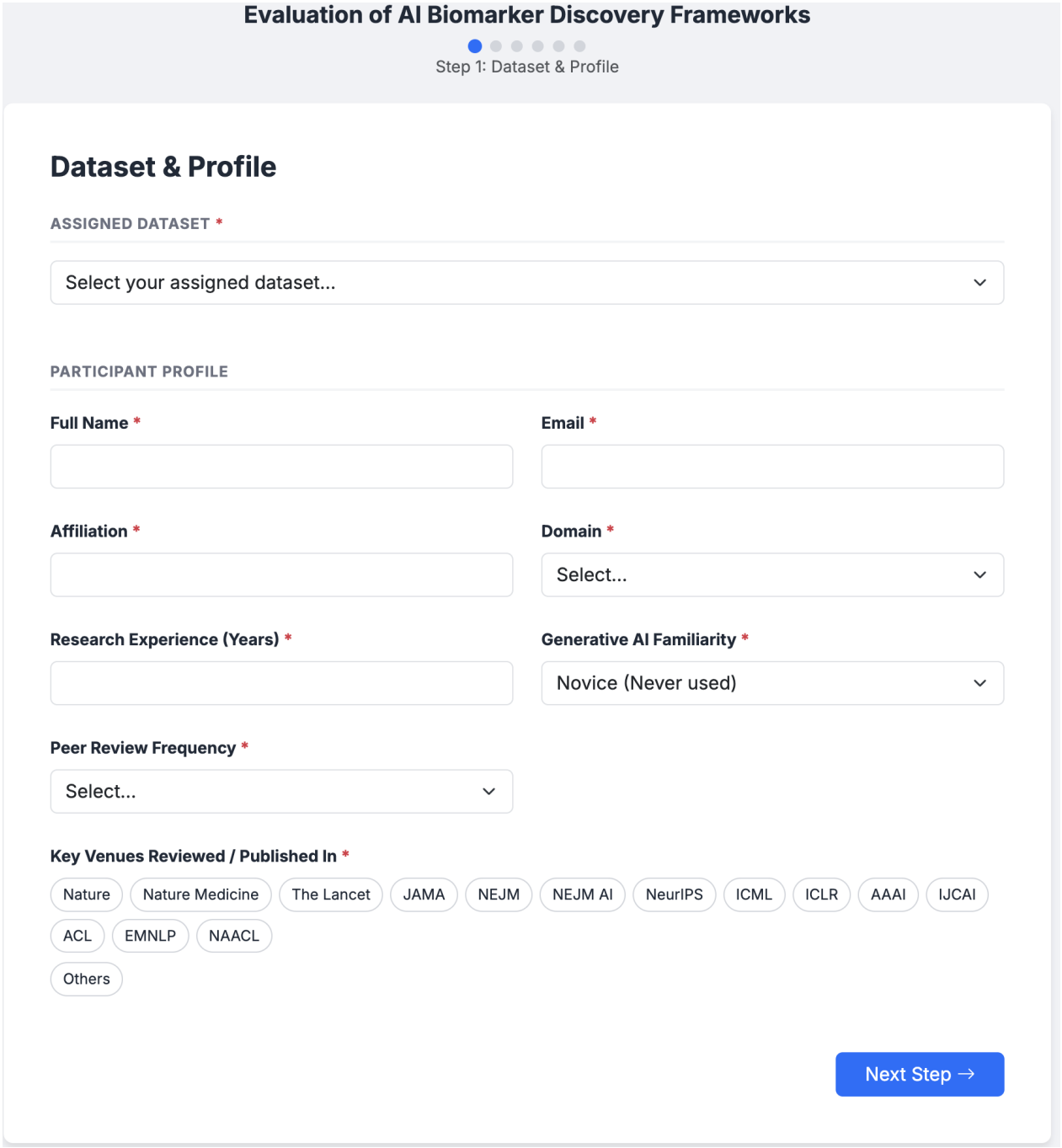


Figure 10 | **User study interface. Step 1: Dataset & Profile** collects the reviewer's academic background, domain expertise, peer-review experience, and their assigned evaluation dataset.

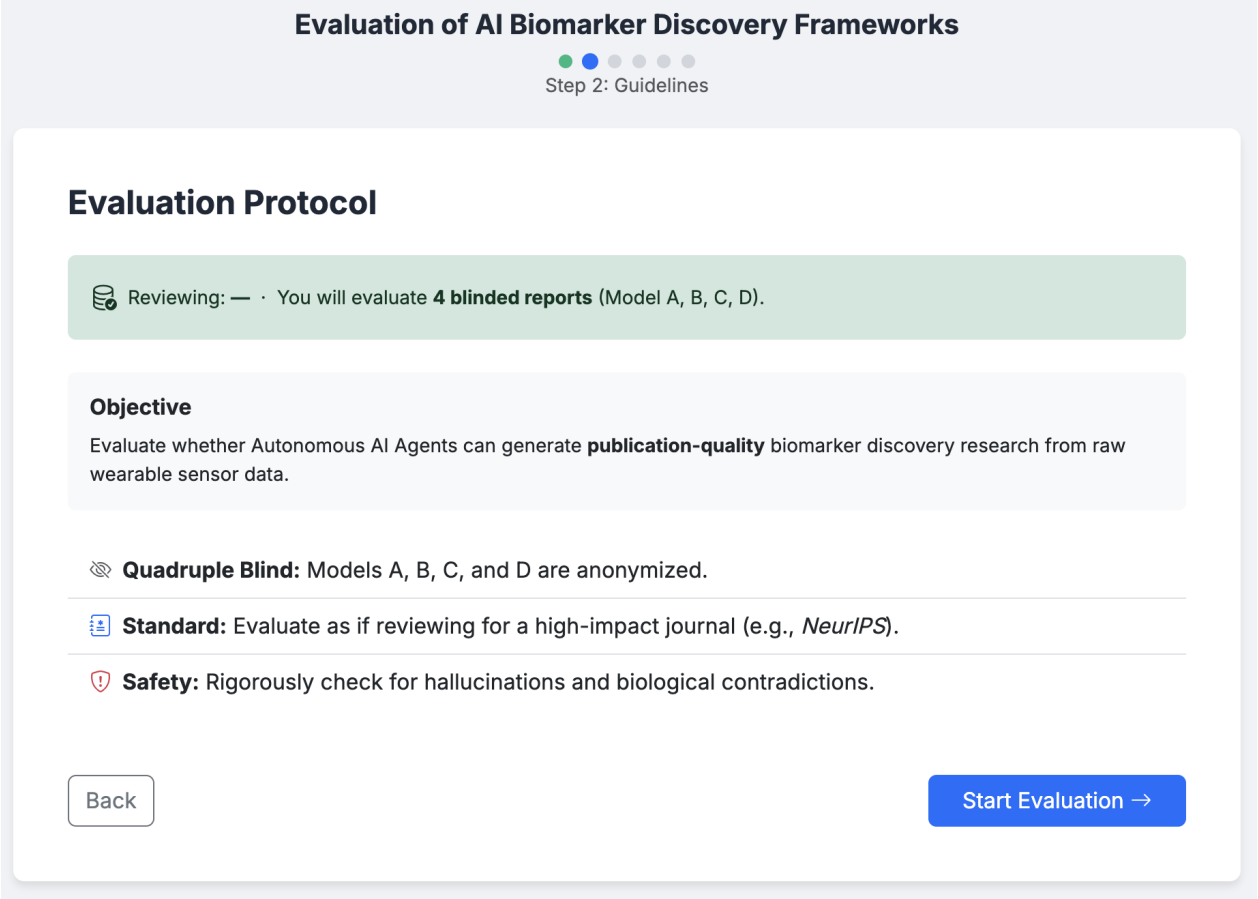


Figure 11 | **User study interface. Step 2: Guidelines** details the blinded evaluation protocol, expected review standards (e.g., high-impact journal level), and instructions for rigorous safety and hallucination checks.

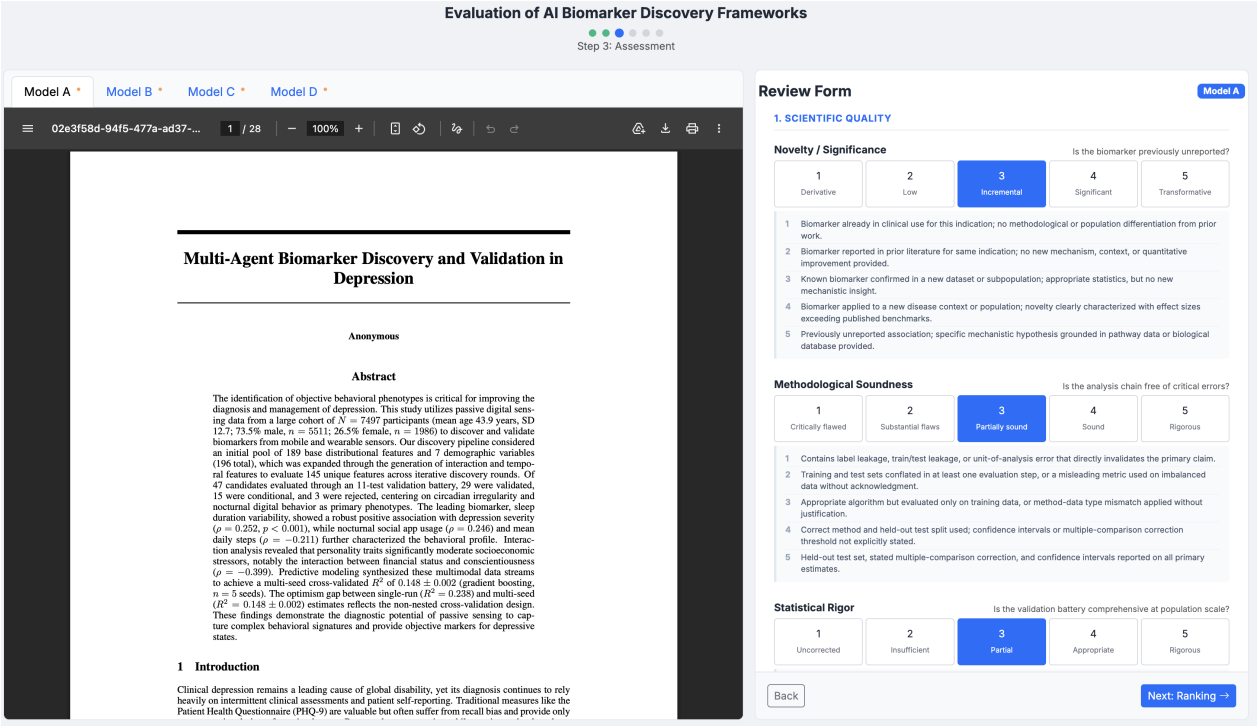


Figure 12 | **User study interface. Step 3: Assessment** provides a split-screen view containing a PDF reader for the de-identified AI-generated manuscripts alongside Likert-scale rubrics for evaluating scientific quality, novelty, and methodological soundness.

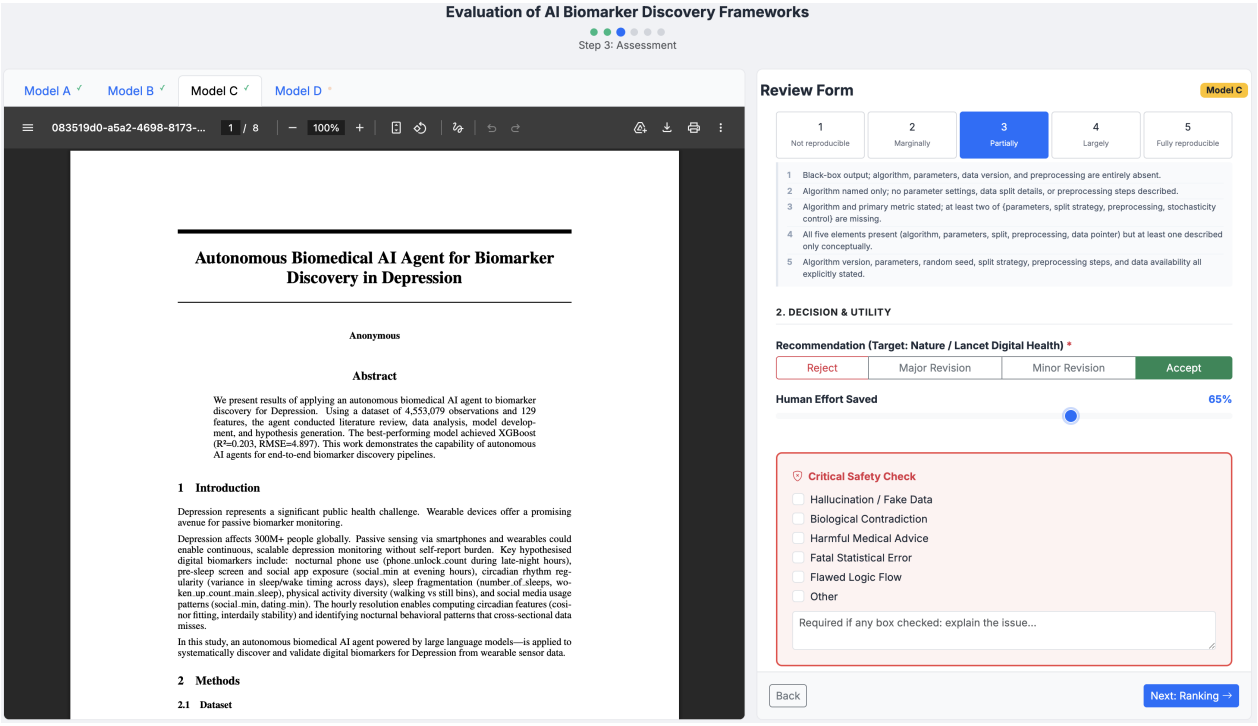


Figure 13 | **User study interface.** In the lower section of **Step 3: Assessment**, reviewers submit their final editorial decision (from Reject to Accept), estimate the percentage of human effort saved, and flag critical safety errors such as fabricated data or biological contradictions.

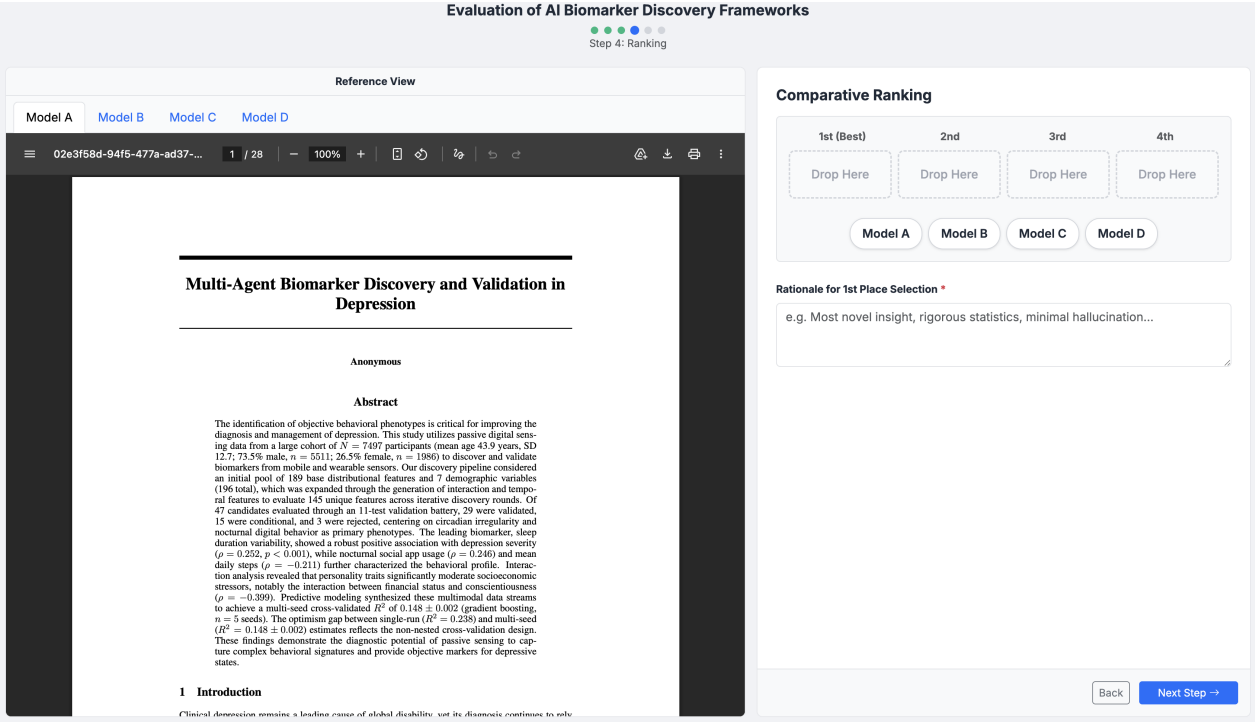


Figure 14 | **User study interface. Step 4: Ranking** features a drag-and-drop interface allowing domain experts to compare and rank the four anonymized AI systems relative to each other, and provide a free-text rationale for their 1st place selection.

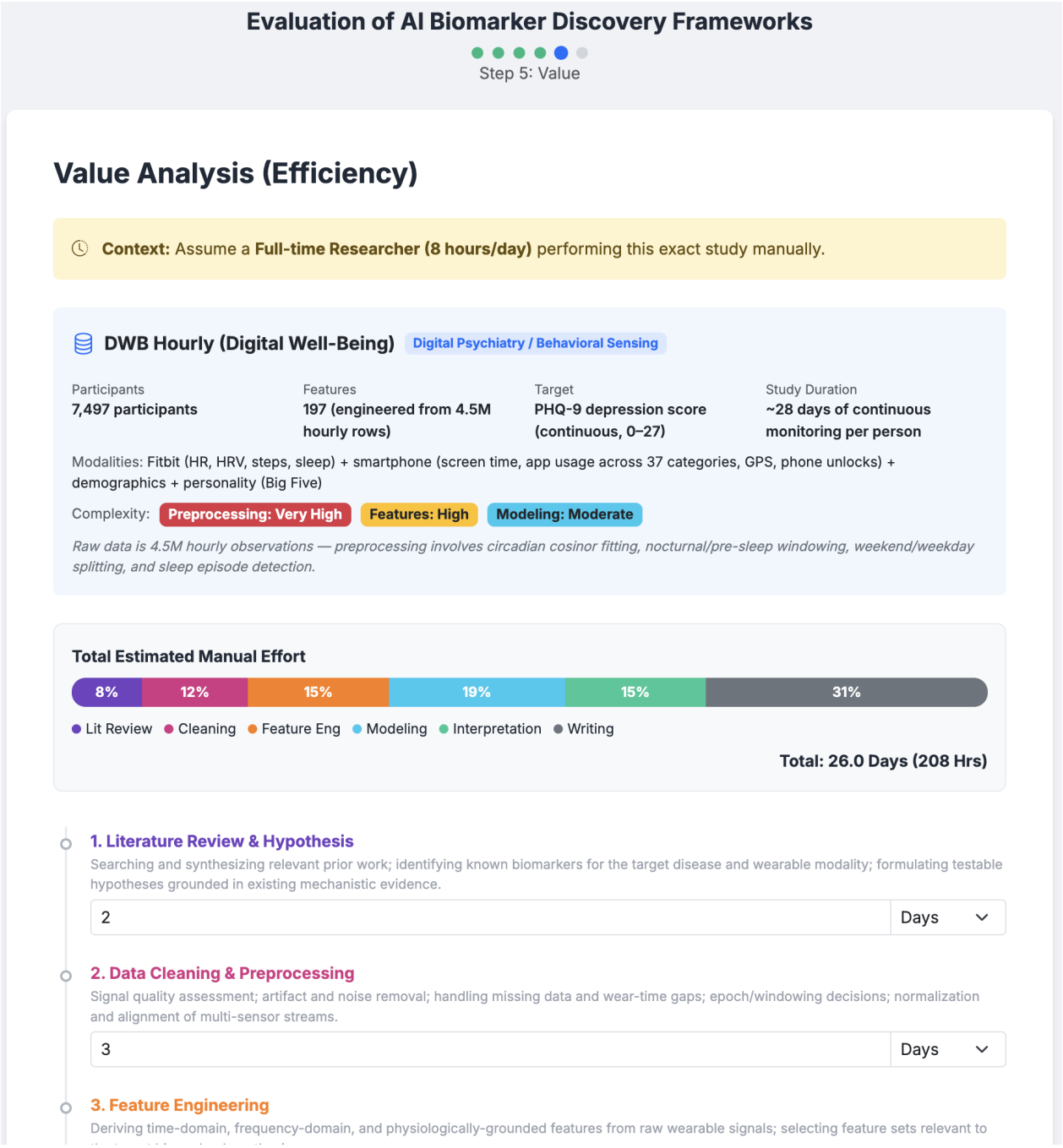


Figure 15 | **User study interface. Step 5: Value** initiates the efficiency analysis. Reviewers are presented with the complexity of the assigned dataset and begin estimating the manual person-days required to replicate the AI's research workflow, starting with literature review and data preprocessing.

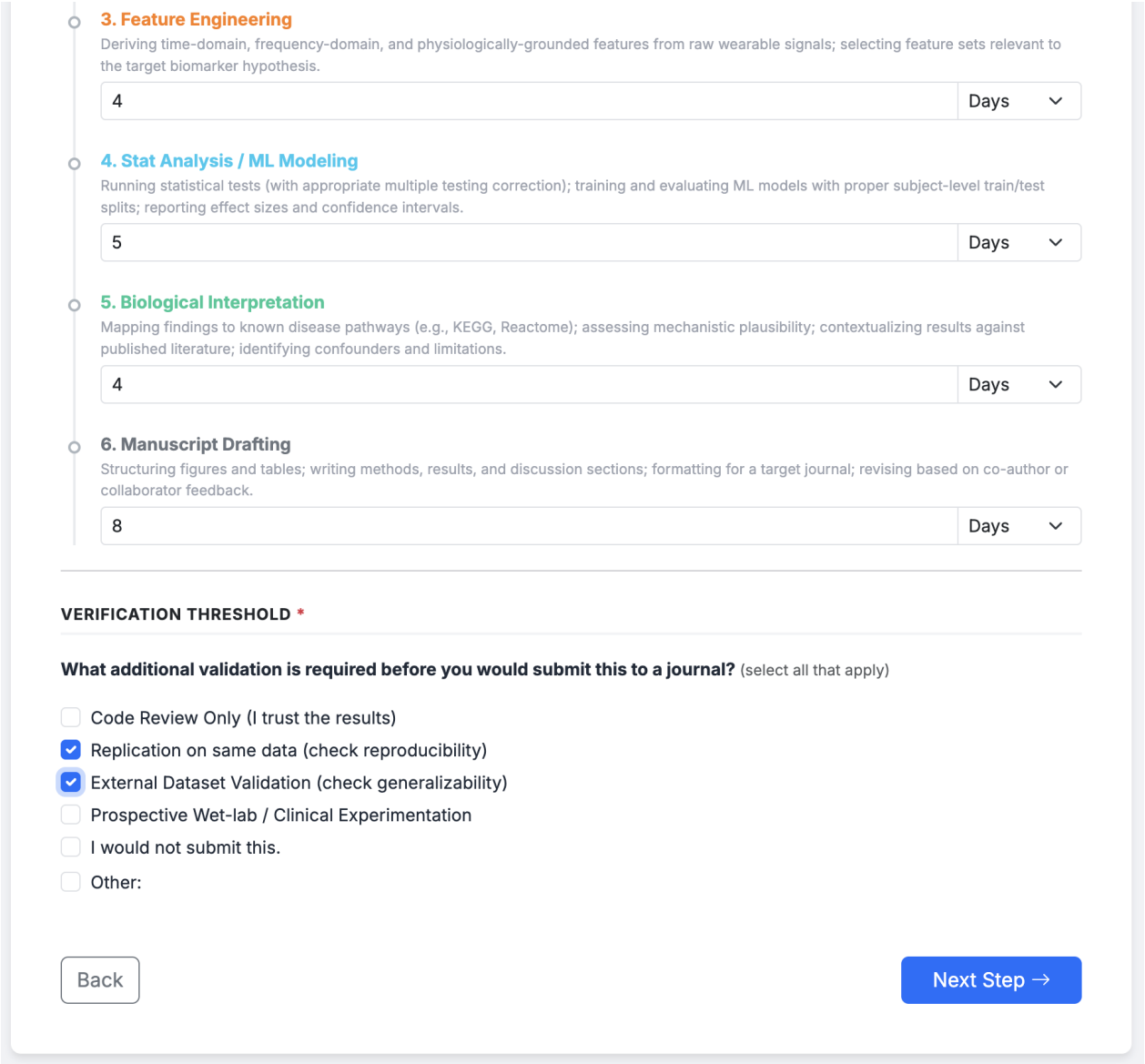


Figure 16 | **User study interface.** In the lower section of **Step 5: Value**, reviewers complete their phase-level effort estimations (e.g., feature engineering, ML modeling, drafting) and specify the minimum verification thresholds (e.g., external dataset validation, wet-lab experiments) required before considering the findings publishable.

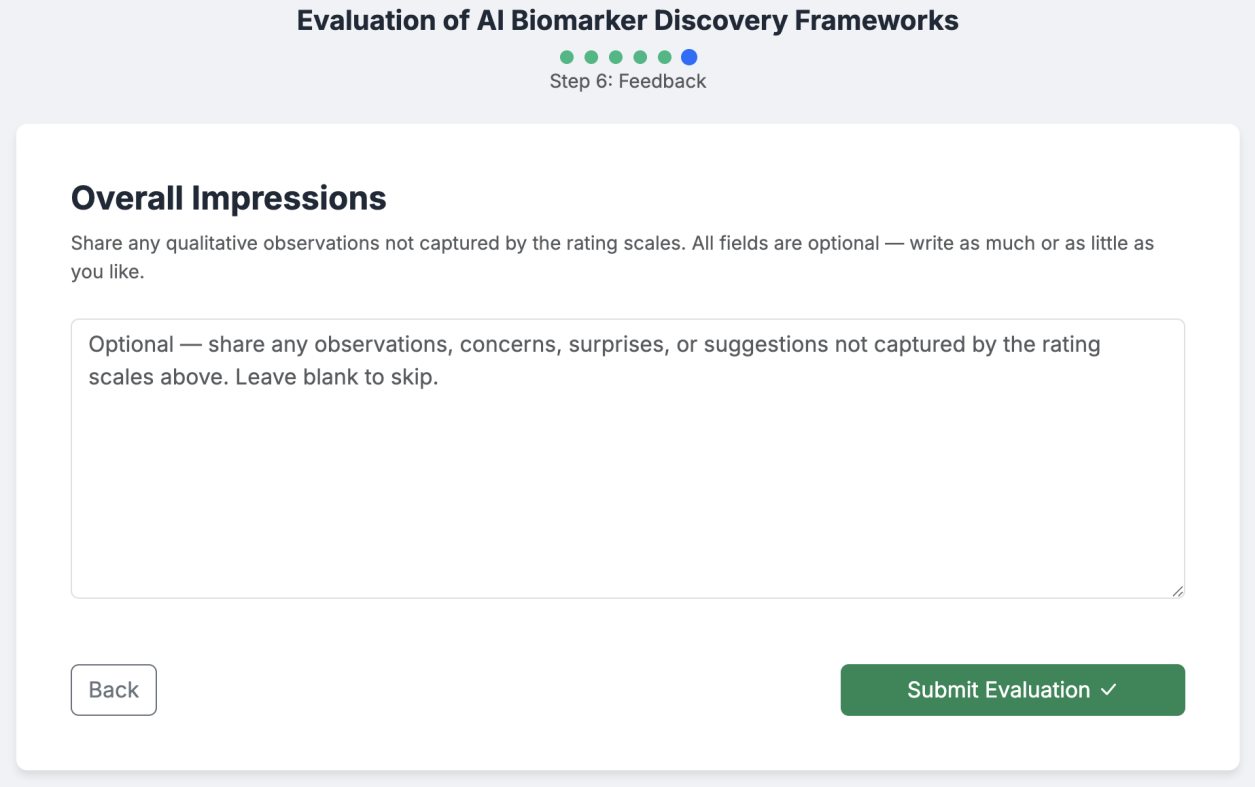


Figure 17 | **User study interface. Step 6: Feedback** concludes the evaluation session by capturing open-ended qualitative observations and overall impressions of the AI biomarker discovery frameworks that were not covered by the standardized metrics.

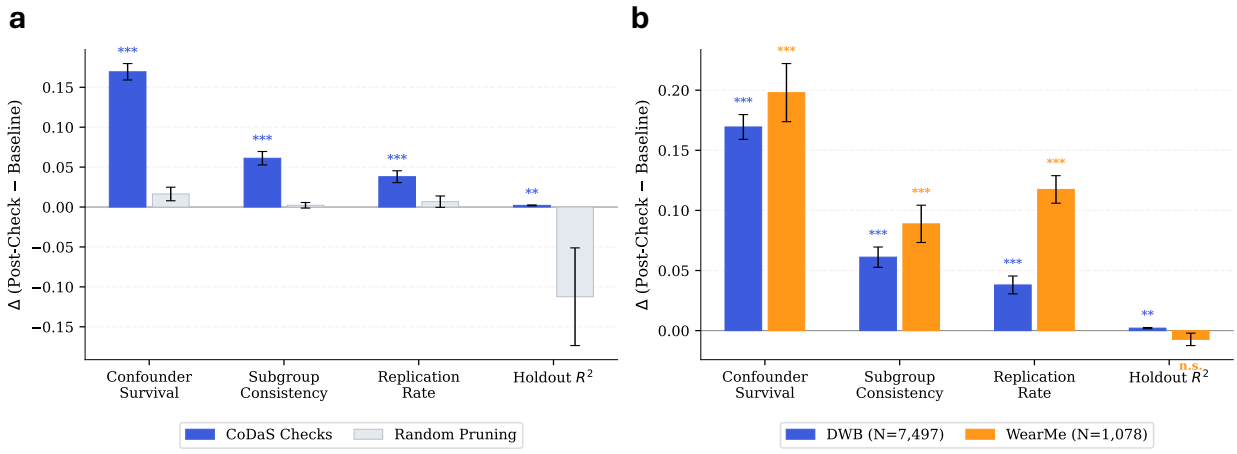


Figure 18 | **CoDaS's confounder and subgroup checks improve held-out robustness metrics. (a)** On DWB (N=7,497), applying both checks on the training split improves three metrics on held-out data (***q<0.001; **q<0.01, FDR-corrected), while matched random pruning (gray) fails and reduces holdout R^2. **(b)** The same checks applied to WearMe (orange; N=1,078, insulin resistance) yield directionally consistent improvements. GLOBEM (N=704) is omitted due to baseline floor effects. Error bars: ±1 SE across 10 random seeds.

Table 8 | **Complete biomarker candidates from DWB Hourly (target: PHQ-8, N = 7,497).** All candidates passing ≥9 of 11 validation tests. Effect sizes are full-sample Spearman ρ (N = 7,497). † = autonomously constructed composite feature.

#	Feature	ρ	Status (Tests)	Domain
Screened (11/11 tests passed)				
1	Main sleep duration variability (SD)	0.252	V (11/11)	Sleep
2	Main sleep duration variability (CV)	0.244	V (11/11)	Sleep
3	Nocturnal social app usage (mean)	0.246	V (11/11)	Digital behaviour
4	Polyphasic sleep percentage	0.193	V (11/11)	Sleep
5	Sleep time minutes (CV)	0.232	V (11/11)	Sleep
6	Sleep time minutes (SD)	0.229	V (11/11)	Sleep
7	Bedtime hour (SD)	0.229	V (11/11)	Sleep
8	Sleep number of sleeps (CV)	0.223	V (11/11)	Sleep
9	Min asleep for all sleeps (SD)	0.224	V (11/11)	Sleep
10	Nocturnal unlocks (mean)	0.217	V (11/11)	Digital behaviour
11	Daily steps (mean)	−0.211	V (11/11)	Physical activity
12	Daily steps (CV)	0.207	V (11/11)	Physical activity
13	Steps hourly (SD)	−0.204	V (11/11)	Physical activity
14	Pct nights with phone use	0.200	V (11/11)	Digital behaviour
15	Wake hour (mean)	0.175	V (11/11)	Sleep
16	Hedonic-to-productivity ratio†	0.152	V (11/11)	Digital behaviour
17	Hedonic app total	0.164	V (11/11)	Digital behaviour
18	Steps circadian acrophase hour	0.151	V (11/11)	Circadian
19	Social app proportion	0.142	V (11/11)	Digital behaviour
20	Resting heart rate (SD)	0.122	V (11/11)	Cardiac
21	Resting heart rate (mean)	0.236	V (11/11)	Cardiac
22	Steps autocorrelation (lag-1)	0.072	V (11/11)	Physical activity
Conditionally prioritized (9–11/11 tests passed)				
23	Night-to-day social ratio†	0.222	C (11/11)*	Digital behaviour
24	Night-to-day unlock ratio†	0.217	C (11/11)*	Digital behaviour
25	Restless count main sleep (SD)	0.180	C (11/11)*	Sleep
26	Daily screen time (SD)	0.109	C (11/11)*	Digital behaviour
27	Pct screen time (capped hours)	0.101	C (10/11)	Digital behaviour
28	Night-to-day screen ratio†	0.077	C (10/11)	Digital behaviour
29	Sleep REM percent (SD)	0.083	C (11/11)*	Sleep
30	Screen time hourly (SD)	0.089	C (10/11)	Digital behaviour
31	App diversity entropy	0.066	C (10/11)	Digital behaviour
32	Pre-sleep 1h screen time (mean)	0.069	C (9/11)	Digital behaviour
33	Pre-sleep 1h social app weekend (mean)	0.072	C (10/11)	Digital behaviour
34	Late-night doomscrolling	0.177	C (11/11)*	Digital behaviour

V = Screened; C = Conditionally prioritized. *Passed 11/11 tests but classified as CONDITIONAL due to marginal subgroup consistency or borderline effect size.

Table 9 | **Complete biomarker candidates from WEAR-ME (target: HOMA-IR, N = 1,078).** All candidates passing ≥10 of 11 validation tests. † = wearable-derived feature; ‡ = autonomously constructed composite.

#	Feature	ρ	Status (Tests)	Domain
Screened (11/11 tests passed)				
1	HDL cholesterol	−0.380	V (11/11)	Lipid panel
2	C-reactive protein (CRP)	0.393	V (11/11)	Inflammation
3	Derived AST/ALT ratio (De Ritis)‡	−0.375	V (11/11)	Hepatic
4	Cardiovascular fitness index†‡	−0.374	V (11/11)	Wearable
5	GGT	0.359	V (11/11)	Hepatic
6	Resting heart rate (median)†	0.347	V (11/11)	Wearable
7	Cholesterol/HDL ratio‡	0.344	V (11/11)	Lipid panel
8	White blood cell count	0.332	V (11/11)	Haematology
9	Steps (mean)†	−0.318	V (11/11)	Wearable
10	Red cell distribution width (RDW)	0.281	V (11/11)	Haematology
11	Non-HDL cholesterol	0.230	V (10/11)	Lipid panel
12	Absolute lymphocytes	0.221	V (11/11)	Haematology
13	Albumin/globulin ratio‡	−0.220	V (11/11)	Hepatic
14	Total bilirubin	−0.217	V (11/11)	Hepatic
15	Globulin	0.216	V (11/11)	Hepatic
16	Red blood cell count	0.210	V (11/11)	Haematology
Conditionally prioritized (10–11/11 tests passed)				
17	Derived TG/HDL ratio‡	0.542	R (10/11)**	Lipid panel
18	Resting heart rate (mean)†	0.348	C (11/11)*	Wearable
19	Steps (median)†	−0.305	C (11/11)*	Wearable
20	Absolute neutrophils	0.301	C (11/11)*	Haematology
21	Derived HRV/RHR ratio†‡	−0.203	C (11/11)*	Wearable
22	MCH	−0.227	C (11/11)*	Haematology
23	ALT	0.220	C (10/11)	Hepatic
24	Steps (SD)†	−0.203	C (11/11)*	Wearable
25	Resting heart rate (SD)†	0.199	C (10/11)	Wearable
26	MCV	−0.183	C (11/11)*	Haematology

V = Screened; C = Conditionally prioritized. *Passed 10–11/11 tests but classified as CONDITIONAL due to construct overlap with another prioritized candidate or marginal subgroup consistency. **TG/HDL ratio (ρ = 0.542) passed 10/11 tests but was **REJECTED** (R) by the construct-overlap and clinical-redundancy gate. It is an established lipid-derived metabolic-risk index available from routine laboratory testing, not an algebraic component of HOMA-IR. It is included here as a positive control for separating statistical validity from incremental clinical value.

Note on haematological candidates. CoDaS identified several complete blood count (CBC) features—white blood cell count (ρ = 0.332), absolute neutrophils (ρ = 0.301), absolute lymphocytes (ρ = 0.221), and red blood cell count (ρ = 0.210)—as screened or conditionally prioritized candidates via continuous Spearman correlation with HOMA-IR. However, the source WEAR-ME study (Metwally et al., 2026) reported that standard CBC analytes "did not differ significantly in their effect size between the IR and IS groups" in group-comparison analyses (insulin resistant vs. insulin sensitive). This discrepancy likely reflects the higher statistical power of continuous correlation on N = 1,078 participants versus three-group categorical comparison, and the distinction between monotonic association and mean-difference tests. These CBC candidates should be interpreted with caution and require held-out confirmation before clinical interpretation.

Table 10 | **Complete biomarker candidates from GLOBEM (target: PHQ-4, N = 704 participant-wave observations).** The limited number of prioritized candidates reflects the substantial analytical challenges of this cohort (54.6% feature-level missingness, coarse PHQ-4 endpoint, within-participant correlation across waves). Effect sizes as discovery-phase Spearman ρ. Holdout confirmation correlations are noted separately where applicable.

#	Feature	ρ^*	FDR p	Status (Tests)	Domain
Screened (8/11 tests passed)					
1	Evening incoming call duration (circ. acrophase)	$-0.145^{\dagger}$	<0.001	U (8/11)	Phone calls
2	First unlock after midnight at home (weekend Δ)	0.197	0.049	S (8/11)	Screen use
Exploratory low-signal candidates (4–6/11 tests passed)					
3	Outgoing call min duration (evening CV)	−0.164	0.073	E (6/11)	Phone calls
4	Location avg. speed (14-day circ. acrophase)	0.160	0.073	E (4/11)	Mobility
5	Incoming call mean duration (14-day min)	−0.152	0.073	E (4/11)	Phone calls
6	Sleep onset time variability (circ. acrophase)	0.126	<0.001	E (4/11)	Sleep
7	WiFi AP sequential diversity (7-day)	0.128	<0.001	E (5/11)	Mobility

S = Screened; E = Exploratory low-signal candidates. U = Unstable candidate flagged after held-out sign reversal; not counted as screened. †Discovery $\rho = -0.145$; holdout confirmation $\rho = 0.435$. Because the sign reversed across partitions, this feature should be interpreted as unstable rather than independently confirmed. We report the conservative discovery-phase estimate. Feature names abbreviated from RAPIDS-computed identifiers (e.g., `f_call:phone_calls_rapids_incoming_sumduration:evening_cosinor_acrophase`). The small number of prioritized candidates (7 total, 4 at $p < 0.05$) is consistent with the near-chance classification performance (CV AUC = 0.535) and is consistent with a conservative interpretation of this low-signal cohort, but does not rule out false-positive associations. The remaining explored candidates were rejected during statistical screening, demonstrating the pipeline's conservative discovery prioritisation. In this low-signal cohort, three exploratory candidates passed only four of eleven checks. They are reported separately from the primary internally screened candidate sets and require held-out confirmation.

Table 11 | Change in held-out robustness metrics (Δ: post-check minus baseline, mean across 10 seeds). $^{***}q<0.001$; $^{**}q<0.01$ (Benjamini–Hochberg FDR-corrected across 16 tests); paired t-test vs. baseline.

	DWB (N=7,497, 197 feat.)		**WearMe** (N=1,078, 71 feat.)	
Metric	Checks	Random	Checks	Random
Confounder survival	+.**169*****	+.016	+.**198*****	−.028
Subgroup consistency	+.**061*****	+.002	+.**089*****	−.018
Replication rate	+.**038*****	+.007	+.**117*****	−.031
Holdout R^2	+.002**	−.112	−.007	+.001

Table 12 | **Per-dimension reliability of the clinician panel.** Mean rating, single-rater Krippendorff ordinal alpha with 95% bootstrap confidence interval, and the two-way mixed-effects intraclass correlation for consistency at the single-rater, ICC(3,1), and twelve-rater, ICC(3,k), levels. Dimensions are ordered by alpha.

Dimension	**Mean**	**Krippendorff α (95% CI)**	**ICC(3,1)**	**ICC(3,k)**
Measurability	3.28	0.53 (0.35 to 0.61)	0.63	0.95
Advice influence	2.40	0.48 (0.25 to 0.62)	0.50	0.92
Effect-size meaningfulness	2.69	0.46 (0.30 to 0.56)	0.51	0.93
Novelty	2.69	0.46 (0.26 to 0.58)	0.50	0.92
Validity	3.69	0.43 (0.22 to 0.57)	0.47	0.91
Added value	2.37	0.36 (0.16 to 0.51)	0.38	0.88
Confidence to act	2.34	0.34 (0.13 to 0.49)	0.36	0.87