An Agentic System for Automated Data Curation and Analysis in Large-Scale Biobanks

Chang-Uk Jeong* CHANG-UK.JEONG@PENNMEDICINE.UPENN.EDU
University of Pennsylvania, USA

Jaesik Kim* JAESIK.KIM@PENNMEDICINE.UPENN.EDU
University of Pennsylvania, USA

Jaehyun Joo JAEHYUN.JOO@PENNMEDICINE.UPENN.EDU
University of Pennsylvania, USA

Byounghan Lee QUDGKS96@AJOU.AC.KR
Ajou University, South Korea

Yang-Gyun Kim† APPLE88400@GMAIL.COM
Kyung Hee University Hospital at Gangdong, South Korea

Dokyoon Kim† DOKYOON.KIM@PENNMEDICINE.UPENN.EDU
University of Pennsylvania, USA

Abstract

The translation of clinical and lifestyle concepts into computable phenotypes is a critical yet manually intensive bottleneck in leveraging large-scale biomedical datasets like the UK Biobank. This process is slow, requires deep domain expertise, and suffers from a lack of scalability and reproducibility, especially for clinicians unfamiliar with large-scale data analysis. We propose and develop an autonomous, dual-component agentic system designed to automate the research workflow from hypothesis to report. The first component, the large language model (LLM)-based data preprocessing framework, systematically searches the UK Biobank's public data dictionary, translating high-level clinical and lifestyle concepts into machine-readable rules. The second component, the Analysis Agent, autonomously executes the statistical analysis plan and synthesizes the findings. The framework is further validated by successfully phenotyping and analyzing several clinical and lifestyle screeners. This work demonstrates a viable end-to-end system that enhances scalability and democratizes complex data analysis with transparency, representing a foundational step toward a new paradigm of AI-driven scientific discovery.

Keywords: AI Agent, UK Biobank, Large Language Model, Phenotyping

Data and Code Availability The UK Biobank data underlying this study cannot be shared publicly due to participant privacy and ethical restrictions. A source code is publicly available at https://github.com/ukjung21/ukb-agent.

Institutional Review Board (IRB) This study does not require IRB approval.

1. Introduction

The advent of large-scale biomedical repositories, such as the UK Biobank (UKB) (Bycroft et al., 2018), has provided unprecedented opportunities for population health research. These resources contain deep genetic, imaging, and clinical data for hundreds of thousands of individuals, enabling the study of complex interactions between lifestyle, environment, and disease. A critical component of leveraging these datasets is phenotyping: the process of translating abstract health concepts or clinical instrument scores into precise, computable definitions using the available data fields.

However, the traditional process of phenotyping is a significant bottleneck (Li et al., 2024). First, it is a manual, resource-intensive endeavor. For instance, when attempting to define a lifestyle with a

* These authors contributed equally
† co-corresponding authors

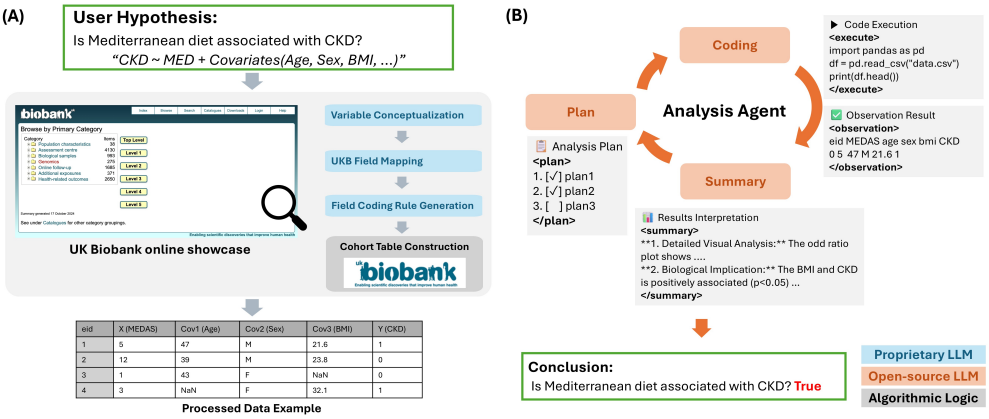


Figure 1: An overview of UKB Agent. The system is composed of two main components. (A) The LLM-based data preprocessing workflow translates a high-level user hypothesis into a curated, analysis-ready dataset by automatically generating phenotyping rules from the UK Biobank data dictionary. (B) The Analysis Agent takes the curated data and autonomously executes a statistical analysis plan to test the hypothesis, producing a final conclusion. It operates in a Human-in-the-loop collaboration mode, where the user confirms specific analysis details such as how to handle missing data or which statistical models to apply before the analysis is performed.

medical questionnaire or screening system using available UK Biobank fields, a researcher must manually search for suitable fields that can serve as proxies for each item in the questionnaire. This process is not only susceptible to inconsistencies but can also lead to the omission of crucial data fields. Furthermore, phenotyping algorithms for the same condition often vary significantly between studies, which complicates meta-analyses (Patel et al., 2022; Torralbo et al., 2025; Kong et al., 2022; Wei et al., 2024). To address these issues, several computational tools have been developed (Hanscombe et al., 2019; Kiral et al., 2020; Yeung et al., 2022). While these tools have been developed to do either data processing or analysis in UK Biobank data, they still depend on a manually defined and curated phenotype and are not entirely automated.

Recently, the paradigm of autonomous AI agents has emerged as a powerful tool for complex problem-solving in scientific domains (Gottweis et al., 2025; Lu et al., 2024; Gridach et al., 2025). These agents, often powered by Large Language Models (LLMs), can devise plans, use tools, and iteratively work towards a goal with minimal human intervention. Groundbreaking research such as Biomni has demonstrated the potential of multi-agent systems to perform complex bioinformatics tasks by planning and executing code in a self-correcting manner (Huang et al., 2025). Similarly, other studies have shown agents capable of designing experiments and searching literature, heralding a new era of automated scientific discovery (Boiko et al., 2023; Swanson et al., 2025; Gao et al., 2024).

To address the challenges in phenotyping and analysis, we propose an agentic framework for automated data curation and analysis in large-scale biobanks. While the core agentic logic—including relevant field refinement, coding rule generation, automated analysis—is designed to be portable across biobank infrastructures, we present the UK Biobank as the first validation case study for this framework. This system, referred to as the UKB Agent in this context, is designed to automate the research lifecycle from hypothesis to validation, aiding clinicians and researchers who may be unfamiliar with the intricacies of large-scale biobank data architectures. By presenting a scenario analyzing the association between a lifestyle factor and a disease outcome, this study introduces a more scalable, transparent, and accessible approach to biomedical data science. We believe this agent-based methodology is a promising step toward accelerating the pace of discovery and uncovering novel insights into the determinants of health and disease.

2. Method

Our framework is designed to automate the research lifecycle, from the formulation of a structured hypothesis to the final analytical report. The entire process is initiated by the user, who must provide a hypothesis structured with three key components:

- **Exposure**: A high-level concept for the exposure of interest, typically the name of a clinical or lifestyle screening instrument (e.g., 'MED', 'PHQ-9', 'Alcohol consumption').
- **Outcome**: A list of ICD-9 and ICD-10 codes to define disease outcome (e.g., [585.9, N18.0, N18.5] for Chronic Kidney Disease).
- **Covariates**: A list of confounding factors for statistical adjustment (e.g., 'age', 'sex', 'BMI').

For instance, a user hypothesizes an *"association between CKD (defined by specific ICD codes) and Mediterranean diet (as measured by MED), with covariates age, sex, and BMI."*

To automate the full research lifecycle from hypothesis to conclusion, we designed and implemented a dual-component framework. The first component, the LLM-based data preprocessing workflow, is engineered to translate abstract variables from a research hypothesis into computable definitions using the UKB data dictionary. The second component, the Analysis Agent, receives the curated, analysis-ready dataset from the Phenotyping Framework. It then autonomously performs the statistical analyses required to test the user's initial hypothesis.

2.1. LLM-based Data Preprocessing Workflow

The primary objective of the data preprocessing workflow is to find relevant fields in the UK Biobank for each variable within a given hypothesis and to generate rational rules for their coding using LLMs. We used the proprietary LLM, `gpt-5-mini-2025-08-07` model, in this stage. For each variable, the framework iteratively executes the following phases to produce a set of variable coding rules. These rules are subsequently applied to the Biobank to generate the final curated data for analysis (Figure 1A).

2.1.1. Description of the UK Biobank Data Showcase

All data processing and mapping operations in our workflow are performed using the publicly available UK Biobank Data Showcase (https://biobank.ndph.ox.ac.uk/showcase/). This showcase serves as a comprehensive data dictionary that details the thousands of variables available to researchers.

The showcase is organized hierarchically. At the highest level are *Category*, which group related information into broad domains such as "Biological samples" or "Genomics". Within each *Category* are numerous Data *Field*, which represent the individual variables. Each Data *Field* is annotated with its descriptive *Notes*, and for categorical variables, its *Coding* information. The *Coding* provides a direct mapping between numerical codes and their human-readable meanings.

2.1.2. Phase 1: Variable Conceptualization

The data preprocessing workflow begins with the automated comprehension of each variable. The goal is to deconstruct such a concept into a structured specification. Lifestyle phenotypes, such as the Mediterranean Diet (MED) score, are composite scores derived from multiple pieces of information, requiring detailed deconstruction. To address this, we employ LLMs augmented with web search capabilities. The system obtains the full name, description, a complete list of questionnaire items, and the precise encoding logic for each item via web search. Simple phenotypes like age or sex are identified as single, direct measurements and bypass this intensive deconstruction.

2.1.3. Phase 2: Semantic UKB *Field* Mapping

The core of the workflow is the accurate mapping of each variable to its proxies within the UKB data. With a structured definition for each variable component, the workflow maps them to the relevant UKB data fields using a Breadth-First Search-LLM (BFS-LLM) algorithm. The process begins with a BFS traversal of the UKB *Category* hierarchy to ensure a comprehensive search. For every *Category*, both it and its sub-categories are gathered as a branch. The model is provided with complete context, which includes the names, descriptions, and field lists of all categories within the branch, to assess if the entire branch is relevant to the target variable. If a *Cate-*

gory and its sub-categories are considered to be irrelevant, the entire branch is pruned from the search, significantly reducing the search space.

Upon reaching a relevant, *Field*-bearing *Category* after BFS traversal, the system performs a final *Field*-level Evaluation. For each relevant *Category*, the system examines its *Fields*. It enriches each *Field* by scraping its detailed *Notes* and *Coding* information from the UKB showcase. This enriched context is fed back to `gpt-4.1-mini` for high-speed, large-volume initial field relevance screening. After this initial pass, `gpt-5-mini`, employing advanced reasoning for final validation and selection, reviews all potentially relevant fields for a single item and selects only the most suitable ones that serve as proxies.

To further enhance efficiency, we have also implemented an embedding-based search method as an optional alternative to the initial *Fields* relevance screening with BFS-LLM approach. This method uses the `abhinand/MedEmbed-base-v0.1` model to compute semantic embeddings of *Fields* descriptions and performs rapid similarity-based retrieval. This option significantly reduces search time while maintaining reasonable accuracy, providing users with a flexible trade-off between speed and comprehensiveness.

2.1.4. Phase 3: Field Coding Rule Generation

In this phase, the verified mappings are translated into executable JSON-based rules. For each item-to-field mapping, an LLM is prompted with the item's encoding criteria and the coding details of the mapped UKB *Field*. The LLM's task is to generate a precise rule that connects each field's specific value to a numeric score. For example, it translates "Score 1 for consuming 4 or more tablespoons of olive oil daily" into a machine-readable condition like `{"field_id": "26110", "conditions": {"operator": ">=", "value": 4}, "score": 1}`. Then LLM ensures that all relevant fields are logically combined to produce a single value based on the encoding criteria.

2.1.5. Phase 4: Cohort Table Construction

Once the rule sets for all variables in the hypothesis are generated, they are systematically applied to the raw UK Biobank dataset. This computational phase executes the JSON rules against the large-scale data table. For each participant, the workflow calculates the necessary variables, determines the outcome status based on ICD codes, and extracts the values for all specified covariates. The final output is an analysis-ready cohort table where each row represents a unique participant and each column represents a fully processed and curated variable.

2.2. The Analysis Agent

Once the dataset is curated by the preprocessing workflow, it is passed to the Analysis Agent, which is responsible for conducting the scientific inquiry from statistical analysis to final reporting. Instead of a simple reactive loop, the agent is built upon a sophisticated, stateful, graph-based architecture that ensures a structured and transparent, and secure analytical framework.

To facilitate user interaction with the system, we developed an intuitive web-based interface using the `Streamlit` framework. The analytical workflow is initiated when a user provides an initial request, such as a research hypothesis and the path to their dataset, through an interactive chat prompt.

2.2.1. Core Architecture: A Stateful Graph-Based Approach

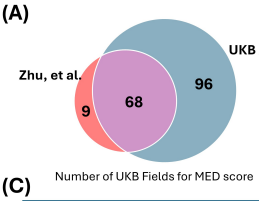
The agent's core is implemented as a state graph using the `LangGraph` library, functioning as a state machine that transitions between two primary states: *generation* and *execution*. The *generation* state serves as the reasoning engine, where the agent plans its next move, interprets results, or synthesizes findings. The *execution* state is the action engine, where the agent interacts with its tools, primarily a Python interpreter.

Crucially, to uphold the strict data privacy requirements of the UK Biobank, all analytical tasks are performed by a locally-hosted, open-source LLM (`ChatOllama` with the `gpt-oss:120b` and `gemma3:27b` model). This design ensures that no sensitive, participant-level data is ever transmitted outside the researcher's secure computational environment, guaranteeing data integrity and confidentiality.

2.2.2. The Structured Reasoning and Action Cycle

As shown in Figure 1B, the agent operates on a structured reasoning and action cycle, which is more rigorous than a generic *plan-act-observe* paradigm (Yao

(A)



Number of UKB Fields for MED score

(B)

Food Component	Vegetables	Fruit	Olive oil	Red and processed meat	White meat	Butter and margarine	Sugary drinks	Wine	Legumes	Nuts	Seafood	Sweets and desserts	Sofrito
Fields Recall	2/2 (100%)	2/2 (100%)	1/1 (100%)	4/4 (100%)	1/1 (100%)	6/7 (86%)	1/1 (100%)	6/6 (100%)	4/7 (57%)	4/4 (100%)	6/6 (100%)	31/36 (86%)	-
Additional fields	35	11	-	11	4	5	8	-	2	2	5	11	2

(C)

Question	Field	Generated Rule	Rationale
Do you consume 2 or more servings of vegetables per day?	**104060**: Mixed vegetable intake	$\in \{300, 2\} \rightarrow 1$	Codes 300 (3+) and 2 (2) indicate consumption of 2 or more servings of vegetables per day
Do you consume 3 or more servings of fruit (including fruit juice) per day?	**26089**: Apples and pears	$\geq 240 \rightarrow 1$	Standard UK fruit portion ≈ 80 g. Three servings = 3 × 80 g = 240 g.
Do you consume less than 1 serving of red and processed meat per day?	**26066**: Beef	$< 100 \rightarrow 1$	The item awards a point for consuming <1 serving of red/processed meat per day. A commonly used single serving of red meat is 100 g.
Do you consume less than 1 serving of sugary drink per day?	**26127**: Sugar-sweetened beverages and other sugary drinks	$< 250 \rightarrow 1$	The question asks for <1 serving/day. A standard beverage serving is approximated as 250 mL (~250 g, assuming density ≈1 g/mL).
Do you consume 2 or more servings of wine per week?	**4407**: Average monthly red wine intake	$\geq 9 \rightarrow 1$	2 servings/week = 2 * 52 / 12 ≈ 8.67 glasses per month
Do you consume 3 or more servings of legumes per week?	**26101**: Legumes and pulses	$\geq 35 \rightarrow 1$	Field 26101 reports grams consumed in the past 24 hours. To approximate '3 or more servings per week' convert servings to a daily equivalent: using a typical UK cooked legume serving of 80 g, 3 servings/week = 240 g/week ≈ 34.3 g/day.

Figure 2: Detailed outputs of the automated phenotyping process for the Mediterranean Diet (MED) score. (A) A summary of the concordance between UKB fields identified by the UKB Agent and those from a prior expert-curated study (B) A summary of the concordance for each of the 13 MED items. (C) Examples of the machine-readable rules and the corresponding natural language rationale generated by the agent.

et al., 2023). Every action is preceded by an explicit reasoning step, enclosed in `<think>` tags, where the agent articulates its rationale. This process, analogous to Chain-of-Thought prompting, makes the agent's decision-making process fully transparent. Following this, the agent must choose one of four predefined actions, each denoted by a specific tag:

1. **Planning (`<plan>`):** The agent first establishes a step-by-step analytical strategy in a checklist format. This plan is revisited and updated dynamically as the analysis progresses.
2. **Code Execution (`<execute>`):** To perform data analysis, the agent writes and runs Python code within a secure interpreter.
3. **Result Interpretation (`<summary>`):** After executing code, the agent receives the output as an `<observation>`, which can be text, a table, or a plot. The agent then uses the `<summary>` action to perform self-reflection, interpreting the results from both a detailed visual or statistical perspective and a broader biological or clinical context.
4. **Final Hypothesis validation (`<solution>`):** Once all steps in the plan are marked as complete, the agent invokes the `<solution>` action to synthesize all findings, interpretations, and results into a cohesive final report that addresses the user's initial hypothesis.

2.2.3. Human-in-the-Loop for Analytical Consistency

To enhance analytical utility and ensure run-to-run consistency, the agent integrates a critical Human-in-the-loop checkpoint. This is particularly crucial for large-scale biobank analysis, where common analytical decisions—such as the strategy for handling missing data or the selection of specific statistical models—are highly consequential and often guided by researcher preference. Without this user-guided step, the agent's autonomous choices might vary between sessions, leading to inconsistent results when comparing different analytical runs.

Operationally, this checkpoint is triggered after the agent performs an initial Exploratory Data Analysis (EDA). The agent then pauses its operation and, via the web interface, presents the user with explicit proposals for the subsequent data handling and statistical modeling steps. The user must then confirm or modify these choices. Once the user provides this confirmation, the agent locks in these decisions

and proceeds with the analysis, guaranteeing that the workflow aligns perfectly with the researcher's intended methodology.

2.2.4. Iterative Refinement and Visual Analysis

The state-graph architecture facilitates a robust process of iterative refinement and self-correction. If a code execution step results in an error, the agent observes the error message, re-enters the generation state to diagnose the problem, and attempts to fix the code in a subsequent `<execute>` action.

A key capability of the agent is its capacity for multimodal visual analysis. It is designed to automatically capture any plots, encode them as base64 images, and append them to the message history. This allows the LLM to see the visualizations it creates. The agent can then analyze patterns, trends, and distributions directly from the plots such as interpreting an odds ratio plot or a histogram and integrate these visual insights into its scientific reasoning and reporting within the `<summary>` step, leading to a more profound and human-like analytical process.

3. Results

To evaluate the efficacy of our framework, we tasked the agent with real-world exposure phenotyping challenges, such as Mediterranean diet adherence (MED), Generalized Anxiety Disorder-7 (GAD-7), and Recent Depressive Symptoms-4 (RSD-4) using the UK Biobank (UKB) dataset. We then compared our system's output against a manually curated definition from a peer-reviewed study. Subsequently, to assess the autonomous analysis capabilities of our analysis agent, we compared the agent's analysis result on the curated data with an analysis conducted by a human expert.

3.1. Successful Variable Conceptualization

We evaluated the agent's ability to autonomously discover relevant clinical constructs from high-level terminology. When provided with general concepts such as "anxiety screening" or "Mediterranean diet adherence", the system successfully performed web-based searches to identify standardized assessment instruments. Specifically, the agent autonomously discovered the Generalized Anxiety Disorder-7 (GAD-7) scale for anxiety assessment and the MEDAS-14 (Mediterranean Diet Adherence Screener) for diet evaluation.

3.2. Case Study: Automated Phenotyping of the Mediterranean Diet (MED)

3.2.1. Validation of UKB Field Mapping

Our system's primary task was to identify relevant UKB data fields that could serve as proxies for each of the 13 items in the MED score. To ensure a fair comparison, we bypassed the web-based information extraction. Instead, the MED score information from Zhu et al. (2025) was directly provided to our system. The results, summarized in Figure 2A, demonstrate a remarkable concordance between the fields identified by our analysis agent and those selected by human experts in Zhu et al. (2025).

As shown in Figure 2B, for most MED components, the agent successfully identified most of the UKB fields used in Zhu et al. (2025) or their alternatives:

- **100% coverage:** In 9 out of 13 categories, the agent utilized all the fields from Zhu et al. (2025) and, in addition, used additional fields that were also manually confirmed to be related to each item.
- **Majority coverage:** In **Butter and margarine**, **Legumes**, and **Sweets and Desserts**, the agent utilized a majority of the fields chosen by human experts and, in addition, used other related fields.
- **Alternative fields extracted:** While Zhu et al. (2025) did not map any fields for **sofrito**, the agent reasonably checked for the use of 'tomato-based sauce' in 'Field 20088: Types of spreads/sauces consumed' and 'Field 10430: Tinned tomato intake' as a proxy for **sofrito**.

Notably, for several items like **Vegetables** and **Sweets and Desserts**, our agent not only successfully identified the fields from prior work Zhu et al. (2025) but also proposed a substantially larger set of novel fields. We manually verified the relevance of these additional fields. A comprehensive list of all identified fields is provided in Appendix A. This suggests that the agent may offer a more comprehensive search than manual methods, potentially increasing the robustness of the resulting phenotype. The extensive list of food codes for 'Sweets and desserts'

showcases the agent's ability to navigate the granular detail of the UKB data dictionary effectively.

3.2.2. Validation of Automated Rule Generation and Rationale

Beyond simply identifying the correct fields, the agent was tasked with generating the encoding logic required to translate raw data into a MED score. Figure 2C provides examples of these generated rules and the agent's rationale. The evaluation confirms that the agent's logic is not only correct but also demonstrates a sophisticated ability to handle the nuances of phenotyping.

The agent consistently formulated valid rules based on sound reasoning:

- **Direct Mapping:** For simple cases like **vegetable consumption** (Field 104050), the agent correctly interpreted the categorical codes and generated a logical rule ($\in \{300, 2\} \rightarrow 1$) to award a point for consuming two or more servings.
- **Handling Unit and Frequency Discrepancies:** The agent demonstrated its ability to bridge gaps between the MED score's requirements and the UKB data's structure.
  - For **red meat**, the MED item is based on "servings," while the UKB field (26065) is in grams. The agent rationally inferred a standard serving size of 100g to create the rule ($< 100 \rightarrow 1$).
  - For **wine**, the agent correctly converted the weekly serving requirement to a monthly value to match the UKB field's frequency, resulting in a sound threshold of ≥ 9 glasses per month.

In summary, these results show that our agent can successfully replicate and even potentially enhance the manual process of phenotyping. The generated rules are not arbitrary but are founded on logical conversions, validating the agent's capability as a reliable tool for complex biomedical data analysis.

To evaluate the capabilities of our Analysis Agent, we tasked it with a representative biomedical research hypothesis: investigating the association between Mediterranean diet (measured by the generated MED score) and Chronic Kidney Disease (CKD), using the curated dataset derived from our Phenotyping Framework. The agent operated entirely autonomously, starting from the initial data exploration to the final synthesis of results, with its process and findings detailed below.

3.3. Phenotyping of Other Exposures

We systematically replicated more exposure phenotypes from peer-reviewed studies that publicly documented their UK Biobank field codes. Specifically, we also replicated multiple definitions of lifestyle variables, such as alcohol consumption, as documented by Kim et al. (2024). We replicated two widely-used screening systems—the Generalized Anxiety Disorder-7 (GAD-7) and the Recent Depressive Symptoms-4 (RSD-4)—following the methodology described by Dutt et al. (2022).

This expanded analysis demonstrates that our automated phenotyping framework can successfully handle diverse variable types, lifestyle factors such as alcohol consumption and sedentary behavior, and validated screening instruments (GAD-7, RSD-4). Our system achieved perfect replication accuracy in field mapping and reasonably captured extra supplementary fields as well. The detailed results are summarized in Appendix B

3.4. Agent-Driven Analysis of the User's Hypothesis with Human-in-the-loop

We tasked the Analysis Agent with the user's hypothesis, investigating the association between the Mediterranean diet (MED) score and Chronic Kidney Disease (CKD) using the curated dataset. A critical component of our framework is the Human-in-the-Loop checkpoint, which ensures analytical consistency and reproducibility. We experimentally confirmed that without this step, the agent might make different autonomous decisions (e.g., choosing data imputation instead of row deletion for missing values), leading to variations in the final analysis results.

In this analysis, the agent's initial EDA identified that 90% of MED score records were missing. The agent paused and presented this to the user, who provided the explicit instruction: "drop rows with any missing value. Proceed with logistic regression." The agent then autonomously executed this user-defined plan. The resulting multivariable logistic regression model was highly significant (LLR p-value: 3.01×10^{-245}, Pseudo R-squared: 0.089). The analysis found a statistically significant protective association for the MED score (OR = 0.859, 95% CI [0.822,

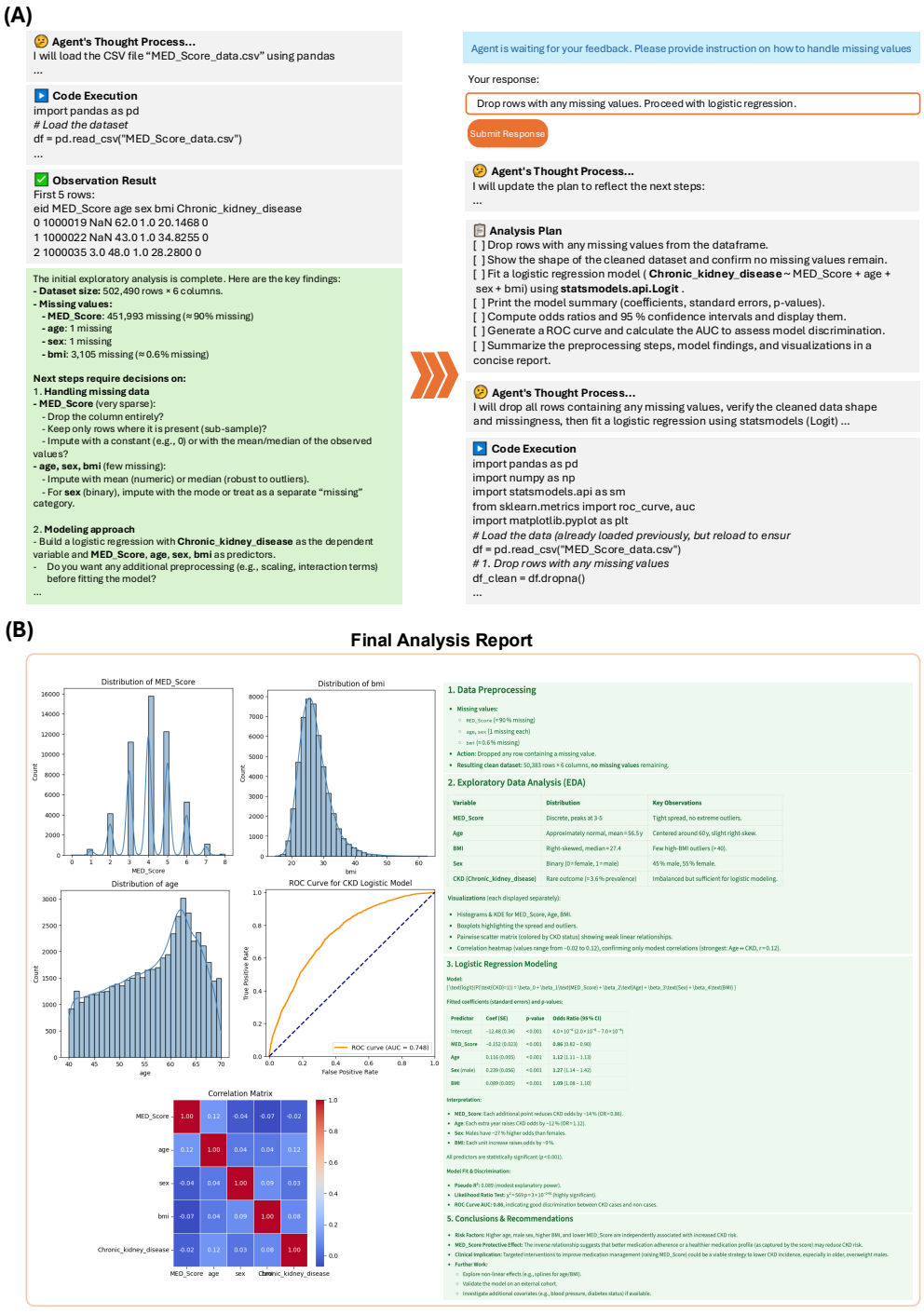


Figure 3: (A) The user interface displaying the agent's transparent, step-by-step analysis (B) Plots and final report created by the Analysis Agent.

Table 1: A result comparison between a human expert and the UKB Agent for CKD

Variable	Source	Estimate	Odds Ratio (OR)	95% CI for OR	p-value
MED Score	Expert	-0.1515	0.859	(0.822, 0.898)	1.83e-11
	UKB Agent	-0.1515	0.859	(0.822, 0.898)	<0.001
Age	Expert	0.1158	1.123	(1.113, 1.133)	<2e-16
	UKB Agent	0.1158	1.123	(1.113, 1.133)	<0.001
Sex (male=1)	Expert	0.2386	1.269	(1.138, 1.416)	8.71e-05
	UKB Agent	0.2386	1.269	(1.138, 1.416)	<0.001
BMI	Expert	0.0891	1.093	(1.082, 1.104)	<2e-16
	UKB Agent	0.0891	1.093	(1.082, 1.104)	<0.001

0.898]), with all covariates being significant predictors ($p < 0.001$). The complete, step-by-step agentic analysis, including all generated code and intermediate observations, is available in the Supplementary file.

Crucially, the results produced by the agent following this Human-in-the-loop directive were bit-for-bit identical to an independent analysis conducted by a human expert following the same protocol (Table 1). This validates the agent's ability to perfectly replicate expert-level analysis when precisely guided by the researcher, guaranteeing both scientific validity and run-to-run consistency.

4. Discussion

A cornerstone of our framework is the systematic workflow for phenotyping, designed to ensure comprehensive field discovery. The manual curation of phenotypes is not only labor-intensive but also susceptible to inconsistencies that can compromise scientific validity. As highlighted in (Huang et al., 2021), minor variations in cohort definitions can lead to statistically significant shifts in study outcomes. Our Hybrid BFS-LLM algorithm mitigates this risk. By combining the exhaustive nature of BFS with the semantic pruning of LLM, our method ensures that no relevant data fields are overlooked. However, the choice of model and search strategy presents significant trade-offs, as shown in Table A1. For the 'Fruit' item, using the BFS-LLM approach, the advanced `gpt-5-mini-2025-08-07` model took over 14 minutes but comprehensively identified 26 fields. In contrast, the faster `gpt-4.1-mini-2025-04-14` model completed in under 4 minutes but found only 2 fields. Our alternative embedding-based search method resolves this trade-off, offering both high speed (1-2 minutes) and high comprehensiveness (27-28 fields) regardless of the model used.

Despite its promise, our framework has limitations. The agent's analytical sophistication, while competent, may not yet match the nuanced judgment of a human expert in selecting optimal statistical models for complex scenarios. Moreover, the validity of the output fundamentally depends on the quality of the initial user-provided hypothesis. The system automates the process but cannot correct for a scientifically flawed premise. Finally, its performance relies on the accuracy of the LLM and the clarity of the public documentation for the biobank data fields. Human oversight, therefore, remains essential for validating the final results.

Looking forward, we envision this framework as a foundational step toward the paradigm of an AI co-scientist (Swanson et al., 2025; Gottweis et al., 2025). Our next plans involve extending the framework's capabilities. First, we will adapt the system to other large-scale biobanks, such as the *All of Us* Research Program (Investigators, 2019), by developing new data-parsing modules. Second, we aim to incorporate multi-modal data types, including genomics and medical imaging, to enable more complex, integrative analyses. The long-term vision, inspired by the agentic science movement (Wei et al., 2025), is to evolve the agent from a hypothesis-tester to a hypothesis-generator. In conclusion, we have developed and demonstrated an autonomous agentic system that successfully automates the end-to-end research lifecycle, from computable phenotype generation in large-scale biobanks to final statistical analysis and reporting. This work represents a foundational step toward a new paradigm of biomedical research, positioning AI agents as powerful co-scientists capable of accelerating our understanding of the determinants of human health and disease.

Acknowledgement

This work was partially supported by NIGMS R01 GM138597 and NHLBI R01 HL169458, and was conducted using the UK Biobank Resource under Application Number 45556.

References

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624:570–578, 12 2023. ISSN 14764687. doi: 10.1038/s41586-023-06792-0.

Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Alan Young, Mark Effingham, Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly, and Jonathan Marchini. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562:203–209, 10 2018. ISSN 14764687. doi: 10.1038/s41586-018-0579-z.

Rosie K Dutt, Kayla Hannon, Ty O Easley, Joseph C Griffis, Wei Zhang, and Janine D Bijsterbosch. Mental health in the uk biobank: A roadmap to self-report measures and neuroimaging correlates. *Human Brain Mapping*, 43(2):816–832, 2022. doi: https://doi.org/10.1002/hbm.25690. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.25690.

Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and Marinka Zitnik. Empowering biomedical discovery with ai agents. 7 2024. URL http://arxiv.org/abs/2404.02831.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Pushmeet Kohli, Yossi Matias, Andrew Carroll, Kavita Kulkarni, Nenad Tomasev, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R D Costa, José R Penadés, Gary Peltz, Yunhan Xu, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. Towards an ai co-scientist. 2 2025. URL http://arxiv.org/abs/2502.18864.

Mourad Gridach, Jay Nanavati, Khaldoun Zine El Abidine, Lenon Mendes, and Christina Mack. Agentic ai for scientific discovery: A survey of progress, challenges, and future directions. 3 2025. URL http://arxiv.org/abs/2503.08979.

Ken B. Hanscombe, Jonathan R.I. Coleman, Matthew Traylor, and Cathryn M. Lewis. Ukbtools: An r package to manage and query uk biobank data. *PLoS ONE*, 14, 5 2019. ISSN 19326203. doi: 10.1371/journal.pone.0214311.

Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li, Lin Qiu, Gavin Li, Junze Zhang, Di Yin, Shruti Marwaha, Jennefer N. Carter, Xin Zhou, Matthew Wheeler, Jonathan A. Bernstein, Mengdi Wang, Peng He, Jingtian Zhou, Michael Snyder, Le Cong, Aviv Regev, and Jure Leskovec. Biomni: A general-purpose biomedical ai agent, 6 2025. URL http://biorxiv.org/lookup/doi/10.1101/2025.05.30.656746.

Yidi Huang, William Yuan, Isaac S Kohane, and Brett K Beaulieu-Jones. Illustrating potential effects of alternate control populations on real-world evidence-based statistical analyses. *JAMIA Open*, 4(2):ooab045, 06 2021. ISSN 2574-2531. doi: 10.1093/jamiaopen/ooab045. URL https://doi.org/10.1093/jamiaopen/ooab045.

AURP Investigators. The "all of us" research program. *New England Journal of Medicine*, 381(7):668–676, 2019.

Min Seo Kim, Injeong Shim, Akl C. Fahed, Ron Do, Woong-Yang Park, Pradeep Natarajan, Amit V. Khera, and Hong-Hee Won. Association of genetic risk, lifestyle, and their interaction with obesity and obesity-related morbidities. *Cell Metabolism*, 36(7):1494–1503.e3, Jul 2024. ISSN 1550-4131. doi: 10.1016/j.cmet.2024.06.004. URL https://doi.org/10.1016/j.cmet.2024.06.004.

Isabell Kiral, Nathalie Willems, and Benjamin Goudey. Ukbcc: a cohort curation package for uk biobank, 7 2020. URL http://biorxiv.org/lookup/doi/10.1101/2020.07.12.199810.

Dongdong Kong, Tim R. McVicar, Mingzhong Xiao, Yongqiang Zhang, Jorge L. Peña-Arancibia, Gianluca Filippa, Yuxuan Xie, and Xihui Gu. phenofit:

An r package for extracting vegetation phenology from time series remote sensing. *Methods in Ecology and Evolution*, 13:1508–1527, 7 2022. ISSN 2041210X. doi: 10.1111/2041-210X.13870.

Xiangnan Li, Yaqi Huang, Shuming Wang, Meng Hao, Yi Li, Hui Zhang, and Zixin Hu. Lukb: preparing local uk biobank data for analysis. *Bioinformatics Advances*, 4, 2024. ISSN 26350041. doi: 10.1093/bioadv/vbae176.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. 9 2024. URL http://arxiv.org/abs/2408.06292.

Riyaz S. Patel, Spiros Denaxas, Laurence J. Howe, Rosalind M. Eggo, Anoop D. Shah, Naomi E. Allen, John Danesh, Aroon Hingorani, Cathie Sudlow, and Harry Hemingway. Reproducible disease phenotyping at scale: Example of coronary artery disease in uk biobank. *PLoS ONE*, 17, 4 2022. ISSN 19326203. doi: 10.1371/journal.pone.0264828.

Kyle Swanson, Wesley Wu, Nash L. Bulaong, John E. Pak, and James Zou. The virtual lab of ai agents designs new sars-cov-2 nanobodies. *Nature*, 7 2025. ISSN 0028-0836. doi: 10.1038/s41586-025-09442-9.

Ana Torralbo, Jonathan M. Davitte, Damien C. Croteau-Chonka, Cai Ytsma, Chris Tomlinson, Natalie K. Fitzpatrick, Sheng Chia Chung, Ghazaleh Fatemifar, Adrian S. Cortes, Tom G. Richardson, Matthew Barclay, Julia Carrasco-Zanini, Chris Finan, Harry Hemingway, Aroon D. Hingorani, Valerie Kuan, Claudia Langenberg, Georgios Lyratzopoulos, R. Thomas Lumbers, Maik Pietzner, Anoop D. Shah, Johan H. Thygesen, Natalie Zelenka, John C. Whittaker, Margaret G. Ehm, and Spiros Denaxas. A computational framework for defining and validating reproducible phenotyping algorithms of 313 diseases in the uk biobank. *Scientific Reports*, 15, 12 2025. ISSN 20452322. doi: 10.1038/s41598-025-05838-9.

Jiaqi Wei, Yuejin Yang, Xiang Zhang, Yuhan Chen, Xiang Zhuang, Zhangyang Gao, Dongzhan Zhou, Guangshuai Wang, Zhiqiang Gao, Juntai Cao, Zijie Qiu, Xuming He, Qiang Zhang, Chenyu You, Shuangjia Zheng, Ning Ding, Wanli Ouyang, Nanqing Dong, Yu Cheng, Siqi Sun, Lei Bai, and Bowen Zhou. From ai for science to agentic science: A survey on autonomous scientific discovery, 2025. URL https://arxiv.org/abs/2508.14111.

Wei-Qi Wei, Robb Rowley, Angela Wood, Jacqueline MacArthur, Peter J Embi, and Spiros Denaxas. Improving reporting standards for phenotyping algorithm in biomedical research: 5 fundamental dimensions. *Journal of the American Medical Informatics Association*, 31(4):1036–1041, January 2024. ISSN 1527-974X. doi: 10.1093/jamia/ocae005. URL https://doi.org/10.1093/jamia/ocae005. _eprint: https://academic.oup.com/jamia/article-pdf/31/4/1036/57148411/ocae005.pdf.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. 3 2023. URL http://arxiv.org/abs/2210.03629.

Ming Wai Yeung, Pim van der Harst, and Niek Verweij. ukbpheno v1.0: An r package for phenotyping health-related outcomes in the uk biobank. *STAR Protocols*, 3, 9 2022. ISSN 26661667. doi: 10.1016/j.xpro.2022.101471.

Kai Zhu, Rui Li, Pang Yao, Hancheng Yu, An Pan, Jo Ann E. Manson, Eric B. Rimm, Walter C. Willett, and Gang Liu. Proteomic signatures of healthy dietary patterns are associated with lower risks of major chronic diseases and mortality. *Nature Food*, 6:47–57, 1 2025. ISSN 26621355. doi: 10.1038/s43016-024-01059-x.

Appendix A. Figure and Table

Food Component	Vegetables	Fruit (including fruit juice)	Olive oil	Red and processed meat	White meat	Butter and margarine	Sugary drinks	Wine	Legumes	Nuts	Seafood	Sweets and desserts	Sofrito
Zhu, et al. (*Nat Food*, 2025)	**1289, 1299**	**1309, 1319**	**20090**	**1349, 1369, 1379, 1389**	**1359**	**101310, 101350, 101390, 101430, 101470, 101510**, 101550	**100170**	**1568, 1578, 1608, 4407, 4418, 4451**	20088, 103270, 103310, **104000, 104010, 104110, 104120**	**102410, 102420, 102430, 102440**	**1329, 1339, 103150, 103200, 103210, 103220**	100540, 100550, 101260, **101970, 101980, 101990, 102010, 102020**, 102030, **102050, 102060, 102070, 102120, 102140, 102150,** 102170, **102180, 102190, 102200, 102210, 102210, 102220, 102250, 102260, 102270, 102280, 102290, 102310, 102310, 102320, 102330, 102340, 102350, 102360, 102370, 102380**	?
Ours (BFS-LLM)	**1289, 1299,** 103990, 104000, 104010, 104030, 104060, 104070, 104080, 104090, 104110, 104120, 104130, 104140, 104150, 104160, 104170, 104180, 104190, 104200, 104210, 104230, 104240, 104250, 104260, 104270, 104280, 104290, 104300, 104310, 104320, 104330, 104340, 104350, 104360, 104370, 104380	**1309, 1319,** 26089, 26090, 26091, 26092, 26093, 26094, 26095, 100190, 100200, 100210, 100220	**20090**	**1349, 1369, 1379, 1389,** 26066, 26100, 26117, 26122, 103010, 103020, 103030, 103040, 103070, 103080, 103090	**1359,** 26069, 26121, 103050, 103060	20087, 26062, 26063, 101300, **101310, 101350, 101390, 101430, 101470, 101510**, 104040	26095, 26127, **100170**, 100180, 100200, 100220, 100230, 100530, 100550	**1568, 1578, 1608, 4407, 4418, 4451**	26101, **104000, 104010, 104110, 104120**, 104280	26107, 26108, **102410, 102420, 102430, 102440**	**1329, 1339, 103150**, 103160, 103170, 103180, 103190, **103200, 103210, 103220**, 103230	26064, 26068, 26080, 26084, 26085, 26140, **101970, 101980, 101990, 102010, 102020, 102050, 102060, 102070**, 102080, 102090, **102120**, 102130, **102140, 102150, 102180, 102190, 102200, 102210, 102220**, 102230, **102250, 102260, 102270, 102280, 102290,** 102300, **102310, 102320, 102330, 102340, 102350, 102360, 102370, 102380**	20088, 104350
Ours (Semantic Embedding)	**1289, 1299**, 26098, 26125, 26144, 26146, 26147, 103990, 104000, 104010, 104020, 104030, 104050, 104060, 104070, 104080, 104090, 104110, 104120, 104130, 104140, 104150, 104160, 104170, 104180, 104190, 104200, 104210, 104230, 104240, 104250, 104260, 104270, 104280, 104290, 104300, 104310, 104320, 104330, 104340, 104350, 104360, 104370, 104380	**1309, 1319,** 26091, 26092, 26093, 26095	2654, **20090**, 26009, 26032, 103980	**1349, 1369, 1379, 1389,** 26066, 26104, 26117, 26122, 103000, 103010, 103020, 103030, 103040, 103070, 103080	1349, **1359**, 1369, 1379, 1389, 26066, 26104, 26117, 26121, 26122, 103020, 103030, 103040, 103060	101300, **101310, 101430, 101470, 101510**, 101550, 103980, 104040	26001, 26003, 26095, 26126, 26127, **100170**, 100180, 100190, 100200, 100220, 100530, 100550	1558, **1568, 1578, 1608, 4407, 4418, 4451**, 20117, 100590, 100630, 100670, 100720,	26101, **104000, 104010, 104110, 104120**, 104280	26106, 26107, 26108, 102400, **102410, 102420, 102430, 102440,** 102450, 103290	**1329, 1339,** 26132, **103150**, 103160, 103170, 103180, 103190, **103200, 103210, 103220**, 103230	26011, 26080, 26085, 26140, **102120**, 102130, **102190**, 102230, **102250, 102260, 102280, 102290, 102330, 102340, 102350, 102360, 102370**	20088, 104340, 104350

Figure A1: Fields comparison between Zhu et al. (2025) and our systems for each component in MED score

Table A1: Comparison of methods and models in terms of the time and performance.

Method	Model	Time	Count	Items
BFS-LLM	gpt-5-mini-2025-08-07	14'16"	26	26089(Apples and pears), 26090(Berries), 26091(Citrus), 26092(Dried fruit), 26095(Fruit juice), 26093(Other fruit), 26094(Stewed fruit), 100190(Orange juice intake), 100200(Grapefruit juice intake), 100210(Pure fruit/vegetable juice intake), 100220(Fruit smoothie intake), 104410(Stewed fruit intake), 104430(Dried fruit intake), 104440(Mixed fruit intake), 104450(Apple intake), 104460(Banana intake), 104470(Berry intake), 104520(Melon intake), 104530(Orange intake), 104540(Satsuma intake), 104560(Pear intake), 104570(Pineapple intake), 104580(Plum intake), 104590(Other fruit intake), 1309(Fresh fruit intake), 1319(Dried fruit intake)
	gpt-4.1-mini-2025-04-14	3'58"	2	1309(Fresh fruit intake), 1319(Dried fruit intake)
Embedding-based	gpt-5-mini-2025-08-07	2'24"	28	1309(Fresh fruit intake), 1319(Dried fruit intake), 100026(Daily dietary data not credible), 104400(Fruit consumers), 104450(Apple intake), 104460(Banana intake), 104470(Berry intake), 104440(Mixed fruit intake), 104430(Dried fruit intake), 104520(Melon intake), 104540(Satsuma intake), 104560(Pear intake), 104500(Grape intake), 104510(Mango intake), 104550(Peach/nectarine intake), 104570(Pineapple intake), 104480(Cherry intake), 104420(Prune intake), 104580(Plum intake), 104410(Stewed fruit intake), 100210(Pure fruit/vegetable juice intake), 100220(Fruit smoothie intake), 26089(Apples and pears), 26090(Berries), 26092(Dried fruit), 26093(Other fruit), 26094(Stewed fruit), 26095(Fruit juice)
	gpt-4.1-mini-2025-04-14	1'03"	27	1309(Fresh fruit intake), 1319(Dried fruit intake), 104450(Apple intake), 104460(Banana intake), 104540(Satsuma intake), 104560(Pear intake), 104550(Peach/nectarine intake), 104570(Pineapple intake), 104480(Cherry intake), 104420(Prune intake), 104580(Plum intake), 104440(Mixed fruit intake), 104470(Berry intake), 104430(Dried fruit intake), 104490(Grapefruit intake), 104500(Grape intake), 104410(Stewed fruit intake), 104400(Fruit consumers), 100220(Fruit smoothie intake), 100210(Pure fruit/vegetable juice intake), 104590(Other fruit intake), 26093(Other fruit), 26092(Dried fruit), 26094(Stewed fruit), 26089(Apples and pears), 26090(Berries), 26091(Citrus)

Appendix B. More phenotypes varing complexity

B.1. Alcohol Consumption Replication

Query: Consumption of wine $\leq$ 150ml per day

- **Reference Fields:** 1568 (Average weekly red wine intake), 1578 (Average weekly champagne plus white wine intake), 4407 (Average monthly red wine intake), 4418 (Average monthly champagne plus white wine intake)
- **Selected Fields:** 100590 (Red wine intake), 100630 (Rose wine intake), 100670 (White wine intake), 100720 (Fortified wine intake), 20096 (Size of red wine glass drunk), 20097 (Size of rose wine glass drunk), 20095 (Size of white wine glass drunk), 1568 (Average weekly red wine intake), 1578 (Average weekly champagne plus white wine intake), 1608 (Average weekly fortified wine intake), 4407 (Average monthly red wine intake), 4418 (Average monthly champagne plus white wine intake), 4451 (Average monthly fortified wine intake), 1558 (Alcohol intake frequency.), 100560 (Other drink intake)
- **Rule Rationale:** Return 1 when the rule establishes that typical daily wine consumption is $\leq$ 150 ml based on prioritized sources: (1) yesterday counts with size checks, (2) weekly averages scaled to daily with size checks, (3) monthly averages scaled to daily with size checks, (4) direct 'no alcohol yesterday' or 'never drinks' indicators. If none of the conditions indicating $\leq$ 150 ml/day are met or data are missing/uninterpretable, the rule returns 0.

Query: Consumption of 4% ABV pint beer $\leq$ 1 per day

- **Reference Fields:** 1588 (Average weekly beer plus cider intake)
- **Selected Fields:** 1588 (Average weekly beer plus cider intake), 4429 (Average monthly beer plus cider intake), 100710 (Beer/cider intake), 26067 (Beer and cider), 100580 (Alcohol consumed), 1558 (Alcohol intake frequency.)
- **Rule Rationale:** If any of the available, transformed sources indicate beer/cider consumption $\leq$ 1 pint per day (or an explicit non-drinker flag yields 0), classify as 1 ($\leq$ 1 pint/day). Otherwise classify as 0. Missing/unavailable sources simply fail their respective conditions; the OR logic ensures classification if any reliable indicator supports $\leq$ 1 pint/day.

Query: Consumption of spirits $\leq$ 45ml per day

- **Reference Fields:** 1598 (Average weekly spirits intake), 4440 (Average monthly spirits intake)
- **Selected Fields:** 1598 (Average weekly spirits intake), 4440 (Average monthly spirits intake), 100730 (Spirits intake), 26138 (Spirits), 100580 (Alcohol consumed)
- **Rule Rationale:** Evaluate whether any available, non-missing source yields an estimated daily spirits volume $\leq$ 45 ml. Order/structure rationale: each source is checked only if it carries a usable numeric signal (we exclude known negative/missing codes using explicit checks). For weekly/monthly measures we convert measures $\rightarrow$ ml and normalize to daily; for categorical yesterday responses we map codes to numeric measures then to ml; for 24h grams we convert to ml via 1 g $\approx$ 1 ml. An explicit 'No' to 'Alcohol consumed yesterday' (100580 == 0) is treated as 0 ml and thus satisfies $\leq$ 45 ml. If none of these conditions are true (no quantitative evidence of $\leq$ 45 ml), the rule returns 0 (not $\leq$ 45 ml). This is intentionally conservative: only explicit quantitative evidence or an explicit 'No' leads to classification as $\leq$ 45 ml/day.

Query: Consumption of fortified wine $\leq$ 90ml per day

- **Reference Fields:** 1608 (Average weekly fortified wine intake), 4451 (Average monthly fortified wine intake)
- **Selected Fields:** 26151 (Fortified wine), 100720 (Fortified wine intake), 1608 (Average weekly fortified wine intake), 4451 (Average monthly fortified wine intake), 20414 (Frequency of drinking alcohol)
- **Rule Rationale:** Priority-based evaluation: (1) If a 24-hour fortified-wine ml estimate exists, classify based on that ($\leq$ 90 ml $\Rightarrow$ 1). (2) Else if yesterday's glass count exists, convert to ml and

classify. (3) Else if weekly average exists, convert to ml/day and classify. (4) Else if monthly average exists, convert and classify. (5) Else if participant reports 'Never' drinking, infer 0 ml/day and classify as $\leq$ 90. If none of the sources provide usable information, the overall rule returns 0 (value_if_false) to avoid falsely labeling as low consumption; alternatively users may set such cases to missing during downstream analyses. Missing-code handling: in preprocessing, UKB missing codes (-1, -3, -10, -818) must be converted to null so that the equality checks to '*' reflect missingness.

B.2. Sedentary Behavior Replication

Query: Sum of time spent watching TV or using computer less than 2 hours per day

- **Reference Fields:** 1070 (Time spent watching television (TV)), 1080 (Time spent using computer)
- **Selected Fields:** 1070 (Time spent watching television (TV)), 1080 (Time spent using computer), 40031 (Sedentary - Day hour average), 40043 (Sedentary - Day average), 1120 (Weekly usage of mobile phone in last 3 months)
- **Rule Rationale:** The rule returns 1 (true) when any available data source indicates combined TV+computer time is less than 2 hours/day by the prioritized hierarchy: (A) direct combined self-report (total_tv_computer_hours) ¡ 2; (B) single self-report (TV or computer) ¡ 2 when the other is missing; (C) accelerometer-derived evening sedentary hours (from 40031) ¡ 2 as an objective evening-window proxy; (D) derived total sedentary hours/day from 40043 ¡ 2 as a broader objective fallback; (E) conservative mobile-phone-hours/day from 1120 ¡ 2 as final fallback. Missing special codes (-1,-3,-10) are treated as NULL and do not count toward numeric comparisons. If none of the conditions can be evaluated as true (including the case that no data are available), the rule resolves to 0 to provide a defined output for downstream analyses. For transparency, analysts may wish to additionally record which source produced the classification (primary vs fallback) and to perform sensitivity analyses excluding the weakest proxies (accelerometry and phone proxies).

B.3. GAD-7 Replication

Item 1. **Query:** Feeling nervous, anxious, or on edge

- **Reference Fields:** 20506 (Recent feelings or nervousness or anxiety)
- **Selected Fields:** 28735 (Feeling anxious, nervous or on edge over the last 2 weeks), 30484 (Frequency of feeling nervous, anxious or on edge in last 2 weeks), 29058 (Recent feelings or nervousness or anxiety), 20506 (Recent feelings or nervousness or anxiety), 23045 (Very nervous mood over last week), 2070 (Frequency of tenseness / restlessness in last 2 weeks), 1970 (Nervous feelings)
- **Rule Rationale:** Selected fields (ordered by priority and reason):

  1) 28735 — "Feeling anxious, nervous or on edge over the last 2 weeks" (n=195,605). Chosen as primary because it directly matches the wording, timeframe (last 2 weeks, same as GAD-7), and has the largest respondent count among exact matches.

  2) 30484 — "Frequency of feeling nervous, anxious or on edge in last 2 weeks" (n=179,118). Direct match of wording/timeframe from a different online module; used as the first fallback when 28735 is missing.

  3) 29058 — "Recent feelings or nervousness or anxiety" (n=170,619). Direct mental well-being item matching wording/timeframe; used as the next fallback.

  4) 20506 — "Recent feelings or nervousness or anxiety" (n=157,235). Same content but different encoding (1-4). Included as another fallback; will be remapped to 0-3.

  5) 23045 — "Very nervous mood over last week" (n=211,849). Different timeframe (last week vs last 2 weeks) and different response scale (1-5), but semantically highly related. Included as a tertiary fallback when all exact 2-week items are missing.

  6) 2070 — "Frequency of tenseness / restlessness in last 2 weeks" (n=501,274). This asks about tenseness/restlessness (closely related symptom

domain and same 2-week timeframe). Included as a lower-priority fallback because wording is not identical but clinically relevant and high coverage.

7) 1970 — "Nervous feelings" (binary lifetime/trait on touchscreen; n=501,278). This is a trait / lifetime question (not 2-week frequency) but directly asks about being a 'nervous person'. Included only as the last-resort fallback to maximize coverage; mapped conservatively to indicate presence vs absence.

Create a single numeric value per participant equal to (priority_weight + mapped_item_score) for each candidate field; take the max across fields so that (because weights are set in decreasing priority) the highest-priority available response is selected even when lower-priority responses have higher raw scores.

Item 2. **Query:** Not being able to stop or control worrying

- **Reference Fields:** 20509 (Recent inability to stop or control worrying)
- **Selected Fields:** 28736 (Not being able to stop or control worrying over the last 2 weeks), 29059 (Recent inability to stop or control worrying), 30485 (Frequency of not being able to stop or control worrying in last 2 weeks), 20509 (Recent inability to stop or control worrying), 20537 (Frequency of difficulty controlling worry during worst period of anxiety), 20539 (Frequency of inability to stop worrying during worst period of anxiety)
- **Rule Rationale:** Harmonise each field to the canonical 0-3 GAD-7 item scale, map all known UKB missing codes and 'prefer not to answer'/'do not know' variants to null, then compute the maximum non-null value across the harmonised item instances. Max is chosen because (a) it returns a single interpretable 0-3 score, (b) it is robust to missingness (any available instance yields a value), and (c) it is conservative in capturing the highest reported symptom frequency across assessments/timeframes—useful for analyses where presence/severity matters. The aggregation uses only the selected fields to avoid dilution by non-equivalent items.

Item 3. **Query:** Worrying too much about different things

- **Reference Fields:** 20520 (Recent worrying too much about different things)
- **Selected Fields:** 29060 (Recent worrying too much about different things), 20520 (Recent worrying too much about different things), 28737 (Little interest or pleasure in doing things over the last 2 weeks), 29058 (Recent feelings or nervousness or anxiety), 20506 (Recent feelings or nervousness or anxiety), 30484 (Frequency of feeling nervous, anxious or on edge in last 2 weeks)
- **Rule Rationale:** Compute the (weighted) average of all available harmonised sources mapped to the 0-3 scale. Primary exact-match field (29060) is duplicated to weight it more heavily: when present it dominates the average. The final aggregated numeric value is mapped to the integer 0-3 via thresholds: $\geq 2.5 \rightarrow 3$, $\geq 1.5 \rightarrow 2$, $\geq 0.5 \rightarrow 1$, else 0. This yields a single item score consistent with the original 0-3 scoring, while maximising data coverage through inclusion of equivalent items and alternate codings (which are harmonised to 0-3). Missing/opt-out codes are mapped to null so they are ignored in the averaging. If no selected fields are available, the aggregation yields null (no data).

Item 4. **Query:** Trouble relaxing

- **Reference Fields:** 20515 (Recent trouble relaxing)
- **Selected Fields:** 29061 (Recent trouble relaxing), 20515 (Recent trouble relaxing), 29062 (Recent restlessness), 20516 (Recent restlessness), 2070 (Frequency of tenseness / restlessness in last 2 weeks), 29058 (Recent feelings or nervousness or anxiety)
- **Rule Rationale:** All selected fields are transformed to the same 0-3 scale and missing/opt-out codes are mapped to null (so they are ignored by aggregation). The aggregation 'max' then returns the highest non-null value among these harmonized inputs. This implements a prioritized fallback implicitly (direct item responses are identical or aligned to the scale; related items will only influence the result when the direct item is absent). Using 'max' minimizes the

chance of returning an artificially low score when one instrument captured higher symptom severity; it favors capturing present symptom severity across instruments when direct item is missing.

Item 5. **Query:** Being so restless that it is hard to sit still

- **Reference Fields:** 20516 (Recent restlessness)
- **Selected Fields:** 29062 (Recent restlessness), 20516 (Recent restlessness), 29009 (Recent changes in speed/amount of moving or speaking), 20518 (Recent changes in speed/amount of moving or speaking), 2070 (Frequency of tenseness / restlessness in last 2 weeks), 120111 (Moving or speaking slower or faster than usual over the last two weeks)
- **Rule Rationale:** Compute the average of all non-missing standardized 0-3 item scores from multiple questionnaire instances and wordings that capture 'restlessness' or closely overlapping fidgety/movement symptoms. Mapping rules convert each instrument's coding to the common 0-3 scale; missing/prefer-not-to-answer/do-not-know codes are mapped to null and thus excluded from the average. Using the average (ignoring nulls) provides an integrated single score between 0 and 3 that maximizes participant inclusion when some questionnaire versions are missing.

Item 6. **Query:** Becoming easily annoyed or irritable

- **Reference Fields:** 20505 (Recent easy annoyance or irritability)
- **Selected Fields:** 29063 (Recent easy annoyance or irritability), 20505 (Recent easy annoyance or irritability), 20494 (Felt irritable or had angry outbursts in past month), 20422 (More irritable than usual during worst period of anxiety), 4653 (Ever highly irritable/argumentative for 2 days), 29050 (Ever had period extreme irritability), 20502 (Ever had period extreme irritability), 1940 (Irritability), 28726 (Currently suffering from problems relating to mood, anxiety and emotions), 29049 (Ever had period of mania / excitability), 29057 (Problems caused by manic or irritable periods - aspects of life were affected), 29056 (Problems caused by manic or irritable periods - treatment was required)
- **Rule Rationale:** If encoded_priority_max exists (≥ 0) we derive the integrated 0-3 score by taking encoded_priority_max modulo 10 (because each transformed field was encoded as offset + base_score with offsets being multiples of 10). This implements the fallback: the highest-priority available field determines the encoded value (via offset), and modulo extracts the original 0-3 score. If encoded_priority_max is missing (no relevant field available or all mapped to null), the rule returns NULL. All UKB missing codes (e.g., -3, -818, -1, -121, -4) were explicitly mapped to null in the categorical mappings above so they are treated as missing and will not contribute to encoded_priority_max.

Item 7. **Query:** Feeling afraid as if something awful might happen

- **Reference Fields:** 20512 (Recent feelings of foreboding)
- **Selected Fields:** 29064 (Recent feelings of foreboding), 20512 (Recent feelings of foreboding), 29058 (Recent feelings or nervousness or anxiety), 20506 (Recent feelings or nervousness or anxiety), 28735 (Feeling anxious, nervous or on edge over the last 2 weeks), 30484 (Frequency of feeling nervous, anxious or on edge in last 2 weeks), 29060 (Recent worrying too much about different things), 20520 (Recent worrying too much about different things), 29059 (Recent inability to stop or control worrying), 20509 (Recent inability to stop or control worrying)
- **Rule Rationale:** Aggregate (max) across harmonised 0-3 values drawn from the direct item, its assessment-centre equivalent (transformed), and conceptually-close anxiety items (and their assessment-centre equivalents, transformed). Max is chosen to capture the highest reported severity across available sources (conservative approach) and to provide a single numeric value when any source is present. Missing/opt-out codes are mapped to null before aggregate so they do not affect the result.

B.4. RSD-4 Replication

Item 1. **Query:** Frequency of depressed mood in last 2 weeks

- **Reference Fields:** 2050 (Frequency of depressed mood in last 2 weeks)
- **Selected Fields:** 2050 (Frequency of depressed mood in last 2 weeks), 28738 (Feeling down, depressed or hopeless over the last 2 weeks), 30487 (Frequency of feeling down, depressed or hopeless in last 2 weeks), 29003 (Recent feelings of depression), 120105 (Feeling down, depressed, or hopeless over the last two weeks), 20510 (Recent feelings of depression)
- **Rule Rationale:** Compute the maximum harmonized item value across equivalent 'feeling down/depressed/hopeless in last 2 weeks' questions. Each field has been transformed to the canonical 1-4 scale (1=Not at all, 4=Nearly every day) with explicit mapping of UKB missing codes to null. 'Max' was chosen to capture the highest reported recent frequency across different instruments (a conservative approach towards identifying recent depressed mood when multiple assessments exist).

Item 2. **Query:** Frequency of unenthusiasm/disinterest in last 2 weeks

- **Reference Fields:** 2060 (Frequency of unenthusiasm / disinterest in last 2 weeks)
- **Selected Fields:** 2060 (Frequency of unenthusiasm / disinterest in last 2 weeks), 20514 (Recent lack of interest or pleasure in doing things), 28737 (Little interest or pleasure in doing things over the last 2 weeks), 30486 (Frequency of having little interest or pleasure in doing things in last 2 weeks), 120104 (Little interest or pleasure in doing things over the last two weeks), 29002 (Recent lack of interest or pleasure in doing things)
- **Rule Rationale:** If the aggregated 'rds4_value' (the max of transformed equivalent items) is within the valid 1-4 range then return that value as the integrated Item 2 score. If no selected source provides a valid response (all are NULL / missing / prefer-not-to-answer), return NULL. Returning the maximum across available equivalent items preserves severity information and maximizes coverage because it accepts any valid contribution from the prioritized set of fields.

Item 3. **Query:** Frequency of tenseness/restlessness in last 2 weeks

- **Reference Fields:** 2070 (Frequency of tenseness / restlessness in last 2 weeks)
- **Selected Fields:** 2070 (Frequency of tenseness / restlessness in last 2 weeks), 30484 (Frequency of feeling nervous, anxious or on edge in last 2 weeks), 29062 (Recent restlessness), 20516 (Recent restlessness), 20506 (Recent feelings or nervousness or anxiety), 28735 (Feeling anxious, nervous or on edge over the last 2 weeks), 29061 (Recent trouble relaxing), 20515 (Recent trouble relaxing), 20505 (Recent easy annoyance or irritability)
- **Rule Rationale:** Map each selected source to a common 1-4 scale representing frequency in the past 2 weeks (1=Not at all ... 4=Nearly every day). Use aggregation_type 'max' to select the highest available mapped frequency across sources for each participant. This preserves the construct (frequency of tenseness/restlessness) while maximizing data coverage across assessment branches and questionnaires; it treats multiple parallel items as fallbacks and prioritizes the strongest recent report.

Item 4. **Query:** Frequency of tiredness/lethargy in last 2 weeks

- **Reference Fields:** 2080 (Frequency of tiredness / lethargy in last 2 weeks)
- **Selected Fields:** 2080 (Frequency of tiredness / lethargy in last 2 weeks), 20519 (Recent feelings of tiredness or low energy), 29005 (Recent feelings of tiredness or low energy), 120107 (Feeling tired or having little energy over the last two weeks), 30575 (Frequency of fatigue over the last two weeks), 30568 (Feelings of tiredness during waking time)
- **Rule Rationale:** After mapping each field to a common 1-4 scale, 'max' captures the highest reported frequency/severity of tiredness across sources. The post-aggregation conditions simply return the aggregated numeric severity when it equals 1-4, and null otherwise (catch-all). This

approach increases the number of participants with a usable score while remaining transparent about transformations and handling of missing/special codes.