SLEEP AS A WINDOW: AUDITABLE LLM DECODING OF AGING SIGNATURES FROM INTEGRATED SLEEP PROFILES

Anonymous authors
Paper under double-blind review

ABSTRACT

Biological aging progresses at different rates across organs, creating patterns of vulnerability that chronological age alone cannot capture. Sleep offers an information-rich window into this heterogeneity, with each participant's full clinical record providing context for its interpretation. Existing sleep-age models usually yield a single estimate, whereas large language model (LLM) outputs can be difficult to trace to patient evidence. We introduce an evidence-auditable LLM framework that decodes integrated sleep profiles into an overall sleep age and ten organ-resolved age representations. A frozen registry curated by clinical sleep experts and trace-validated reports preserve input provenance. The framework also audits the accompanying physiological chains for major contradictions. We evaluated the framework in 9,123 participants from SHHS, MrOS, and the real-world XH longitudinal cohort. After adjustment for chronological age and conventional risk factors, mortality HRs per year of sleep gap were 1.054 (95% CI 1.031–1.079) in SHHS and 1.030 (1.008–1.052) in MrOS. Ten-year mortality was higher in the top than in the bottom gap decile (SHHS 25.4% versus 8.9%, MrOS 25.5% versus 13.0%). Models combining chronological age with endpoint-matched organ gaps outperformed those using non-matched gaps in C-index (mean proportion outperformed: SHHS 92%, MrOS 88%, XH 84%). This pattern supports LLM decoding of organ-specific aging signatures. In XH, overall sleep gaps tracked incident disease and cumulative disease burden, and endpoint-matched gaps added predictive value beyond multiple clinical baseline models. Together, these findings position integrated sleep profiles as an auditable window onto systemic aging heterogeneity and its future health consequences.

1 INTRODUCTION

Physiological reserve and disease susceptibility vary substantially among individuals of the same chronological age and across organs within an individual. Multi-organ studies have begun to quantify this divergence, showing that organ-specific age gaps relate to organ-linked disease and mortality (Tian et al., 2023; Oh et al., 2023). Sleep provides a particularly revealing observation window: overnight recordings capture coordinated physiological activity, including neural, respiratory, cardiovascular, and autonomic signals, under a shared behavioral state. Sleep-EEG and polysomnography models can estimate chronological age (Sun et al., 2019; Brink-Kjaer et al., 2022), and their age gaps have been associated with mortality (Paixao et al., 2020; Brink-Kjaer et al., 2022) and dementia (Ye et al., 2020). In clinical practice, however, nocturnal physiology is interpreted alongside demographics, lifestyle, examination, questionnaires, medical history, and medication. We refer to this person-level record as an *integrated sleep profile*. Phenome-wide links between sleep characteristics and traits across physiological systems further motivate this joint view (Kohn et al., 2025).

An integrated sleep profile raises a question that extends beyond recovering chronological age: **can it reveal both a shared signature of whole-person aging and organ-specific aging patterns?** The overall sleep age summarizes the shared aging signature expressed across nocturnal physiology and its clinical context, but a single scalar cannot localize organ-specific vulnerability. Conversely, a vector of organ-resolved ages is scientifically meaningful only if its internal differences correspond to future

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

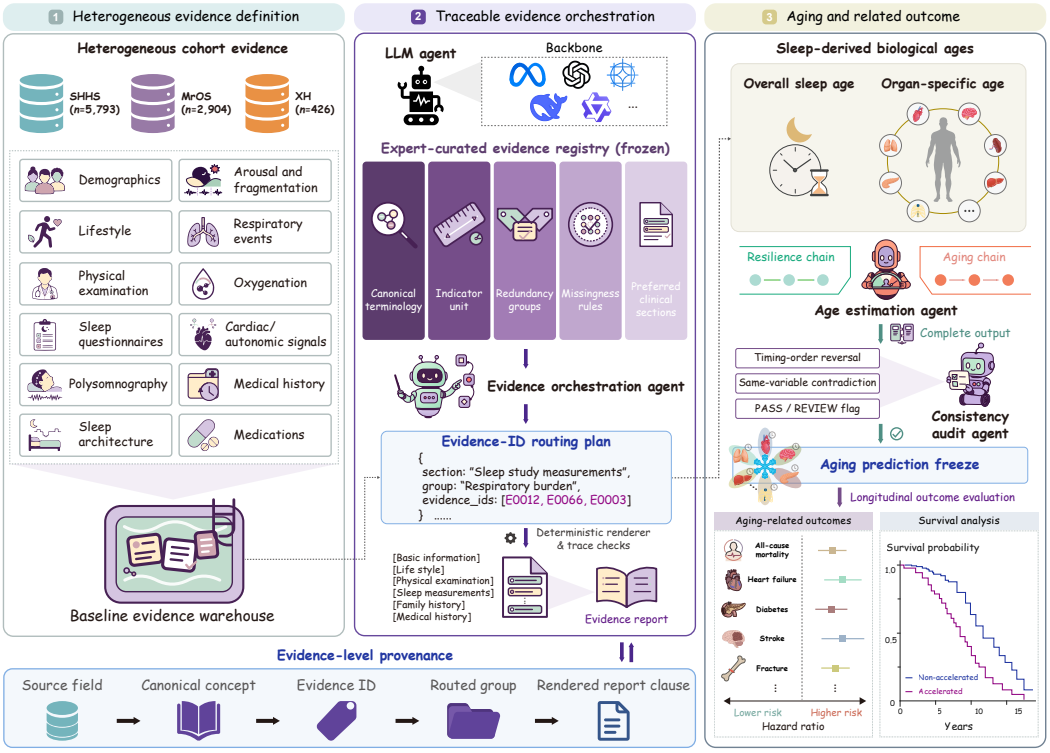

Figure 1: Overview of evidence-auditable aging-representation decoding from integrated sleep profiles. Clinical sleep experts curate a frozen registry spanning sleep measurements, demographics, lifestyle, examination, medical history, questionnaires, and medications. The Evidence Orchestration Agent selects typed Evidence IDs. Registry-guided validation, exact rendering, and trace checks preserve the selected values, units, and provenance. An Age Estimation Agent decodes one overall sleep age and ten organ-resolved ages. The Consistency Audit Agent independently screens the complete frozen estimator output for prespecified serious contradictions. Longitudinal outcomes are subsequently used to evaluate the unchanged age representation.

organ-linked outcomes. We therefore test organ specificity by asking whether endpoint-matched gaps discriminate the corresponding outcomes better than non-matched gaps in the same participants.

Answering this question across cohorts requires more than assembling a common feature table. SHHS, MrOS, and XH differ in variable names, units, measurement coverage, questionnaires, examinations, and medical-history records. Reducing them to their feature intersection discards informative cohort-specific evidence, whereas unrestricted narrative synthesis can obscure where a value came from or how it was represented. Health-report prompting has produced overall and organ-specific age estimates at population scale (Li et al., 2025). At this scale, explicit field-level provenance becomes important when heterogeneous cohort variables are translated into generated representations. This challenge is consequential because clinical LLMs can be sensitive to information structure and can mishandle numerical evidence even when their medical knowledge is strong (Singhal et al., 2023; Hager et al., 2024).

We address this challenge with an evidence-auditable multi-agent framework built on LLMs (Figure 1). Clinical sleep experts curate a frozen registry that links each admissible variable to an Evidence ID, canonical concept, unit, report section, and source. The Evidence Orchestration Agent routes admissible Evidence IDs. Registry-guided plan validation, exact rendering, and trace checks then produce a source-linked report. The Age Estimation Agent decodes the report into one overall sleep age and ten organ-resolved ages, together with inspectable physiological chains. The Consistency Audit Agent reads the complete estimator output and assigns a PASS or REVIEW flag when it detects a prespecified serious contradiction. This architecture preserves cohort-specific evidence in traceable reports for decoding overall sleep and organ-resolved aging signatures.

Our contributions are as follows:

- **Evidence-auditable decoding.** We link heterogeneous clinical variables to source-traceable LLM inputs through an expert-curated evidence registry. Separate checks of report fidelity and physiological consistency make the evidence supplied to the estimator and its generated interpretations independently inspectable.
- **Shared and organ-resolved aging signatures.** We assess the clinical relevance of the eleven-age representation through prespecified organ–outcome mappings and comparisons among all gaps within each endpoint. This tests whether organ assignment carries prognostic information alongside the overall sleep-aging signal.
- **Multi-cohort longitudinal evaluation.** We examine the representations across two community cohorts and a real-world follow-up cohort, spanning mortality, incident organ disease, and health deficits. Clinical risk models and supervised age predictors provide complementary comparisons of their prognostic value.

2 Related Work

Aging representations from sleep assessment. Sleep-derived representations span handcrafted EEG brain-age models, deep raw-PSG age regression, and recent multimodal PSG foundation models (Sun et al., 2019; Brink-Kjaer et al., 2022; Thapa et al., 2026; Bilal et al., 2026). Their reported targets include mortality and incident dementia (Paixao et al., 2020; Brink-Kjaer et al., 2022; Sun et al., 2026). Recent foundation models instead support future-disease prediction and outcome-linked risk stratification (Thapa et al., 2026; Bilal et al., 2026). Personal-health LLMs have also been adapted to wearable sleep and fitness data for coaching and sleep-quality tasks (Khasentino et al., 2025), whereas health-report prompting has produced overall and organ-specific age estimates at population scale (Li et al., 2025). Transparent AI in sleep settings has further shown the value of retaining interpretable physiological evidence across flexible monitoring scenarios (Hu et al., 2025). The present study connects these directions by decoding integrated sleep profiles into one overall sleep age and ten organ-resolved ages with traceable evidence provenance. Longitudinal evaluation tests the prognostic relevance of these representations, with sleep measurements and their accompanying clinical context considered jointly. Supplementary Section A.1 compares the inputs, supervision, output granularity, and validation of these approaches.

Age gaps depend on how the underlying age estimate was constructed. Following Li et al. (2025), we define each LLM gap directly as predicted age minus chronological age. The estimator is not fitted within a study cohort to minimize chronological-age error, so we preserve this definition and include chronological age in downstream association and incremental-value models. For conventional sleep-feature regressors trained against chronological age, we apply the age-bias correction of Smith et al. (2019). Both constructions quantify deviation from chronological age and are compared using the same participants and follow-up within each endpoint. Their formal relation and interpretation are detailed in Supplementary Section A.

Evidence-grounded medical LLM and agent systems. Strong performance on medical knowledge tasks (Singhal et al., 2023) has encouraged the use of LLMs in clinical workflows. These workflows, however, expose sensitivity to information order, instruction following, and numerical evidence (Hager et al., 2024), alongside difficulties with incomplete clinical context (Li et al., 2024). Static benchmarks can further overstate reliability when clinical knowledge changes or evaluation data leak into training (Wang et al., 2026a). Agentic approaches structure clinical reasoning through complementary roles (Kim et al., 2024), drawing on broader methods that interleave reasoning with actions and tool-grounded critique (Yao et al., 2023; Gou et al., 2024). Their reliability still requires evaluation across multistep workflows, given the variation in reasoning, decision-making, and instruction following observed in agent benchmarks (Liu et al., 2024). In health applications, longitudinal wearable analysis combines multistep reasoning with expert error evaluation (Merrill et al., 2026), while health-report prompting produces overall and organ-specific age estimates at population scale (Li et al., 2025). These developments motivate modular clinical reasoning and highlight the need to retain explicit links between patient-specific representations and their heterogeneous source evidence.

Maintaining these links requires checking generated content beyond the valid structure enforced by grammar-constrained decoding (Geng et al., 2023). Claim-level evaluation makes content assessable

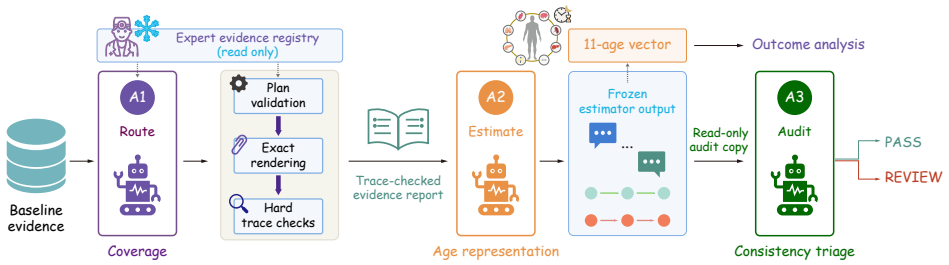

Figure 2: Participant-level evidence-auditable decoding protocol. An expert-curated, read-only registry defines admissible clinical concepts. The Evidence Orchestration Agent routes Evidence IDs. Registry-guided validation, exact rendering, and trace checks preserve the selected values, units, and provenance. The Age Estimation Agent produces one overall sleep age and ten organ-resolved ages for outcome analysis. The Consistency Audit Agent independently examines the complete frozen estimator output and assigns PASS or REVIEW for prespecified serious contradictions.

by decomposing responses into atomic statements (Min et al., 2023) or reference-checked claim triplets (Hu et al., 2024). Explicit citations link claims to sources (Gao et al., 2023b), while evidence-based revision addresses unsupported text (Gao et al., 2023a). Retrieval-guided critique brings this use of external evidence into generation itself (Asai et al., 2024). Clinical evaluation extends these checks to unsafe responses and plausible-sounding medical hallucinations (Han et al., 2024; Pandit et al., 2025). Contradiction detection adds a complementary check of whether statements within the same output are mutually compatible (Mündler et al., 2024). Detecting such problems, however, does not guarantee that model-generated revisions are reliable. Intrinsic self-correction can fail without external feedback (Huang et al., 2024), whereas sound external verification can outperform unconstrained self-critique (Stechly et al., 2025). Even an internally consistent reasoning chain need not faithfully reflect how a prediction was produced (Turpin et al., 2023). For aging representations, this distinction motivates separate assessment of input fidelity and generated physiological interpretations. Our framework preserves patient values, units, and labels through expert-defined Evidence IDs and trace-checked reports before estimation. The Consistency Audit Agent independently examines the complete output for physiological contradictions without altering the age estimates.

3 Evidence-Auditable Decoding of Integrated Sleep Profiles

Expert-Curated Evidence Registry. Clinical sleep experts reviewed the cohort dictionaries and specified a frozen registry for each admissible variable: Evidence ID, provenance, canonical label, unit, report section, redundancy group, and missingness policy. The registry spans sleep measurements, demographics, lifestyle, examination, questionnaires, medical history, and medications. It provides the shared clinical semantics for cohort-specific evidence, while patient values remain in the corresponding evidence pools. Outcome and follow-up variables are reserved for longitudinal evaluation.

Evidence Orchestration Agent and Trace Validation. The Evidence Orchestration Agent receives the available registry entries and returns a section-wise Evidence-ID plan. A registry-based controller validates the identifiers and resolves registered alternatives. The renderer then transfers the corresponding source value, canonical label, and unit verbatim. A trace records every planned, rendered, omitted, and unavailable item. Together, these operations connect expert-defined clinical concepts to participant-level evidence through explicit source links.

Before decoding, automated checks align the plan, rendered report, registry, and evidence pool. Plans with an extra or missing ID, label mismatch, changed value or unit, duplicate, section violation, or schema failure are returned for correction or rerendering. These checks establish report fidelity and retain a source-resolving record for each discrepancy.

Age Estimation Agent. The Age Estimation Agent receives the traceability-checked profile and jointly returns an 11-component age representation: one overall sleep age and ten organ-resolved

ages for the cardiovascular, neurological, respiratory, metabolic, skeletal, hepatic, renal, immune, urinary, and digestive systems. For participant i, the signed gap for age estimate k is $g_{i,k} = \widehat{a}_{i,k} - a_i$, where k indexes the overall sleep age or one of the ten organ-resolved ages, and a_i is chronological age. Each age estimate is accompanied by aging and resilience chains and a narrative justification, exposing the generated physiological interpretation for direct audit.

Consistency Audit Agent. After schema validation, the complete estimator output is frozen. Its vector of one overall sleep age and ten organ-resolved ages is used unchanged in longitudinal analyses. The Consistency Audit Agent independently examines the full frozen output and applies two prespecified hard-consistency checks. The first identifies an impossible order defined by within-night sleep-measurement timing. The second identifies a named variable that is simultaneously unknown and assigned a state, or assigned explicitly opposing states. Outputs meeting either criterion receive a REVIEW flag. The remaining outputs receive PASS. The audit annotation and structured issue type are retained for clinical inspection and do not alter the frozen age estimates. Longitudinal outcomes are introduced only in the subsequent statistical evaluation.

Figure 2 summarizes the participant-level communication and data flow through evidence orchestration, age estimation, and consistency audit.

4 MULTI-COHORT EVALUATION

4.1 COHORTS AND ENDPOINTS

SHHS contributed 5,793 community participants with baseline polysomnography and prospective mortality and cardiocerebrovascular outcomes (Quan et al., 1997). MrOS contributed 2,904 older men with mortality, cardiovascular, neurological, and fracture follow-up (Orwoll et al., 2005; Dean et al., 2016). XH contributed 426 participants from a real-world clinical sleep cohort with longitudinal diagnosis records. Only the baseline integrated sleep profile entered the model. Outcomes and follow-up were used solely for evaluation.

The cohorts differ in ascertainment, chronological-age structure, and endpoint capture, so each was analyzed separately. SHHS and MrOS provide mortality and cardiovascular, neurological, and fracture outcomes, whereas XH provides incident diagnosis categories after endpoint-specific exclusion of prevalent disease. This heterogeneity tests transfer across distinct clinical and research settings. Cohort characteristics are summarized in Figure D.1 and Table D.1.

4.2 LONGITUDINAL AND ORGAN-RESOLVED EVALUATION

We assessed chronological-age anchoring and calibration, longitudinal association, and organ resolution. Cox models estimated associations between each one-year larger prespecified gap and time-to-event outcomes. In SHHS and MrOS, we fitted four nested specifications: gap alone, gap plus chronological age, additional adjustment for sex where variable, race, BMI, smoking, and alcohol, and additional adjustment for baseline cardiovascular disease, with fracture history also included in MrOS. Missing SHHS and MrOS covariates were handled by five multiple imputations. In XH, the first-new-diagnosis Cox analysis used three nested specifications: gap alone, gap plus chronological age, and gap plus chronological age and log-transformed baseline diagnosis burden. Cumulative diagnosis burden used the same adjustment sequence in negative-binomial models with log follow-up as an offset.

The XH first-new-diagnosis composite and cumulative-burden analysis used a broad eligible outcome set containing 244 clinically filtered, organ-mapped diagnosis categories with at least one post-index event. For cumulative burden, each participant's outcome was the number of distinct eligible categories that first appeared after baseline. The separate organ-resolved analyses used 20 clinically representative outcomes with at least ten events after excluding participants who had the corresponding diagnosis at baseline.

The overall sleep gap was prespecified for all-cause death, the first XH diagnosis, and cumulative XH burden. Each organ-linked endpoint was paired before analysis with its corresponding organ-resolved gap, for example the cardiovascular gap for cardiovascular outcomes, the neurological gap for cerebrovascular outcomes, and the skeletal gap for fractures. The complete XH mapping

covers the 20 organ-linked outcomes. HRs are reported per one-year larger gap with 95% confidence intervals, and multiplicity was controlled within cohort and model specification.

Kaplan–Meier analyses compared nearly equal thirds, the bottom and top 10% by overall sleep-gap rank, and a fixed five-year split. Ties were ordered by participant identifier. Curves used all available follow-up, with log-rank tests comparing the first and last strata. Adjusted associations were estimated separately using continuous-gap Cox models. Organ resolution was tested by adding each of the eleven gaps separately to a Cox model using only chronological age. Within each endpoint, every gap was evaluated in the same participants and cross-validation folds. Repeated five-fold out-of-fold Harrell C-index quantified improvement over that baseline. For each endpoint, the prespecified matched gap was compared with the mean of the ten non-matched gaps. Ranks were retained as descriptive diagnostics.

4.3 Comparative and Aging-Relevance Validation

We compared seven LLM backbones on the same 426 XH profiles and eleven-age schema, using the adjusted first-new-diagnosis association and endpoint-matched C-indices across 20 organ-linked outcomes. We focused on open-weight backbones because they support deployment in institution-controlled computing environments, an important option for protecting clinical data.

Six conventional regressors learned chronological age from a common 18-variable subset in healthy SHHS participants and transferred to MrOS and XH. Their gaps received prespecified age-bias correction fitted before outcome linkage (Smith et al., 2019). LLM gaps were defined directly as estimated minus chronological age (Li et al., 2025). Across 48 endpoint evaluations, we evaluated the overall sleep gap for every endpoint and an endpoint-matched readout that used the overall sleep gap for mortality and the corresponding organ-resolved gap for each organ-linked outcome. Every method used the same participants and follow-up within each endpoint.

To test whether the overall sleep gap captured aging-related information, we related it to SHHS and MrOS frailty indices and XH clinical-deficit burden using rank correlations and age- and sex-adjusted standardized associations. In XH, we also measured C-index changes after adding the endpoint-matched gap to three increasingly informative clinical models. Full specifications are in the Supplementary Information.

4.4 Consistency Audit Agent Evaluation

An independent model-based developmental reference annotated all 426 XH outputs using the same two prespecified hard-consistency errors and without access to the audit decisions. We report the review rate, Wilson 95% intervals, and a two-sided Fisher exact test. These annotations characterize development-stage triage behavior. Independent annotation by clinical sleep experts remains a separate validation step.

5 Results

5.1 Overall Sleep Gap Was Associated with All-Cause Mortality

SHHS overall sleep-gap groups showed graded survival stratification, with Kaplan–Meier 10-year mortality increasing across low, intermediate, and high rank-based tertiles (11.4%, 18.6%, and 24.0%, Figure 3a). The high versus low log-rank test yielded $P = 3.87 \times 10^{-26}$. A fixed five-year threshold yielded 11.2% versus 21.3% (Figure 3b), whereas the bottom and top deciles yielded 8.9% versus 25.4%, a 16.5-percentage-point difference (Figure 3c). The three views capture graded percentile stratification, an absolute threshold independent of percentile estimation, and separation at the distributional extremes. Supplementary Figures J.1 and K.1 show finer-grained SHHS strata and the corresponding MrOS threshold analyses, respectively.

Continuous-gap Cox models then quantified the associations after full conventional-risk adjustment (Figure 4). Each one-year larger overall sleep gap was associated with all-cause mortality in SHHS (1,302 deaths among $n = 5{,}776$, HR 1.054, 95% CI 1.031–1.079) and MrOS (2,039 deaths among $n = 2{,}466$, HR 1.030, 95% CI 1.008–1.052). The four nested adjustment trajectories, from gap-only through medical-history adjustment, are reported in Supplementary Figure C.1.

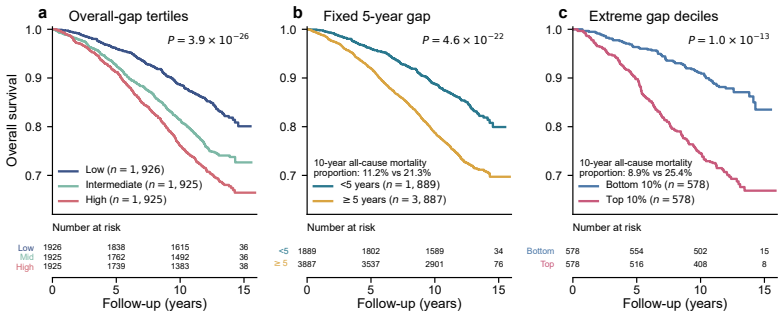

Figure 3: SHHS all-cause survival stratified by overall sleep gap over the full available follow-up ($n = 5{,}776$, 1,302 deaths). **a**, Three nearly equal groups by ascending gap rank. **b**, A fixed five-year gap threshold. **c**, Lowest versus highest 10% by gap rank ($n = 578$ each). Ties were ordered by participant identifier in a and c. Panels b and c report Kaplan–Meier 10-year mortality. Two-sided log-rank tests compare low versus high in a, < 5 versus ≥ 5 years in b, and bottom versus top decile in c. Risk tables report participants who remained event-free and uncensored immediately before each displayed time point.

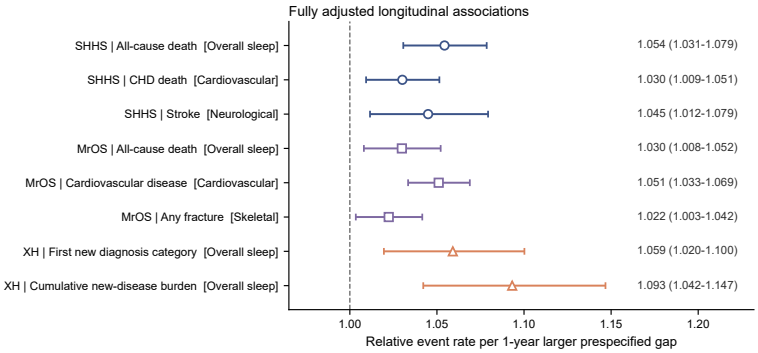

Figure 4: Multi-cohort longitudinal associations of the overall sleep gap and organ-resolved gaps. Points show fully adjusted relative event rates per one-year larger prespecified gap with 95% CIs. Estimates are HRs except XH cumulative burden, which is an incidence-rate ratio. SHHS models adjust for chronological age, sex, BMI, race, smoking, alcohol, and baseline CVD. MrOS models adjust for chronological age, BMI, race, smoking, alcohol, baseline CVD, and fracture history (male-only cohort). XH models adjust for chronological age and baseline diagnosis burden. Participants are the independent units.

5.2 Organ-Resolved Gaps Tracked Multi-Organ Outcomes

Fully adjusted organ-resolved gaps tracked cardiovascular, neurological, and skeletal outcomes. Cardiovascular-gap associations included CHD death, CVD death, and heart failure in SHHS and CHD, CVD, heart failure, and arrhythmia in MrOS. The neurological gap tracked SHHS stroke, and the skeletal gap tracked MrOS any and major osteoporotic fractures. Figure 4 shows representative associations. Supplementary Figure C.1 reports all endpoints and adjustment sets. All listed associations passed within-specification FDR correction.

In XH, the overall sleep gap was associated with the first new diagnosis from the 244-category eligible outcome set after adjustment for chronological age and baseline diagnosis burden (HR 1.059 per year, 95% CI 1.020–1.100). It was also associated with the cumulative number of distinct eligible diagnoses that first appeared after baseline (incidence rate ratio 1.093 per year, 1.042–1.147). Endpoint-matched discrimination across 20 clinically representative organ-linked outcomes with at least ten events is reported in Supplementary Figure C.2. The first-new-diagnosis and cumulative-burden estimates, together with the fitted participant-level burden curve, appear in Supplementary Figures C.4b–d and F.2.

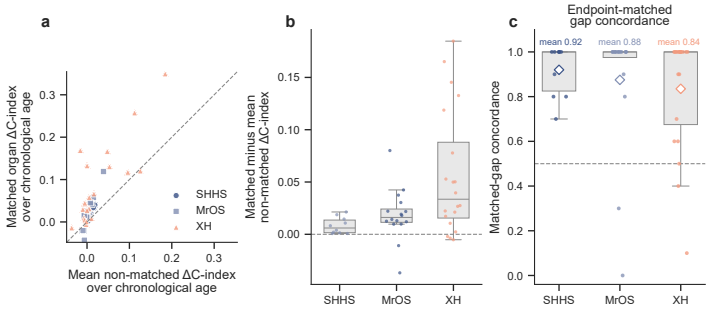

Figure 5: Organ correspondence in endpoint-specific discrimination beyond chronological age. **a**, Endpoint-level out-of-fold ΔC-index from adding the prespecified matched gap to chronological age versus the mean for the other ten gaps. **b**, Within-endpoint difference between these quantities. **c**, Fraction of alternative gaps with a lower ΔC-index, with ties counting one half. Each endpoint used identical participants and cross-validation folds. Boxes show the interquartile range and median, circles show endpoints, diamonds show cohort means, and the dashed line marks 0.5.

5.3 Outcome Associations Varied Across LLM Backbones

The XH comparison used one complete eleven-age output per participant from each backbone, fixed before outcomes were analyzed ($n = 426$). For the first-new-diagnosis composite, the fully adjusted HR per one-year overall sleep gap was 1.059 for Llama 3.3-70B and 1.002–1.042 for six alternatives. Across the 20 endpoint-matched organ outcomes, macro-average C-index was 0.668 for Llama 3.3-70B and 0.648 for the nearest comparator, Gemma 4 31B. Endpoint-level results are shown in Supplementary Figure F.1.

5.4 Organ Resolution and Conventional Age-Regression Baselines

Organ correspondence varied in magnitude but was consistent across cohorts. In SHHS, it was expressed primarily as rank consistency rather than a large absolute gain: the endpoint-matched gap improved C-index beyond chronological age by 0.014 on average, compared with 0.006 for the ten alternative gaps. This modest 0.008 mean advantage was positive for all ten endpoints, and the matched gap exceeded 92% of alternatives on average. Mean advantages were 0.018 in MrOS and 0.057 in XH, with corresponding concordance values of 0.88 and 0.84 (Figure 5). Full endpoint-by-gap results are shown in Supplementary Figure I.1.

We next compared outcome discrimination under each estimator's prespecified gap construction. All methods used identical participants and follow-up within each of 48 endpoint-specific evaluations. Mean gap-only C-index was 0.569 for the overall sleep gap and 0.615 for the endpoint-matched LLM representation. Among six traditional chronological-age regressors, extra-trees had the highest mean C-index (0.551). Outcome-level distributions and summary values are provided in Supplementary Figure C.3 and Table C.1, with input availability in Supplementary Figure E.1.

5.5 Incremental Discrimination and Deficit Accumulation

Within each XH endpoint, paired models used identical participants and cross-validation folds. Adding the endpoint-matched gap to the clinical model yielded a mean ΔC-index of 0.070 across 20 organ outcomes, with 18 of 20 point estimates above zero. After including the first-new-diagnosis composite, the macro-average C-index increases were 0.082, 0.088, and 0.067 relative to chronological age, chronological age plus sex, and the clinical model, respectively (Supplementary Table H.1).

The overall sleep gap also aligned with independently constructed baseline deficit measures. After adjustment for chronological age and sex where variable, standardized coefficients were 0.240 in SHHS, 0.183 in MrOS, and 0.368 for XH's clinical-deficit burden. Confidence intervals and rank correlations are reported in Supplementary Table E.1 and Figure C.4a, supporting that the overall sleep gap captures aging-related health variation beyond chronological age.

5.6 Traceability and Consistency Audit Agent Performance

All 9,123 outputs contained one overall sleep age and ten jointly returned organ-resolved ages. Registry and trace checks verified every rendered Evidence ID, label, value, and unit. Overall sleep age correlated with chronological age in SHHS, MrOS, and XH ($r = 0.978$, 0.934, and 0.971). Calibration and eleven-age distributions appear in Supplementary Figures D.2 and G.1.

The frozen Consistency Audit Agent returned PASS for 391 of 426 XH outputs and routed 35 (8.2%) for REVIEW. Independent developmental annotations, generated without access to the audit decisions, identified prespecified hard-consistency errors in 24/35 REVIEW outputs (68.6%) and 13/391 PASS outputs (3.3%). A two-sided Fisher exact test gave $P = 9.84 \times 10^{-22}$. This review rate yielded 64.9% sensitivity, 97.2% specificity, 68.6% positive predictive value, and 96.7% negative predictive value (Supplementary Figure L.1). The agent evaluated the physiological consistency of each complete output. The frozen eleven-age vector remained unchanged in all longitudinal analyses.

6 Discussion and Limitations

6.1 Interpreting the Aging Representations

Integrated sleep profiles yielded complementary shared and organ-resolved views of aging. The overall sleep gap was associated with mortality, deficit accumulation, and evolving disease burden. The eleven-age vector preserved variation that one estimate would collapse and made organ–outcome correspondence directly testable. The outputs are structured aging signatures of the complete profile, not interchangeable tissue ages.

Organ correspondence appeared in all three cohorts, with different effect magnitudes. SHHS showed a modest absolute advantage but consistently favorable ordering against alternative gaps, whereas MrOS and XH showed larger mean separation. Thus, organ resolution captures relative structure within the eleven-age vector despite differences in endpoints, cohorts, and evidence. Unlike conventional regressors that learn one chronological-age estimate from predefined features, the LLM returns shared and organ-resolved representations jointly. Comparisons in identical endpoint-specific samples showed that these gaps retained outcome-relevant structure beyond chronological age and conventional age-regression representations.

6.2 Auditability, Limitations, and Outlook

Auditability operates at two linked levels. Expert-defined evidence semantics and trace validation preserve the identity, value, and unit presented to the estimator. The Consistency Audit Agent then identifies major contradictions in the physiological interpretation. Separating source fidelity from interpretive coherence enables focused review without changing the prespecified age vector.

SHHS and MrOS provide a narrower range of organ-specific outcomes, whereas XH spans broader diagnoses and comorbidities. Differences in cohort composition, endpoint ascertainment, and evidence availability may influence effect sizes. The prospective design establishes longitudinal relevance, while repeated measurements are needed to resolve within-person trajectories and intervention responses. External evaluation can test portability, clinical sleep expert review can assess consistency auditing in clinical inspection, and intervention studies can ask whether gap changes identify modifiable trajectories. Open-weight backbones provide a practical path to institution-hosted deployment under local clinical-data governance.

7 Conclusion

This study establishes an evidence-auditable route from heterogeneous integrated sleep profiles to one overall and ten organ-resolved aging representations. Across three cohorts, the overall sleep gap stratified mortality and tracked deficit accumulation and disease burden, while matched gaps preferentially discriminated corresponding organ outcomes. Expert-defined evidence semantics preserve source fidelity, and consistency auditing exposes physiological interpretations for review. Linking traceable evidence to prospective aging signatures advances LLM-based aging assessment from opaque estimation toward physiologically structured, testable representation.

AI USE STATEMENT

For manuscript preparation, large language models were used for language editing and polishing.

ETHICS STATEMENT

SHHS and MrOS data were accessed with permission through the National Sleep Research Resource and analyzed under the applicable data-use requirements. The XH study received institutional ethics approval. The institution and approval identifier are withheld during anonymous review and will be provided in the final version. All report inputs were de-identified before model processing, with direct identifiers and other identity-revealing information removed.

REPRODUCIBILITY STATEMENT

The Supplementary Information provides detailed descriptions of the cohort and endpoint definitions, evidence-registry construction and report rendering, complete agent prompts and output schemas, age-gap definitions, model settings, statistical analyses, representative de-identified outputs, and sensitivity analyses.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024. https://proceedings.iclr.cc/paper_files/paper/2024/hash/25f7be9694d7b32d5cc670927b8091e1-Abstract-Conference.html.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Erhan Bilal, Matheus Lima Diniz Araujo, Kristen L. Beck, Catherine M. Heinzinger, Samer Ghosn, Carl Y. Saab, Nancy Foldvary-Schaefer, Jeffrey L. Rogers, and Reena Mehra. A foundation model for sleep-based risk stratification and clinical outcomes. *Nature Communications*, 17:7603, 2026. https://doi.org/10.1038/s41467-026-75326-9.

Andreas Brink-Kjaer, Eileen B. Leary, Haoqi Sun, M. Brandon Westover, Katie L. Stone, Paul E. Peppard, Nancy E. Lane, Peggy M. Cawthon, Susan Redline, Poul Jennum, et al. Age estimation from sleep studies using deep learning predicts life expectancy. *npj Digital Medicine*, 5(1):103, 2022. https://doi.org/10.1038/s41746-022-00630-9.

Dennis A. Dean, Ary L. Goldberger, Robert Mueller, Matthew Kim, Michael Rueschman, Daniel Mobley, Satya S. Sahoo, Chandan P. Jayapandian, Lu Cui, Michael G. Morrical, et al. Scaling up scientific discovery in sleep medicine: the national sleep research resource. *Sleep*, 39(5): 1151–1164, 2016. https://doi.org/10.5665/sleep.5774.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16477–16508, 2023a. https://doi.org/10.18653/v1/2023.acl-long.910.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488, 2023b. https://doi.org/10.18653/v1/2023.emnlp-main.398.

Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. Grammar-constrained decoding for structured NLP tasks without finetuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10932–10952, 2023. https://doi.org/10.18653/v1/2023.emnlp-main.674.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations*, 2024. https://openreview.net/forum?id=Sx038qxjek.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, and Daniel Rueckert. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 30(9):2613–2622, 2024. https://doi.org/10.1038/s41591-024-03097-1.

Tessa Han, Aounon Kumar, Chirag Agarwal, and Himabindu Lakkaraju. MedSafetyBench: Evaluating and improving the medical safety of large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. https://doi.org/10.52202/079017-1054.

Steve Horvath. DNA methylation age of human tissues and cell types. *Genome Biology*, 14(10): R115, 2013. https://doi.org/10.1186/gb-2013-14-10-r115.

Shuaicong Hu, Jian Liu, Yanan Wang, Cong Fu, Jichu Zhu, Huan Yu, and Cuiwei Yang. Transparent artificial intelligence-enabled interpretable and interactive sleep apnea assessment across flexible monitoring scenarios. *Nature Communications*, 16(1):7548, 2025. https://doi.org/10.1038/s41467-025-62864-x.

Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. Knowledge-centric hallucination detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6953–6975, 2024. https://doi.org/10.18653/v1/2024.emnlp-main.395.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations*, 2024. https://proceedings.iclr.cc/paper_files/paper/2024/hash/8b4add8b0aa8749d80a34ca5d941c355-Abstract-Conference.html.

Justin Khasentino, Anastasiya Belyaeva, Xin Liu, Zhun Yang, Nicholas A. Furlotte, et al. A personal health large language model for sleep and fitness coaching. *Nature Medicine*, 31:3394–3403, 2025. https://doi.org/10.1038/s41591-025-03888-0.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. MDAgents: An adaptive collaboration of LLMs for medical decision-making. In *Advances in Neural Information Processing Systems*, volume 37, 2024. https://doi.org/10.52202/079017-2522.

Sarah Kohn, Alon Diament, Anastasia Godneva, Raja Dhir, Adina Weinberger, Yotam Reisner, Hagai Rossman, and Eran Segal. Phenome-wide associations of sleep characteristics in the Human Phenotype Project. *Nature Medicine*, 31(3):1026–1037, 2025. https://doi.org/10.1038/s41591-024-03481-x.

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. MediQ: Question-asking LLMs and a benchmark for reliable interactive clinical reasoning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. https://doi.org/10.52202/079017-0908.

Yanjun Li, Qi Huang, Jin Jiang, Xusheng Du, Wenxin Xiang, Shiqi Zhang, Zean Pan, Liyuan Zhao, Yuyan Cui, Limei Ke, et al. Large language model-based biological age prediction in large-scale populations. *Nature Medicine*, 31:2977–2990, 2025. https://doi.org/10.1038/s41591-025-03856-8.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations*, 2024. https://proceedings.iclr.cc/paper_files/paper/2024/hash/e9df36b21ff4ee211a8b71ee8b7e9f57-Abstract-Conference.html.

Mike A. Merrill, Akshay Paruchuri, Naghmeh Rezaei, Geza Kovacs, Javier Perez, et al. Transforming wearable data into personal health insights using large language model agents. *Nature Communications*, 17:1143, 2026. https://doi.org/10.1038/s41467-025-67922-y.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, 2023. https://doi.org/10.18653/v1/2023.emnlp-main.741.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. In *International Conference on Learning Representations*, 2024. https://proceedings.iclr.cc/paper_files/paper/2024/hash/b0d4b9d3dc41b36cad1d36ddd27858ab-Abstract-Conference.html.

Hamilton Se-Hwee Oh, Jarod Rutledge, Daniel Nachun, Robert Palovics, Olamide Abiose, Patricia Moran-Losada, Divya Channappa, Deniz Yagmur Urey, Kate Kim, et al. Organ aging signatures in the plasma proteome track health and disease. *Nature*, 624(7990):164–172, 2023. https://doi.org/10.1038/s41586-023-06802-1.

Eric Orwoll, Janet Babich Blank, Elizabeth Barrett-Connor, Jane Cauley, Steven Cummings, Kristine Ensrud, Cora Lewis, Peggy M. Cawthon, Robert Marcus, Lynn M. Marshall, Joan McGowan, Kathy Phipps, Sherry Sherman, Marcia L. Stefanick, and Katie Stone. Design and baseline characteristics of the osteoporotic fractures in men (MrOS) study: a large observational study of the determinants of fracture in older men. *Contemporary Clinical Trials*, 26(5):569–585, 2005. https://doi.org/10.1016/j.cct.2005.05.006.

Luis Paixao, Pooja Sikka, Haoqi Sun, Aayushee Jain, Jacob Hogan, Robert J. Thomas, and M. Brandon Westover. Excess brain age in the sleep electroencephalogram predicts reduced life expectancy. *Neurobiology of Aging*, 88:150–155, 2020. https://doi.org/10.1016/j.neurobiolaging.2019.12.015.

Shrey Pandit, Jiawei Xu, Junyuan Hong, Zhangyang Wang, Tianlong Chen, Kaidi Xu, and Ying Ding. MedHallu: A comprehensive benchmark for detecting medical hallucinations in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2858–2873, 2025. https://doi.org/10.18653/v1/2025.emnlp-main.143.

Stuart F. Quan, Barbara V. Howard, Conrad Iber, James P. Kiley, F. Javier Nieto, George T. O'Connor, David M. Rapoport, Susan Redline, John Robbins, Jonathan M. Samet, and Paul W. Wahl. The sleep heart health study: design, rationale, and methods. *Sleep*, 20(12):1077–1085, 1997. https://doi.org/10.1093/sleep/20.12.1077.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023. https://doi.org/10.1038/s41586-023-06291-2.

Stephen M. Smith, Diego Vidaurre, Fidel Alfaro-Almagro, Thomas E. Nichols, and Karla L. Miller. Estimation of brain age delta from brain imaging. *NeuroImage*, 200:528–539, 2019. https://doi.org/10.1016/j.neuroimage.2019.06.017.

Kaya Stechly, Karthik Valmeekam, and Subbarao Kambhampati. On the self-verification limitations of large language models on reasoning and planning tasks. In *International Conference on Learning Representations*, 2025. https://openreview.net/forum?id=4O0v4s3IzY.

Haoqi Sun, Luis Paixao, Jefferson T. Oliva, Balaji Goparaju, Diego Z. Carvalho, Kicky G. van Leeuwen, Oluwaseun Akeju, Robert J. Thomas, Sydney S. Cash, Matt T. Bianchi, and M. Brandon Westover. Brain age from the electroencephalogram of sleep. *Neurobiology of Aging*, 74:112–120, 2019. https://doi.org/10.1016/j.neurobiolaging.2018.10.016.

Haoqi Sun, Sasha Milton, Yi Fang, Hash Brown Taha, Shreya Shiju, Robert J. Thomas, Wolfgang Ganglberger, Matthew P. Pase, Timothy Hughes, Shaun Purcell, Susan Redline, Katie L. Stone, Kristine Yaffe, M. Brandon Westover, and Yue Leng. Machine learning–based sleep electroencephalographic brain age index and dementia risk: An individual participant data meta-analysis. *JAMA Network Open*, 9(3):e261521, 2026. https://doi.org/10.1001/jamanetworkopen.2026.1521.

Rahul Thapa, Magnus Ruud Kjaer, Bryan He, Ian Covert, Hyatt Moore, IV, Umaer Hanif, Gauri Ganjoo, M. Brandon Westover, Poul Jennum, Andreas Brink-Kjaer, Emmanuel Mignot, and James Zou. A multimodal sleep foundation model for disease prediction. *Nature Medicine*, 32(2): 752–762, 2026. https://doi.org/10.1038/s41591-025-04133-4.

Ye Ella Tian, Vanessa Cropley, Andrea B. Maier, Nicola T. Lautenschlager, Michael Breakspear, and Andrew Zalesky. Heterogeneous aging across multiple organ systems and prediction of chronic disease and mortality. *Nature Medicine*, 29(5):1221–1231, 2023. https://doi.org/10.1038/s41591-023-02296-6.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, 2023. https://doi.org/10.52202/075280-3275.

Xidong Wang, Shuqi Guo, Yue Shen, Junying Chen, Jian Wang, Jinjie Gu, Ping Zhang, Lei Liu, and Benyou Wang. LiveClin: A live clinical benchmark without leakage. In *International Conference on Learning Representations*, 2026a. https://proceedings.iclr.cc/paper_files/paper/2026/hash/9f45b2d4ace7ae425e40c7c1d0d37e64-Abstract-Conference.html.

Yunhe Wang, Sihao Xiao, Bowen Liu, Rongtao Jiang, Yuxi Liu, Yian Hang, Li Chen, Runsen Chen, Michael V. Vitiello, Derrick Bennett, et al. Organ-specific proteomic aging clocks predict disease and longevity across diverse populations. *Nature Aging*, 6(1):162–180, 2026b. https://doi.org/10.1038/s43587-025-01016-8.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. https://openreview.net/forum?id=WE_vluYUL-X.

Elissa Ye, Haoqi Sun, Michael J. Leone, Luis Paixao, Robert J. Thomas, Alice D. Lam, and M. Brandon Westover. Association of sleep electroencephalography-based brain age index with dementia. *JAMA Network Open*, 3(9):e2017357, 2020. https://doi.org/10.1001/jamanetworkopen.2020.17357.

A Comparative Landscape of Aging Representations

A.1 Dimensions of comparison

Aging representations differ in what they measure, how their age scale is constructed, whether they return one or several physiological ages, and how their biological relevance is evaluated. Table A.1 organizes representative molecular, sleep-derived, multi-organ, and report-derived approaches along these dimensions. This task-level comparison complements the conceptual synthesis in the main text and makes explicit the distinction between clocks trained to predict chronological age, profile decoding without outcome supervision, and direct outcome representations.

Table A.1: Representative biological-age, sleep-derived, and multi-organ aging representations. Studies are compared by input, supervision, output granularity, longitudinal evaluation, and the form in which their representation remains interpretable or traceable.

Study	Input	Representation and supervision	Longitudinal evaluation	Age output	Interpretability or traceability
Horvath (2013)	DNA methylation across tissues and cell types	Penalized regression fitted to chronological age	Cross-tissue and disease-state validation	One pan-tissue epigenetic age	CpG-weighted molecular clock
Sun et al. (2019)	Sleep EEG in MGH and SHHS	Handcrafted EEG features fitted to chronological age	Repeat-PSG aging and disease associations	One brain age	Feature-defined EEG age
Paixão et al. (2020)	SHHS sleep EEG	Sleep-EEG brain-age index fitted to chronological age	All-cause mortality	One brain age	Prospective validation of a scalar EEG gap
Brink-Kjaer et al. (2022)	Raw PSG from seven cohorts	Deep chronological-age regression	Mortality and life expectancy	One sleep age	Signal attribution for a raw-PSG model
Tian et al. (2023)	UK Biobank imaging, physiological, and blood phenotypes	Normative chronological-age models for three brain and seven body systems	Chronic disease and mortality	Ten system ages	Organ-wise phenotypic feature sets
Oh et al. (2023)	Organ-enriched plasma proteins in five cohorts	Bagged LASSO models fitted to chronological age	Disease, mortality, and dementia progression	Organismal plus 11 organ ages	Proteins mapped to tissue of origin
Li et al. (2025)	Health reports from six population cohorts	LLM prompting without outcome supervision	Mortality and broad incident disease	Overall plus organ-specific ages	Patient-level report reasoning
Sun et al. (2026)	Sleep EEG from five longitudinal cohorts	Sleep-EEG brain-age index and participant-level meta-analysis	Incident dementia	One brain age	Multi-cohort neurological validation
SleepFM (Thapa et al., 2026)	Multimodal raw PSG in Stanford cohorts and SHHS	Self-supervised PSG foundation model	Multiple future diseases	No biological-age vector	Learned multimodal PSG representation
Bilal et al. (2026)	Clinical PSG in STARLIT-10K and SHHS	Transformer representation learning	Mortality and cardiovascular and neurological outcomes	No biological-age vector	Embedding-derived risk strata
Wang et al. (2026b)	Plasma proteomics in UKB, CKB, and NHS	Organ-enriched models fitted to chronological age	Disease onset, progression, and mortality	Organismal plus ten organ ages	Cross-population organ-protein mapping
This work	Integrated sleep profiles in SHHS, MrOS, and XH	LLM decoding of expert-structured evidence without outcome supervision	Mortality, incident disease, and disease burden	One overall sleep age plus ten organ-resolved ages	Evidence IDs, trace-checked reports, and output consistency auditing

A.2 Construction and interpretation of age gaps

For participant i and age component k, the LLM-derived gap is $g_{ik} = \widehat{a}_{ik} - a_i$, where $\widehat{a}_{ik}$ is the decoded age and a_i is chronological age. This is the direct-difference construction used for LLM-derived overall and organ-specific ages by Li et al. (2025). It is appropriate here because the estimator is not trained within SHHS, MrOS, or XH to minimize chronological-age prediction error. The complete age vector is therefore retained as decoded, and chronological age enters the downstream association and incremental-discrimination models explicitly.

This adjustment also clarifies the relation to a linearly residualized gap. If $\widetilde{g}_{ik} = g_{ik} - (\alpha + \beta a_i)$, replacing g_{ik} with $\widetilde{g}_{ik}$ in a model that already contains a_i changes the chronological-age coefficient but not the coefficient of the gap. The primary age-adjusted analyses consequently test the same linear gap contribution while preserving the prespecified LLM representation. Conventional sleep-feature

regressors are different because they are fitted directly to chronological age and can exhibit bias due to regression to the mean. Their gaps are corrected according to Smith et al. (2019) before external-cohort evaluation. LLM and supervised-model gaps are then compared on the same endpoint-specific participants and cross-validation folds, following the outcome-comparison strategy of Li et al. (2025).

A.3 Complementary validation axes

The experiments interrogate three distinct properties of an aging representation. Healthy-subset chronological-age regression asks whether a shared feature table can reconstruct an age scale. Association with independently constructed deficit-accumulation indices asks whether the decoded overall sleep gap follows an established measure of accumulated health burden. Prospective mortality, incident-disease, and disease-burden analyses ask whether the frozen representation is informative for subsequent health trajectories after chronological-age adjustment. Finally, the endpoint-matched comparison tests organ resolution directly by evaluating the prespecified organ-linked gap against every alternative gap for the same endpoint and participants. These axes distinguish age reconstruction, general aging content, longitudinal relevance, and organ correspondence rather than treating any one metric as a complete validation of biological age.

B EXPERIMENTAL SETUP

B.1 COHORTS AND ANALYSIS POPULATIONS

SHHS. The Sleep Heart Health Study is a community-based prospective study of sleep-disordered breathing and cardiovascular outcomes in adults aged at least 40 years (Quan et al., 1997). We processed 5,793 participants with an index sleep study and registry-resolvable baseline evidence. Available evidence spans demographics, lifestyle, anthropometry, medical history, questionnaires, and polysomnographic summaries. Mortality follow-up was available for 5,791 participants. The number included in each endpoint analysis varied because incident-outcome models excluded participants with that condition at baseline, and adjusted models additionally required complete follow-up and covariate data.

MrOS. The Osteoporotic Fractures in Men Study is a prospective cohort of men aged at least 65 years (Orwoll et al., 2005). Its sleep ancillary data are distributed through the National Sleep Research Resource (Dean et al., 2016). We processed 2,904 participants with evidence reports. The cohort provides mortality, cardiovascular, neurological, and fracture outcomes. A total of 2,839 predictions linked exactly to the survival/outcome table.

XH. XH is a clinical sleep cohort with baseline sleep reports, examination and medical-history evidence, and longitudinal diagnosis records. The analysis included 426 participants. The model received only de-identified baseline evidence reports, without longitudinal outcome records. Across cohorts, sleep measurements and their accompanying baseline context form the integrated sleep profile. All three final analysis tables were produced by the defined framework and contain the same jointly returned overall sleep age and ten organ-resolved ages. The three cohorts total 9,123 participants.

B.2 EVALUATION QUESTIONS

We separate five questions that are often conflated in generative clinical systems. *Input fidelity* measures completion, structured-output validity, Evidence-ID coverage, value/unit fidelity, final traceability status, and privacy checks. *Chronological-age anchoring and calibration* measure agreement between the overall sleep age and chronological age, fitted slope and intercept, mean gap, and residual age dependence. *Outcome association and discrimination* test frozen gaps against longitudinal outcomes using HRs, incidence-rate ratios, and Harrell C-indices. *Incremental discrimination* asks whether endpoint-matched gaps improve ordering beyond clinical baselines. *Physiological-consistency auditing* asks whether the Consistency Audit Agent concentrates prespecified gross inconsistencies into a small review set. Outcome analyses evaluate the complete integrated profile without component-wise attribution.

The primary claim is traceable representation construction rather than supervised outcome prediction or agent-count superiority. We nevertheless compare the frozen representation with conventional chronological-age regressors and baseline-covariate survival models. The unified SHHS, MrOS, and XH eleven-age tables were frozen without outcome access and analyzed as supplied.

B.3 CONSISTENCY AUDIT AGENT AND DEVELOPMENTAL EVALUATION

The frozen high-threshold gate received each complete eleven-target estimator output and returned either PASS or CLINICIAN_REVIEW. It inspected only two prespecified conditions: an impossible order defined by within-night sleep-measurement timing, and an explicit contradiction in which the same named variable is simultaneously unknown and assigned a state or is assigned opposing states. Evidence support, omitted alternatives, general clinical completeness, and numerical-age magnitude were outside its scope. A review decision did not replace, resample, or modify the candidate age vector.

We characterized the review rate in all 426 XH candidates. An independent model-based developmental reference applied the same two-category rubric to all 426 outputs without access to the audit decisions. We report error prevalence within PASS and REVIEW groups with two-sided Wilson 95% confidence intervals and compare them using a two-sided Fisher exact test. Sensitivity, specificity,

and positive predictive value are retained in Source Data. Because the reference was model based rather than produced by sleep clinicians, this analysis characterizes development-stage triage behavior. Independent annotation by clinical sleep experts remains a separate validation step.

B.4 Calibration and cohort-characteristic analyses

For each cohort, we plotted the final overall sleep age against chronological age with an identity line and ordinary least-squares fit. We reported Pearson r, fitted slope and intercept, mean gap, and gap standard deviation. A complementary plot of gap against chronological age tested residual chronological-age dependence. These analyses were designed to detect systematic bias, not to confirm its absence. Cohort age and BMI distributions were summarized with violins and mean, SD, median, and interquartile range. Values outside 10–80 kg m^{-2} were treated as invalid BMI entries for descriptive reporting only.

B.5 Healthy-reference traditional age models

We reconstructed the conventional SHHS chronological-age regression benchmark from the supplied `ML_age` notebooks. The primary healthy reference excludes baseline or prior cardiovascular disease, diabetes, hypertension, available major disease, sleep treatment, apnea–hypopnea index ≥ 15 events h^{-1}, or at least 5% of sleep time below 90% oxygen saturation. Chronological age and both sleep-breathing measures must be observed. This definition identifies 1,488 SHHS participants, of whom 1,486 are included in the primary framework cohort. No follow-up variable informed this definition.

Elastic Net, radial-basis support-vector regression (SVR-RBF), random forest, extra-trees, histogram gradient-boosting, and a multilayer perceptron (denoted DNN) predict chronological age from the same 18 variables available in all cohorts: sex, BMI, systolic and diastolic blood pressure, total sleep time, sleep efficiency, sleep-onset latency, wake after sleep onset, N1/N2/N3/REM percentages, REM latency, arousal index, obstructive- and central-apnea indices, and mean and nadir SpO_2. The multilayer perceptron is denoted DNN in figures and tables. It follows the supplied notebook specification with 32- and 16-unit tanh hidden layers, Adam optimization, ℓ_2 penalty 10^{-2}, learning rate 5×10^{-4}, batch size 64, and early stopping. Within each SHHS fold, medians, missingness indicators, scaling, and parameters are learned from healthy training participants only. SHHS estimates are five-fold out of fold. MrOS and XH are strict external transfers from the full linked healthy SHHS set. We report MAE, RMSE, Pearson r, R^2, and mean error. Because chronological age is explicitly available to the agents, their chronological-age MAE is an anchoring diagnostic rather than a fair supervised leaderboard against models that must infer chronological age from features alone. For outcome comparison, we first corrected the conventional gaps, defined as predicted age minus chronological age, within each cohort using a second-stage regression on centered chronological age and an orthogonalized quadratic chronological-age term, following the bias-correction principle of Smith et al. (2019). Correction parameters were fitted without any follow-up outcome and before outcome linkage. We then fitted gap-only penalized Cox models for all 48 eligible endpoints (11 SHHS, 17 MrOS, and 20 XH). Within each endpoint, every method used the same participants, events, and follow-up. Following Li et al. (2025), the LLM gaps retained their direct definition as estimated age minus chronological age, whereas the supervised-model gaps used their prespecified age-bias correction. We evaluated two frozen LLM representations: the overall sleep gap for every endpoint, and an endpoint-matched representation using the overall sleep gap for all-cause death and the prespecified organ-resolved gap for each organ-linked endpoint. All methods were compared by absolute gap-only Harrell C-index on the same participants, events, and follow-up for each endpoint. This benchmark tests outcome ordering under the gap construction native to each estimator. The separate age-adjusted analyses and analyses adding gaps to baseline models test information beyond chronological age.

B.6 Validation against aging-related deficit measures

SHHS uses the existing 36-deficit frailty index and requires at least 30 observed deficits. We mapped the same accumulation principle to 33 available MrOS clinical, disease, medication, functional, and self-reported health deficits, again requiring at least 30 observations. XH does not contain 30 uniformly observed non-sleep deficits. We therefore report a 20-item baseline clinical-deficit burden when at least 16 items are observed and do not label it a frailty index. For each cohort's final output,

we estimate Spearman correlation and a robust linear association between standardized overall sleep gap and standardized deficit measure, adjusting for chronological age and sex where sex varies.

B.7 Survival and discrimination analyses

For SHHS/MrOS, we used the prespecified mapping from 28 outcomes to one prespecified gap per outcome. All-cause mortality uses the overall sleep gap, and each organ-linked endpoint uses its prespecified corresponding organ-resolved gap. Examples include cardiovascular gaps for cardiovascular outcomes, neurological gaps for cerebrovascular outcomes, and skeletal gaps for fractures. The XH mapping extends this rule across all prespecified diagnosis categories. Gaps were winsorized at cohort-specific first and 99th percentiles. SHHS and MrOS used four nested Cox specifications: gap alone, gap plus chronological age, additional adjustment for race, BMI, smoking, alcohol, and sex where variable, and additional adjustment for baseline cardiovascular disease, with fracture history also included in MrOS. Covariates were handled by five multiple imputations. Every available positive follow-up time was retained. Participants without a recorded event were right-censored at their last recorded follow-up. No common administrative horizon was imposed. Main figures report HRs per one-year larger gap with 95% confidence intervals. Per-SD estimates are retained in source tables. Benjamini–Hochberg correction (Benjamini & Hochberg, 1995) was applied within cohort and adjustment level across endpoint-matched outcomes.

For all-cause mortality, SHHS participants were ranked by overall sleep gap and partitioned into three consecutive groups of nearly equal size. The low, intermediate, and high groups contained 1,926, 1,925, and 1,925 participants, respectively. Any remainder was assigned to lower-ranked groups. Within each cohort, equal gaps were ordered by ascending participant identifier, so tied values could span a group boundary. Extreme groups contained the first and last 10% of ranked participants, rounding each tail's size to the nearest integer (578 each in SHHS and 247 each in MrOS). Kaplan–Meier curves use all available observed follow-up, and the two-sided log-rank test compares the high and low strata. Covariate-adjusted associations were estimated separately using continuous-gap Cox models. We also tabulated one-year gap-bin counts. Because the 0–1 and 1–2 year bins contained only 5 and 10 participants, respectively, the survival sensitivity pools final SHHS gaps into prespecified wider intervals (< 2, 2–< 4, 4–< 6, 6–< 8, 8–< 10, and ≥ 10 years). Fixed five-year and extreme-decile Kaplan–Meier displays are retained as descriptive threshold sensitivities.

The eligible XH outcome dictionary contained 244 clinically filtered, organ-mapped diagnosis categories with at least one post-index event. The first-new-diagnosis composite was the first eligible category after the index study that was absent at baseline. A sensitivity definition that allowed post-index recurrence of a category already present at baseline produced the same 256 events among 426 participants. Models were unadjusted, adjusted for chronological age, and adjusted for chronological age plus baseline diagnosis burden. Twenty clinically representative organ-specific endpoints with at least ten events after excluding participants with the corresponding diagnosis at baseline used prevalent-disease exclusions, the corresponding organ-age gap, and a Cox penalty of 0.01. Benign prostatic hyperplasia was evaluated only in men. FDR was controlled across the 20 final tests. Absolute C-indices used 2,000 participant-level bootstrap resamples with Cox refitting.

For cumulative XH burden, we counted the number of distinct categories from the same 244-category eligible outcome dictionary that first appeared after baseline. Negative-binomial models use log follow-up as an offset and report incidence-rate ratios per one-SD overall sleep gap. Models are unadjusted, chronological-age adjusted, and adjusted for chronological age plus log-transformed baseline diagnosis burden. Patient-specific annualized observed rates are displayed descriptively. As a model-based sensitivity, we overlay adjusted predicted rates over -2 to $+2$ SD of gap with chronological age and baseline burden held at their cohort means.

To test organ–outcome correspondence, we evaluate all eleven candidate gaps for every eligible non-mortality endpoint using the same participants. Risk-score direction is selected in each training fold and evaluated out of fold over ten repeated five-fold splits. Each candidate augments a Cox model using only chronological age. The competition metric is the out-of-fold C-index increment ΔC-index over that same chronological-age baseline. The primary descriptive contrast is the within-endpoint matched-gap advantage: the prespecified matched gap's ΔC-index minus the mean ΔC-index of the ten non-matched gaps. Candidate ranks are retained as secondary diagnostics. This prevents residual chronological-age structure from being mistaken for organ specificity.

XH LLM-backbone sensitivity. We focused on open-weight models because they support deployment in institution-controlled computing environments, an important option for protecting clinical data. We compared the current Llama 3.3-70B representation with six alternative LLM backbones that returned the same eleven-age schema for the 426 XH reports: DeepSeek V4 Pro, DeepSeek V4 Flash, Qwen3 235B-A22B, Qwen 3.5 27B, GPT-OSS 120B, and Gemma 4 31B. For each participant and alternative backbone, one complete generation was drawn uniformly with a fixed seed (20260825) before outcomes were analyzed. Llama 3.3 used the primary framework output. Every backbone contributed all eleven ages for the same 426 participants. Analysis gaps were calculated as estimated age minus chronological age.

For the first diagnosis from the 244-category eligible outcome set that was absent at baseline, Cox models reported HRs and robust 95% confidence intervals per one-year overall sleep gap. The three prespecified specifications were unadjusted, adjusted for chronological age, and adjusted for chronological age plus log-transformed baseline diagnosis burden. Each of the 20 clinically representative incident outcomes had at least ten events. Absolute Harrell C-index was estimated from a gap-only penalized Cox model using the corresponding organ-age gap. Every backbone used the same participants within each endpoint. Cross-endpoint means summarize performance across the 20 outcomes.

XH matched-gap incremental discrimination. We separately compared seven penalized Cox specifications within each outcome using the same participants and ten-repeated five-fold splits: chronological age, with and without the endpoint-matched gap, chronological age plus sex with and without the matched gap, the clinical base with and without the matched gap, and the matched gap alone. The clinical base comprised chronological age, sex, and log-transformed baseline diagnosis burden. Baseline burden was the number of distinct eligible diagnosis categories recorded on or before the sleep study. The overall sleep gap was matched to the strict first-new-diagnosis composite, whereas each organ outcome used its prespecified organ gap. Sex was omitted for the male-only benign prostatic hyperplasia analysis because it was invariant. Harrell C-indices were computed from log-risk scores averaged across repeats. The primary nested contrast was clinical base plus matched gap minus clinical base. Its 95% confidence interval and two-sided tail probability used 2,000 paired participant-bootstrap resamples with add-one correction. The two simpler paired contrasts contextualized the same gap increment. Benjamini–Hochberg correction covered the 20 organ outcomes. The composite was reported separately.

B.8 Implementation and audit artifacts

Each cohort analysis table contains chronological age, eleven estimated ages, their signed gaps, and participant identifiers verified against the original cohort order. Quality checks confirmed row counts, chronological ages, recomputed gaps, endpoint linkage, and endpoint-specific analysis populations. Outcomes and follow-up were unavailable during age estimation.

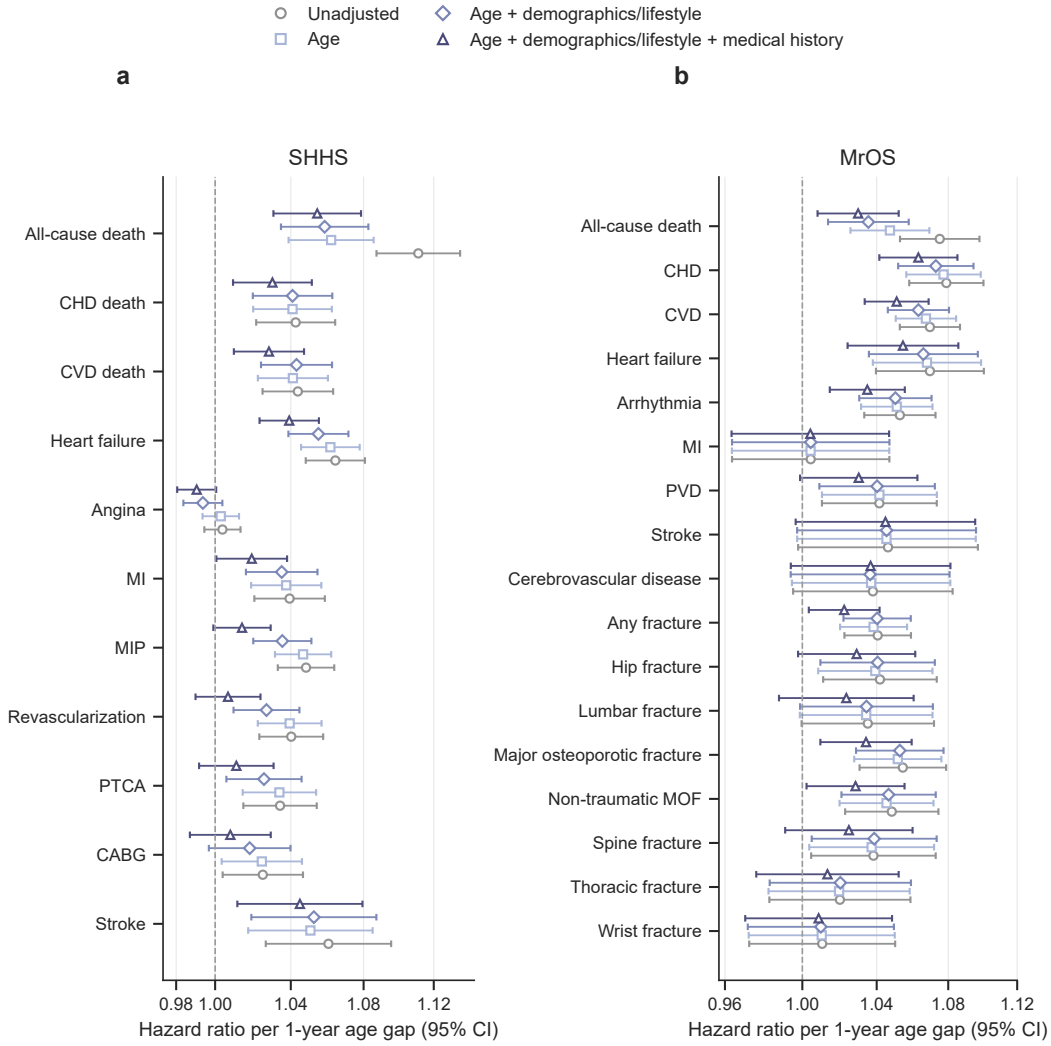

Figure C.1: Endpoint-specific HRs per one-year larger prespecified age gap under four nested adjustment sets. The demographic/lifestyle model includes chronological age, sex where variable, race, BMI, smoking, and alcohol use. The full model additionally includes baseline cardiovascular disease and, in MrOS, fracture history. MrOS is male-only, so sex is not a model covariate. Horizontal bars are 95% confidence intervals. The vertical line denotes HR=1.

C Extended Results

C.1 Mortality associations and survival stratification

Figure C.1 reports HRs per one-year larger prespecified gap under four nested specifications. In SHHS, 10, 10, 10, and 5 of 11 associations passed within-specification FDR correction in the gap-only, chronological-age-adjusted, demographic/lifestyle-adjusted, and fully adjusted models. Corresponding MrOS counts were 11, 11, 11, and 7 of 17. All-cause mortality remained associated with the overall sleep gap after full adjustment in SHHS (1,302 deaths among $n = 5{,}776$, HR 1.054, 95% CI 1.031–1.079, $P = 4.92 \times 10^{-6}$) and MrOS (2,039 deaths among $n = 2{,}466$, HR 1.030, 1.008–1.052, $P = 7.03 \times 10^{-3}$). Across the four specifications, the SHHS HR decreased from 1.111 to 1.062, 1.058, and 1.054. The corresponding MrOS estimates were 1.075, 1.047, 1.035, and 1.030.

Additional survival analyses are reported in Supplementary Figures J.1 and K.1. The former documents the sparse SHHS counts in literal one-year gap intervals and the resulting pooled two-year

display. The latter reports MrOS five-year and extreme-decile sensitivities. The SHHS survival analysis included 5,776 of 5,791 records linked to the survival source after requiring complete follow-up, event, gap, and chronological-age fields. Participants were followed to their recorded event or last observation without a common administrative horizon.

C.2 Multi-organ outcome associations

Endpoint-matched gaps were associated with outcomes across cardiovascular, neurological, and skeletal systems after full adjustment. In SHHS, the cardiovascular gap was associated with heart failure (HR 1.039, 95% CI 1.023–1.055), cardiovascular death (1.028, 1.010–1.047), and coronary heart disease death (1.030, 1.009–1.051), while the neurological gap was associated with stroke (1.045, 1.012–1.079). In MrOS, the cardiovascular gap was associated with coronary heart disease (1.063, 1.041–1.085), cardiovascular disease (1.051, 1.033–1.069), heart failure (1.055, 1.024–1.086), and arrhythmia (1.035, 1.015–1.056). The skeletal gap was associated with any fracture (1.022, 1.003–1.042) and major osteoporotic fracture (1.034, 1.010–1.059). All associations listed above passed within-specification FDR correction. SHHS angina was not positively associated, showing that a prespecified organ match did not imply a uniformly positive association.

Among 426 XH participants, 256 developed a first post-index diagnosis from the 244-category eligible outcome set that was absent at baseline. The overall sleep gap had an SD of 3.198 years. After adjustment for chronological age and baseline diagnosis burden, the HR was 1.202 per one-SD larger gap (95% CI 1.064–1.357, $P = 3.11 \times 10^{-3}$). Expressing the same coefficient per year gave an HR of 1.059 (1.020–1.100). The chronological-age-adjusted model without baseline burden gave a similar per-year HR of 1.063 (1.024–1.104). XH also contributed a cumulative diagnosis-burden analysis over the same 244-category set and 20 clinically representative organ-linked endpoint analyses with at least ten events each. For cumulative burden, the fully adjusted negative-binomial model gave an incidence-rate ratio of 1.330 per SD (1.141–1.549, $P = 2.61 \times 10^{-4}$), corresponding to 1.093 per year (1.042–1.147). Complete effect estimates appear in Figure C.4b,c, and the participant-level rate distribution with adjusted fit appears in Supplementary Figure F.2.

Higher endpoint-matched C-indices among the 20 XH organ-linked outcomes included 0.861 for hepatic insufficiency, 0.790 for myocardial ischemia, 0.741 for diabetes mellitus, and 0.706 for non-alcoholic fatty liver disease, among others (Figure C.2). The two highest estimates were based on relatively few events and consequently had wide bootstrap intervals. Figure C.2 reports all 20 outcomes.

C.3 Sensitivity across LLM backbones

The XH backbone sensitivity analysis used one complete eleven-age output per participant from each model, with all outputs fixed before outcomes were analyzed. All 426 participants were included (Supplementary Figure F.1). For the first-new-diagnosis composite, the Llama 3.3-70B HR per one-year overall sleep gap was 1.055 unadjusted (95% CI 1.018–1.094), 1.063 after chronological-age adjustment (1.024–1.104), and 1.059 after chronological-age and baseline-burden adjustment (1.020–1.100, $P = 3.11 \times 10^{-3}$). Fully adjusted HRs for the six alternative backbones ranged from 1.002 to 1.042. Across the 20 endpoint-matched organ outcomes, the macro-average C-index was 0.668 for Llama 3.3-70B and 0.648 for the nearest alternative, Gemma 4 31B. The endpoint-level distributions show that backbone differences were not uniform across outcomes.

C.4 Organ resolution and conventional age-regression baselines

Organ correspondence varied in magnitude but was consistent across cohorts. In SHHS, it was expressed primarily as rank consistency rather than a large absolute gain: the endpoint-matched gap improved C-index beyond chronological age by 0.014 on average, compared with 0.006 for the ten alternative gaps. This modest 0.008 mean advantage was positive for all ten endpoints. The matched gap ranked among the top three in nine and exceeded 92% of alternatives on average. Mean advantages were 0.018 in MrOS and 0.057 in XH, with corresponding concordance values of 0.88 and 0.84 (Figure 5). Endpoint-level comparisons of all eleven gaps are provided in Supplementary Figure I.1.

We next compared the two LLM representations with six chronological-age regressors trained on healthy SHHS participants. The 18-variable common input set and its cross-cohort availability are documented in Supplementary Figure E.1. All methods used identical participants and follow-up within each of 48 endpoint-specific evaluations. Mean gap-only C-index was 0.569 for the overall sleep gap and 0.615 for the endpoint-matched representation, which used the overall sleep gap for all-cause death and the prespecified organ-resolved gap otherwise. Among traditional gaps, extra-trees had the highest mean C-index (0.551, Figure C.3 and Table C.1). All methods were compared using the same participants, events, and observed follow-up for each endpoint. The LLM rows used gaps calculated directly as estimated age minus chronological age, while the supervised regressors used gaps corrected for age bias using parameters estimated before outcome linkage.

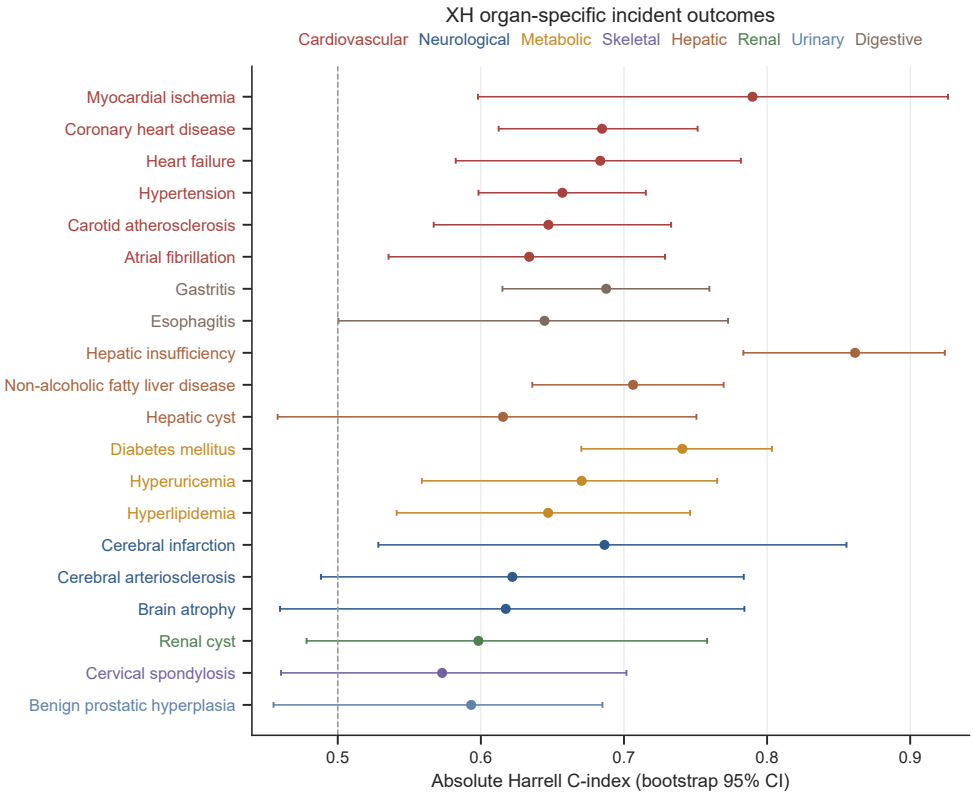

Figure C.2: XH Harrell C-indices for 20 incident outcomes using endpoint-matched organ-resolved gaps. Bars are bootstrap 95% confidence intervals from 2,000 participant-level resamples. Point and label colors denote organ systems. The dashed line marks chance-level discrimination.

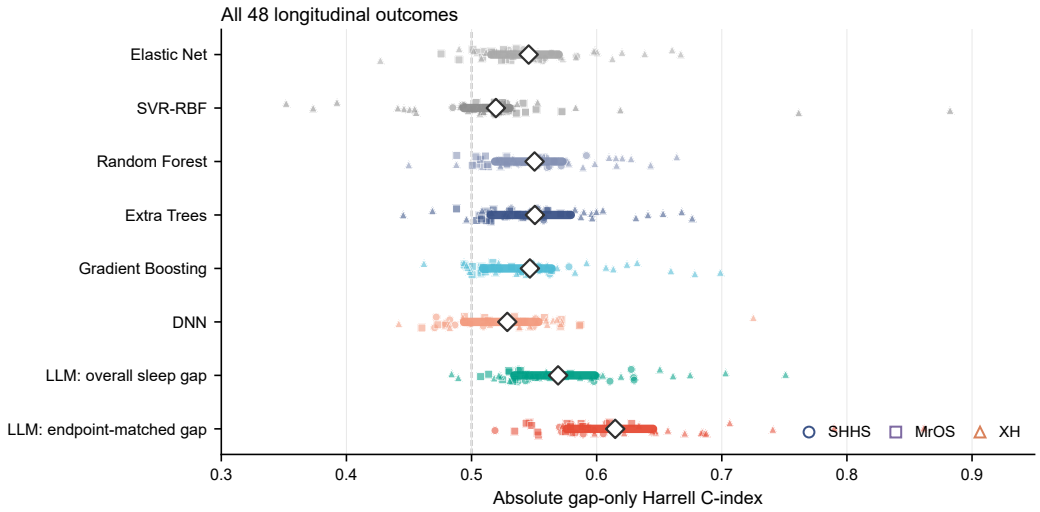

Figure C.3: Absolute gap-only Harrell C-index across 48 eligible outcomes (11 SHHS, 17 MrOS, and 20 XH). The LLM overall row uses the overall sleep gap for every endpoint. The endpoint-matched row uses the overall sleep gap for all-cause death and the prespecified organ-resolved gap for each organ-linked endpoint. Each dot is one endpoint. Horizontal lines show the interquartile range and diamonds show arithmetic means. Markers identify cohorts, and the dashed line denotes chance discrimination. All methods use the same participants and follow-up within each endpoint. LLM gaps equal estimated age minus chronological age. Traditional gaps use their prespecified age-bias correction fitted before outcome linkage.

Table C.1: All-outcome comparison of two LLM gap representations with six traditional chronological-age gaps. Values summarize absolute gap-only Harrell C-index across 48 outcomes (11 SHHS, 17 MrOS, and 20 XH). Every method used the same participants and follow-up within each endpoint. Columns report the mean, minimum, and maximum across outcomes. The endpoint-matched row uses the overall sleep gap for all-cause death and the prespecified organ-resolved gap otherwise. LLM gaps equal estimated age minus chronological age. Traditional gaps use their prespecified age-bias correction fitted before outcome linkage.

Representation	Mean C-index	Minimum	Maximum
LLM: endpoint-matched gap	0.615	0.519	0.861
LLM: overall sleep gap	0.569	0.484	0.751
Random forest	0.550	0.450	0.664
Extra-trees	0.551	0.445	0.676
Gradient boosting	0.547	0.462	0.699
Elastic Net	0.546	0.427	0.667
DNN	0.529	0.442	0.725
SVR-RBF	0.519	0.352	0.882

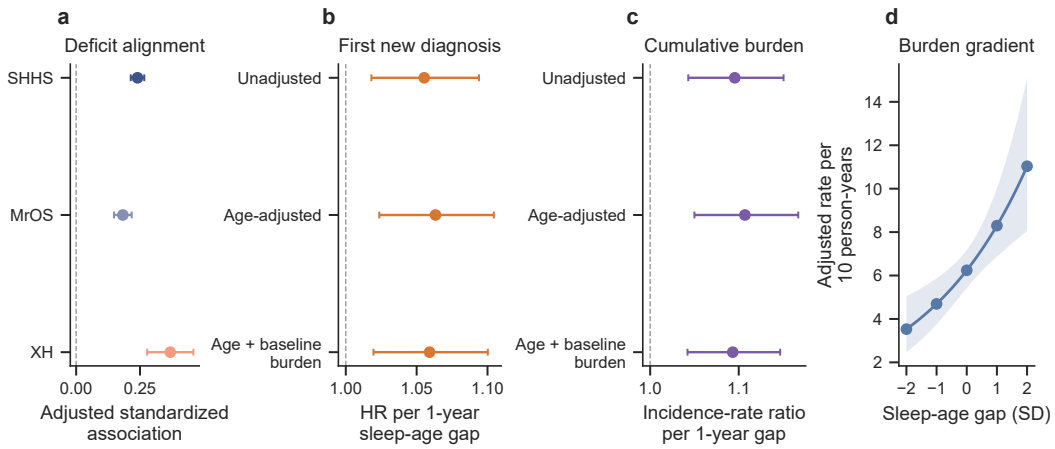

Figure C.4: Deficit alignment and XH longitudinal-burden analyses. Panel a shows adjusted standardized associations between the overall sleep gap and cohort-specific deficit measures. Panels b and c report per-year HRs for first new diagnosis and incidence-rate ratios for cumulative burden. Panel d gives fully adjusted model-based rates at -2, -1, 0, $+1$, and $+2$ SD of overall sleep gap. Bars and bands are 95% confidence intervals.

C.5 Incremental discrimination and deficit accumulation

Within each XH endpoint, paired models used the same participants and cross-validation folds. Adding the endpoint-matched gap to chronological age, sex, and baseline diagnosis burden yielded a mean ΔC-index of 0.070 across 20 organ outcomes. Eighteen of 20 point estimates were above zero (Table H.1). The two negative point estimates were brain atrophy (ΔC-index $= -0.006$) and renal cyst (-0.021), with both intervals crossing zero. For the strict first-new-diagnosis composite, adding the overall sleep gap increased C-index from 0.620 to 0.627 (ΔC-index $= 0.007$, 95% CI -0.010 to 0.023, $P = 0.399$). Across this composite and all 20 organ outcomes, the macro-average gains were 0.082 over chronological age, 0.088 over chronological age plus sex, and 0.067 over the clinical model.

The overall sleep gap was positively related to each cohort's independently constructed baseline deficit measure (Figure C.4a). After chronological age and, where applicable, sex adjustment, standardized coefficients were 0.240 in SHHS (95% CI 0.214–0.267, $P = 8.24 \times 10^{-71}$), 0.183 in MrOS (0.148–0.218, $P = 8.04 \times 10^{-25}$), and 0.368 for XH clinical-deficit burden (0.278–0.459, $P = 1.60 \times 10^{-15}$). Spearman correlations were 0.278, 0.199, and 0.352. The XH measure is termed clinical-deficit burden rather than a frailty index because only 20 defensible baseline deficits were available. Exact definitions and coefficients are listed in Table E.1. This cross-cohort alignment provides complementary evidence that the overall sleep gap captures aging-related health variation beyond chronological age.

C.6 Traceability and Consistency Audit Agent performance

The unified framework tables contained a complete vector for each of 5,793 SHHS, 2,904 MrOS, and 426 XH participants. Every vector jointly included one overall sleep age and ten organ-resolved ages. All age and gap cells were finite, and every supplied gap agreed with estimated age minus chronological age. Registry and trace checks verified planned and rendered Evidence IDs, canonical labels, values, units, duplicates, and section placement. These checks establish input fidelity. The full age distributions are shown in Supplementary Figure G.1.

Overall sleep age tracked chronological age in SHHS, MrOS, and XH, with Pearson correlations of 0.978, 0.934, and 0.971 and fitted slopes of 1.034, 1.045, and 0.989. Mean overall sleep gaps were 5.61, 5.82, and 8.49 years. Correlations between gap and chronological age were weak in magnitude in all three cohorts ($r = 0.150$, 0.112, and -0.045, respectively, with $|r| \leq 0.150$). Supplementary Figure D.2 shows the complete diagnostics.

The Consistency Audit Agent returned PASS for 391 of 426 XH outputs (91.8%) and REVIEW for 35 (8.2%, Supplementary Figure L.1a). In independent developmental annotations generated without access to the audit decisions, 24/35 REVIEW candidates contained a prespecified hard- consistency error (68.6%, Wilson 95% CI 52.0–81.4%), compared with 13/391 PASS candidates (3.3%, 2.0–5.6%). The two-sided Fisher exact test gave $P = 9.84 \times 10^{-22}$ (Supplementary Figure L.1b). At the observed 8.2% review rate, sensitivity was 64.9%, specificity 97.2%, positive predictive value 68.6%, and negative predictive value 96.7% (Supplementary Figure L.1c). The agent audited the physiological consistency of the complete estimator output, while the frozen eleven-age vector remained unchanged in the longitudinal analyses.

D COHORT CONTEXT AND CALIBRATION DIAGNOSTICS

The cohort differences are part of the transfer setting: MrOS is older and male-only, whereas XH is younger and clinically ascertained.

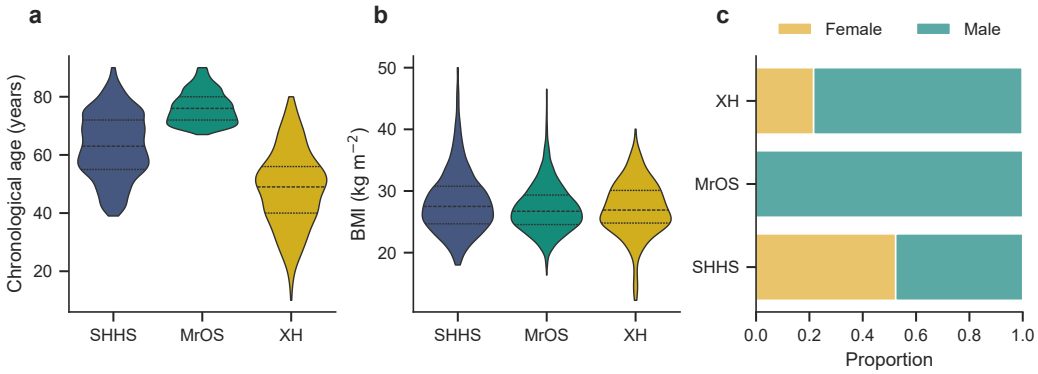

Figure D.1: Cohort characteristics. Violin plots show chronological age (a) and valid BMI (b), with quartiles inside each distribution. Panel c shows recorded sex proportions. Invalid or unavailable BMI values are excluded only from the descriptive BMI panel.

Table D.1: Cohort characteristics. Continuous variables are mean ± SD and median [interquartile range].

Cohort	n	Chronological age, years	BMI, kg m^{-2}	BMI available	Male, n (%)
SHHS	5,793	63.1 ± 11.2 (63 [55–72])	28.2 ± 5.1 (27.5 [24.7–30.8])	5,750	2,760 (47.6)
MrOS	2,904	76.4 ± 5.5 (76 [72–80])	27.2 ± 3.8 (26.7 [24.6–29.3])	2,902	2,904 (100.0)
XH	426	48.3 ± 13.1 (49 [40–56])	27.1 ± 4.4 (26.9 [24.8–30.1])	354	333 (78.2)

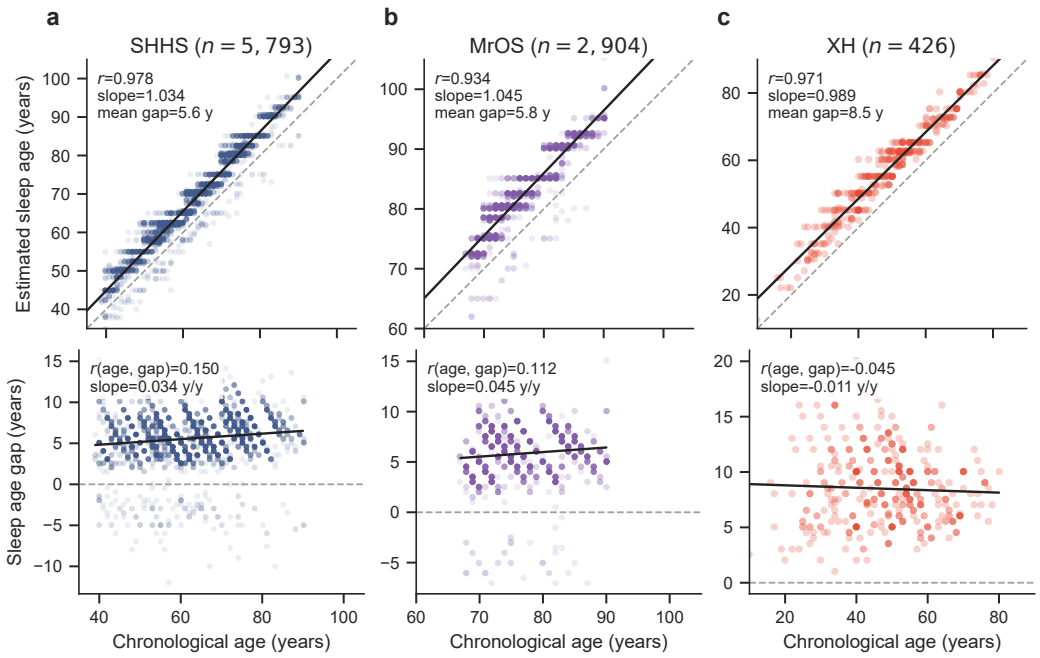

Figure D.2: Calibration and residual age dependence of the overall sleep age. Panels a–c correspond to SHHS, MrOS, and XH. Upper axes show participant-level overall sleep age against chronological age. Dashed lines denote identity and solid lines are cohort-specific linear fits. Lower axes show overall sleep gap against chronological age. Dashed lines denote zero gap and solid lines are linear fits. Each translucent circle is one participant. Insets report Pearson correlation, fitted slope, and mean gap.

E CONVENTIONAL-MODEL CONTEXT

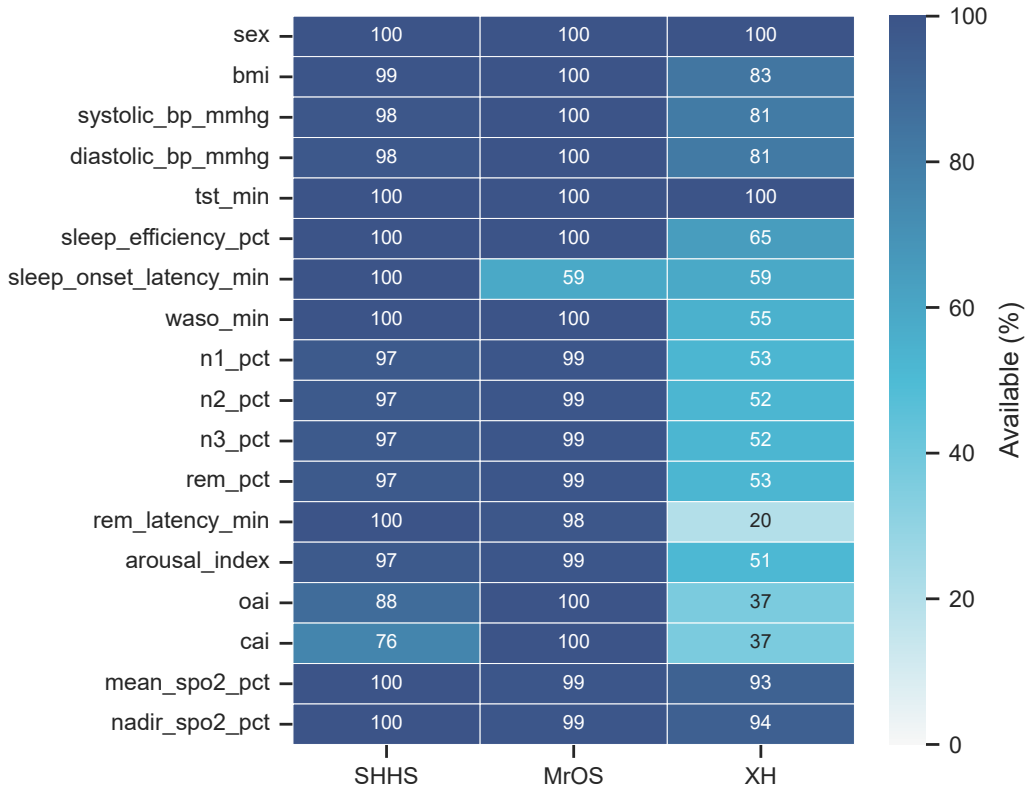

Figure E.1: Availability of the 18 common variables used by six traditional age regressors. Models were trained on healthy SHHS participants. MrOS and XH are external transfers. Missingness handling was fitted within the SHHS training data.

Table E.1: Association of the overall sleep gap with independent baseline health-deficit measures. Coefficients are standardized and adjusted for chronological age and sex where sex varies.

Cohort	Measure	n	Spearman ρ	Adjusted β (95% CI)
SHHS	36-deficit frailty index	5,300	0.278	0.240 (0.214–0.267)
MrOS	33-deficit frailty index	2,903	0.199	0.183 (0.148–0.218)
XH	20-item clinical-deficit burden	426	0.352	0.368 (0.278–0.459)

F XH BACKBONE AND LONGITUDINAL-BURDEN SENSITIVITIES

For each alternative backbone, one complete eleven-age output per participant was fixed before outcomes were analyzed.

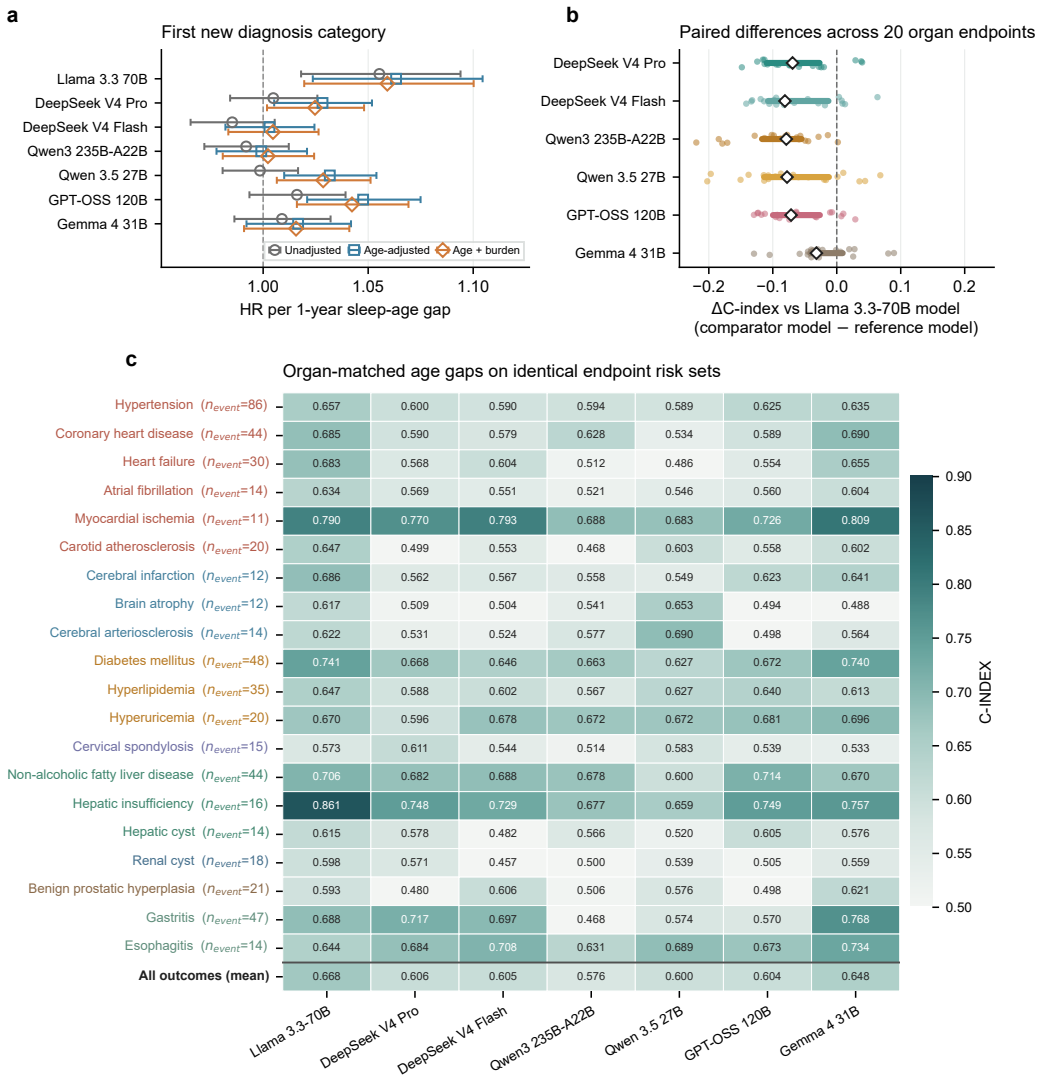

Figure F.1: XH LLM-backbone sensitivity using one complete eleven-age output per participant from each model, fixed before outcomes were analyzed ($n = 426$). Panel a reports HRs per one-year overall sleep gap for the first new diagnosis from the 244-category eligible XH outcome set (256 events). Panel b reports each comparator's endpoint-level C-index minus the Llama 3.3-70B C-index using the same participants and follow-up. Values below zero favor Llama 3.3-70B. Panel c gives absolute organ-matched C-indices across 20 incident outcomes. Its final row reports the unweighted mean across outcomes.

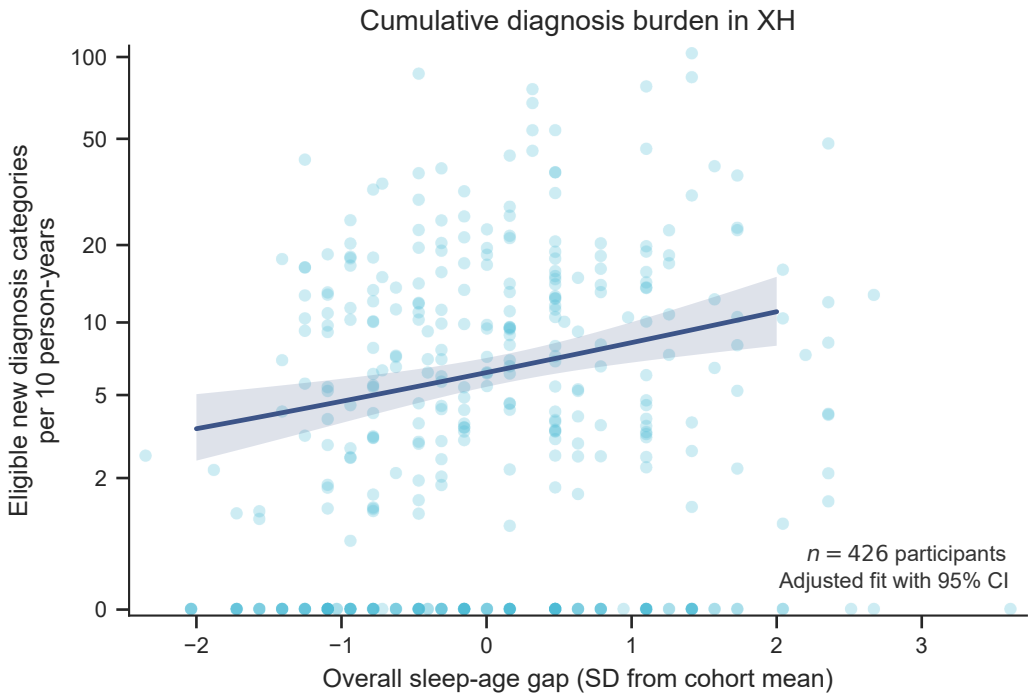

Figure F.2: Continuous XH cumulative diagnosis-burden sensitivity. For each participant ($n = 426$), the numerator is the number of distinct diagnoses from the 244-category eligible XH outcome set that first appeared after baseline. Each translucent circle shows this count per 10 person-years against the standardized overall sleep gap. The solid curve and shaded 95% confidence interval are from the negative-binomial model adjusted for chronological age and baseline diagnosis burden, with those covariates held at their cohort means. The fit is displayed over -2 to $+2$ SD. All participants remain visible. Vertical spacing uses a $\log(1 + y)$ display transform while tick labels retain the original rate scale.

G MULTICOHORT AGE-REPRESENTATION DISTRIBUTIONS

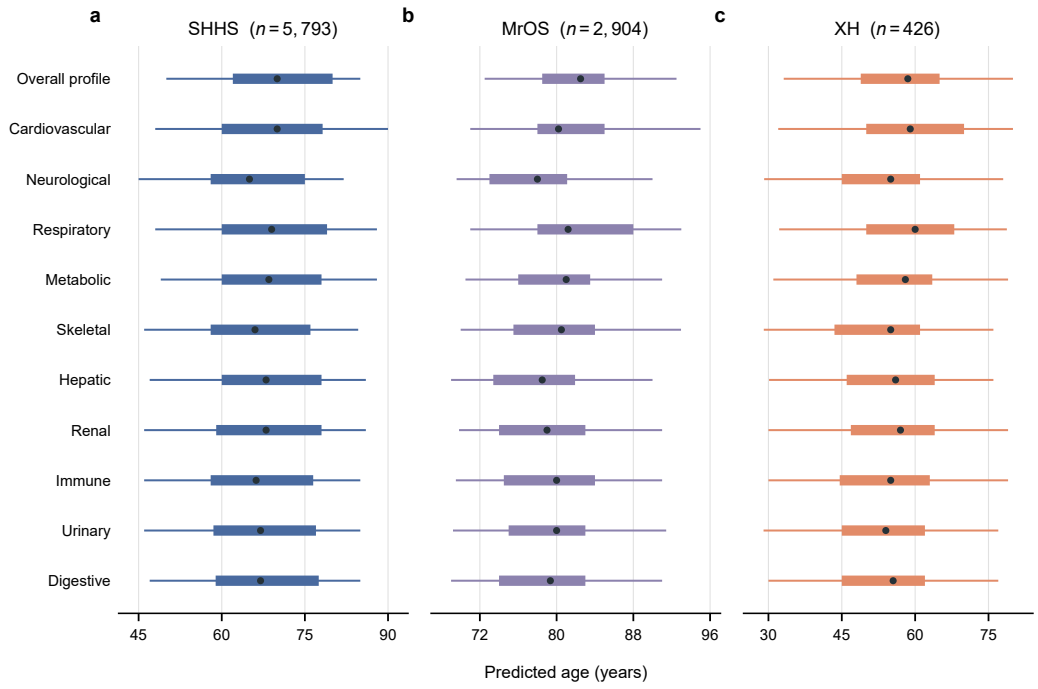

Figure G.1: Distribution of the eleven age representations across cohorts. **a–c**, SHHS ($n = 5{,}793$), MrOS ($n = 2{,}904$), and XH ($n = 426$). Points denote medians, thick bars interquartile ranges, and thin lines the 5th–95th percentiles. Each panel uses a cohort-specific age scale.

H XH MATCHED-GAP INCREMENTAL DISCRIMINATION

We compared seven models across the same XH multi-organ outcomes to test whether an endpoint-matched gap improved discrimination beyond successively richer baselines. Within each outcome, every model used the same participants and repeated cross-validation folds. Sex was omitted for benign prostatic hyperplasia because this outcome was evaluated in men only.

Table H.1: Discrimination of endpoint-matched gaps in XH. Within each endpoint, all seven penalized Cox models use the same participants and ten-repeated five-fold cross-validation splits. The clinical model comprises chronological age, sex, and log-transformed baseline diagnosis burden. Sex is omitted for benign prostatic hyperplasia because this outcome is evaluated in men only. The matched gap is the overall sleep gap for the first-new-diagnosis composite and the corresponding organ-resolved gap for each organ outcome. Entries are Harrell C-indices from averaged out-of-fold risk scores. The final row reports the unweighted mean across all 21 endpoints. Within each row, the highest value is bold and the second-highest distinct value is underlined. Paired-bootstrap intervals and multiplicity-adjusted results are provided in Source Data.

Endpoint	Matched gap	n	Events	Chronological age		Chronological age + sex		Clinical base		Gap only
				Base C	+ gap C	Base C	+ gap C	Base C	+ gap C	C
First new diagnosis category	Overall sleep gap	425	255	0.595	0.610	0.592	0.609	<u>0.620</u>	**0.627**	0.549
Hypertension	Cardiovascular gap	296	86	0.595	**0.661**	0.592	0.659	0.605	<u>0.660</u>	0.650
Coronary heart disease	Cardiovascular gap	360	43	0.766	**0.814**	0.764	<u>0.812</u>	0.756	0.806	0.675
Heart failure	Cardiovascular gap	415	30	0.677	<u>0.740</u>	0.670	0.737	0.727	**0.754**	0.675
Atrial fibrillation	Cardiovascular gap	420	14	<u>0.772</u>	**0.789**	0.747	0.771	0.738	0.753	0.624
Myocardial ischemia	Cardiovascular gap	423	11	0.514	0.773	0.423	0.752	0.671	**0.812**	<u>0.777</u>
Carotid atherosclerosis	Cardiovascular gap	419	20	0.810	0.839	0.820	**0.846**	0.826	<u>0.842</u>	0.640
Cerebral infarction	Neurological gap	406	12	0.866	**0.897**	0.858	<u>0.885</u>	0.855	0.879	0.675
Brain atrophy	Neurological gap	420	12	**0.918**	0.910	<u>0.913</u>	0.908	0.888	0.882	0.589
Cerebral arteriosclerosis	Neurological gap	415	14	**0.874**	<u>0.874</u>	0.858	0.867	0.849	0.857	0.589
Diabetes mellitus	Metabolic gap	391	47	0.589	**0.751**	0.557	<u>0.744</u>	0.544	0.740	0.734
Hyperlipidemia	Metabolic gap	400	35	0.525	**0.643**	0.477	0.626	0.566	0.634	<u>0.635</u>
Hyperuricemia	Metabolic gap	417	20	0.549	0.652	0.530	0.635	0.606	<u>0.665</u>	**0.666**
Cervical spondylosis	Skeletal gap	416	15	**0.582**	<u>0.553</u>	0.508	0.505	0.426	0.434	0.525
Non-alcoholic fatty liver disease	Hepatic gap	389	44	0.583	**0.720**	0.568	<u>0.710</u>	0.607	0.705	0.702
Hepatic insufficiency	Hepatic gap	419	16	0.515	**0.864**	0.446	<u>0.860</u>	0.472	0.844	0.857
Hepatic cyst	Hepatic gap	420	14	<u>0.644</u>	**0.673**	0.598	0.632	0.573	0.595	0.584
Renal cyst	Renal gap	415	18	0.668	0.663	**0.735**	0.716	<u>0.730</u>	0.709	0.567
Benign prostatic hyperplasia	Urinary gap	323	21	<u>0.761</u>	**0.779**	<u>0.761</u>	**0.779**	0.744	0.758	0.563
Gastritis	Digestive gap	391	47	0.539	<u>0.682</u>	0.539	0.677	0.511	0.671	**0.684**
Esophagitis	Digestive gap	414	14	0.430	0.602	0.625	**0.696**	0.585	<u>0.674</u>	0.633
Macro average (all 21 endpoints)	Endpoint-specific	–	–	0.656	**0.738**	0.647	<u>0.735</u>	0.662	0.729	0.647

I ENDPOINT-WISE EVALUATION OF ORGAN CORRESPONDENCE

Each cohort contributes one overall sleep age and ten organ-resolved ages, jointly returned by the framework for every participant. For each organ-linked endpoint, all eleven gaps were evaluated in the same participants and cross-validation folds.

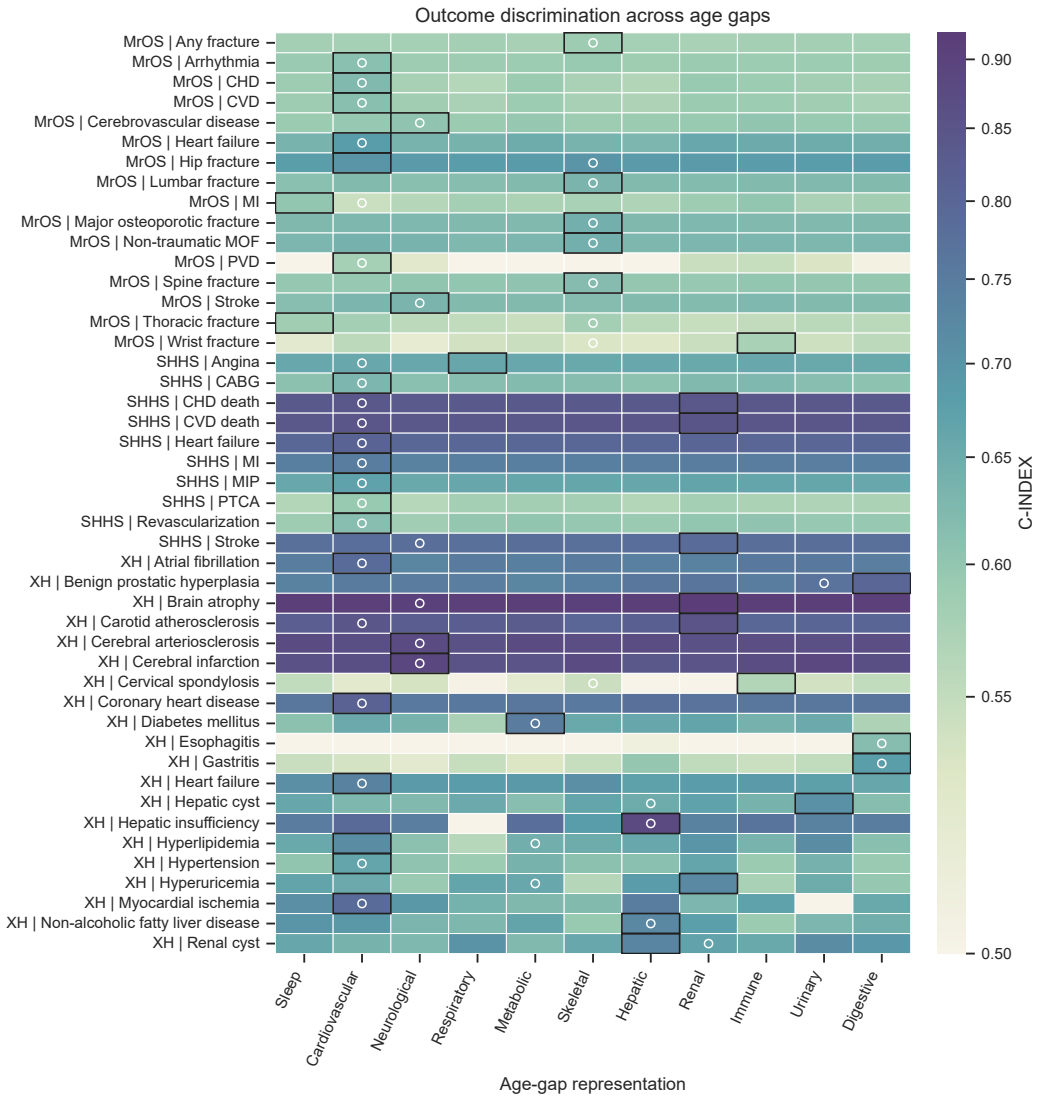

Figure I.1: Endpoint-wise comparison of all eleven age gaps. Each cell shows the repeated out-of-fold Harrell C-index for chronological age plus the indicated gap. Within each endpoint, every gap was evaluated in the same participants and cross-validation folds. A dark outline marks the largest value in a row, and an open circle marks the prespecified endpoint-matched gap. Figure 5 summarizes the matched gap relative to the ten alternatives.

J SHHS SURVIVAL SENSITIVITIES

All curves use the SHHS overall sleep gap and the full available follow-up.

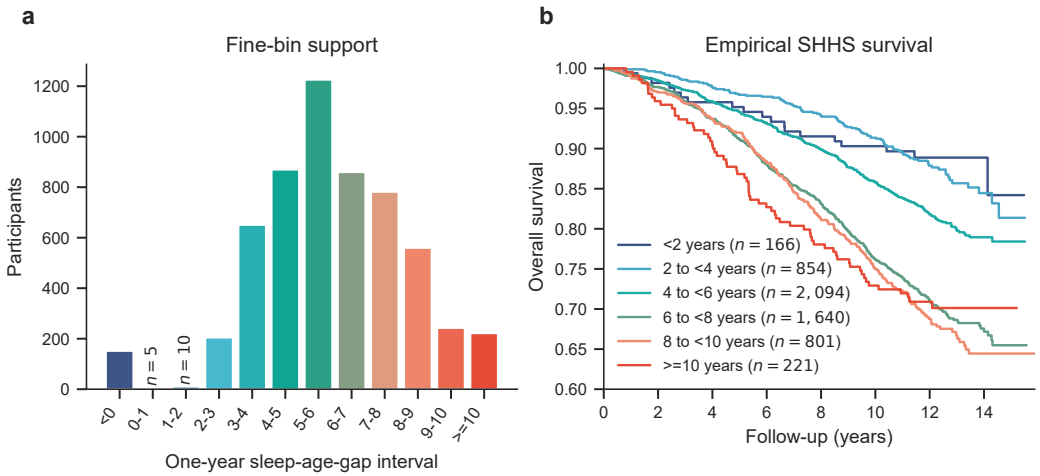

Figure J.1: Survival across SHHS overall sleep-gap intervals. Panel a shows the imbalance in literal one-year bins, including $n = 5$ for 0–1 years and $n = 10$ for 1–2 years. Panel b uses wider intervals and Kaplan–Meier curves for pooled two-year intervals. Local crossings caution against claiming a perfectly monotonic dose response.

K MROS SURVIVAL SENSITIVITIES

MrOS all-cause survival was evaluated using the full available follow-up. The fixed five-year split provides an absolute grouping threshold. The extreme comparison ranks participants by overall sleep gap and takes the lowest and highest 10% (247 participants each). Equal gaps are ordered by participant identifier, allowing tied values to span the rank boundaries.

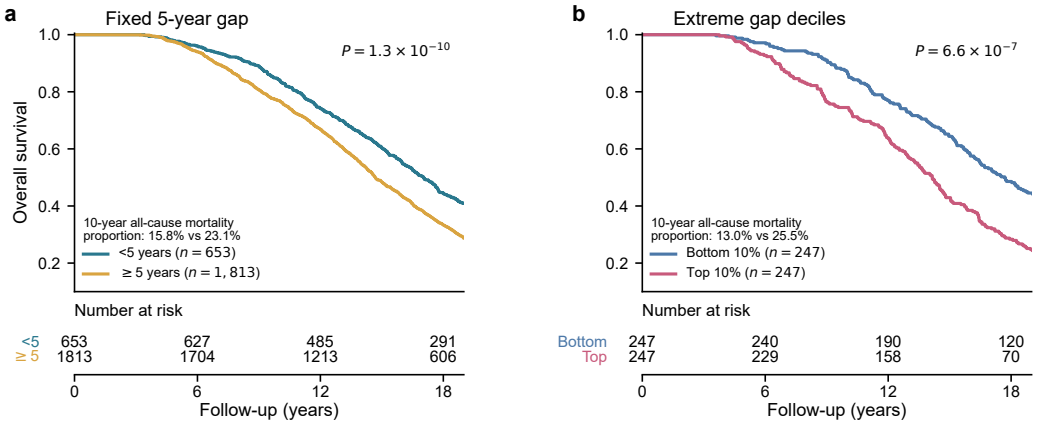

Figure K.1: MrOS all-cause survival sensitivities. Panel a compares overall sleep gaps below versus at least five years. Panel b compares the lowest and highest 10% by gap rank ($n = 247$ each). Risk tables show participants who remain event-free and uncensored immediately before each displayed time point, not the Kaplan–Meier-estimated number of survivors. Displayed P values are from two-sided log-rank tests. At 10 years, Kaplan–Meier estimated mortality was 15.8% versus 23.1% for the fixed split and 13.0% versus 25.5% for the extreme-decile comparison.

L DEVELOPMENTAL EVALUATION OF PHYSIOLOGICAL-CONSISTENCY AUDITING

The frozen Consistency Audit Agent received all 426 XH estimator outputs. It applied the two hard-consistency checks defined in Section 4 and returned PASS or CLINICIAN_REVIEW without changing any age. The full cohort therefore estimates the review rate. An independent model-based developmental reference annotated all 426 outputs for the same two error classes without access to the audit decisions. Error enrichment was assessed using two-sided Wilson 95% confidence intervals and Fisher's exact test. These labels characterize development-stage triage behavior. Independent annotation by clinical sleep experts remains a separate validation step.

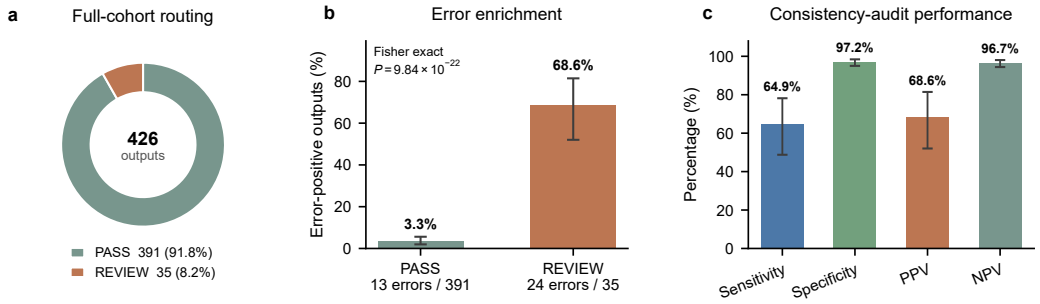

Figure L.1: Developmental evaluation of the Consistency Audit Agent. **a**, PASS and REVIEW dispositions across all 426 XH candidates. This panel shows the review rate, not error prevalence. **b**, Prevalence of reference hard-consistency errors among all PASS and REVIEW cases. Lines denote two-sided Wilson 95% confidence intervals. The displayed P value is from a two-sided Fisher exact test. **c**, Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV), with two-sided Wilson 95% confidence intervals. The agent screens physiological consistency while the frozen age vector remains unchanged. Reference labels came from an independent model-based developmental annotation.

Across all 426 outputs, prespecified errors occurred in 24 of 35 REVIEW outputs (68.6%) and 13 of 391 PASS outputs (3.3%). The two-sided Fisher exact test gave $P = 9.84 \times 10^{-22}$. REVIEW routing captured 24 of 37 reference errors (64.9%), with 97.2% specificity and a 96.7% negative predictive value.

M AGENT CONTRACTS AND COMPLETE PROMPTS

Clinical sleep experts define the registry and its clinical semantics. No agent creates or revises this knowledge layer. The Evidence Orchestration Agent routes Evidence IDs. Registry-guided validation checks the plan, renders canonical values and units, and hard-checks the resulting trace. The Age Estimation Agent jointly returns the complete union-schema vector of overall sleep age and ten organ-resolved ages. The Consistency Audit Agent applies a read-only, high-threshold consistency contract to the complete estimator output. It cannot edit ages, validate evidence support, or select predictions using outcomes.

Evidence Orchestration Agent Prompt

You are the Evidence Orchestration Agent in a multi-agent clinical sleep-aging pipeline.

The local resolver has joined a frozen expert registry to one participant's de-identified baseline evidence warehouse. You receive only registry-eligible evidence. Your task is routing, not prose generation: select nonredundant evidence IDs and place them in the six fixed report sections and clinically coherent group order. The local deterministic renderer will copy the canonical label, exact value and approved display unit.

Preserve cohort heterogeneity. Do not require a metric merely because another cohort has it. The expert registry is the inclusion boundary: route every supplied non-missing record, including negative baseline findings, into its supplied preferred section and clinical group. An item may be omitted only when its value is missing/uninformative or it is an exact duplicate of another supplied record. Records marked required must be used. Do not infer diagnoses and do not create or rewrite any value.

Return exactly one JSON object:

```
{
  "sections": [
    {
      "section": "Basic information | Lifestyle | Physical examination | Sleep study measurements | Family history
| Medical history",
      "groups": [
        {"group": "supplied clinical_group", "evidence_ids": ["evidence_id", "..."]}
      ]
    }
  ],
  "omitted_evidence": [
    {"evidence_id": "evidence_id", "reason": "exact_duplicate | missing_or_uninformative"}
  ]
}
```

Use all six sections exactly once and in the supplied order. Every supplied evidence ID must appear exactly once, either in a section group or in omitted_evidence. Do not return report prose.

Figure M.1: Complete Evidence Orchestration Agent prompt. The agent routes admissible Evidence IDs under the frozen expert registry.

Age Estimation Agent Prompt (1 of 2): Task

Your Role: You are a well-versed scholar in the study of aging, specializing in sleep medicine. You are able to combine various health and sleep indicators to determine a person's sleep biological age by tracing physiological pathways and conducting holistic reasoning.

Background Knowledge: Biological age reflects the body's true functional state, often differing from chronological age. The Overall Sleep Biological Age is a powerful, integrative metric that synthesizes all available sleep and health data. It functions as a key predictor of all-cause mortality, providing a holistic estimate of the body's systemic aging. In parallel, organ-specific ages (e.g., cardiovascular, neurological, metabolic, skeletal, immune, urinary, digestive) reflect the aging trajectory of individual physiological organs and are linked to the risk of organ-specific adverse health outcomes (e.g., heart failure, stroke, chronic obstructive pulmonary disease, type 2 diabetes mellitus, liver disease, kidney disease, and related conditions). Sleep quality is a critical modulator in this process. Poor sleep, characterized by conditions like obstructive sleep apnea-hypopnea syndrome or frequent arousals, accelerates aging, while restorative sleep promotes resilience.

Your Task: Based on the comprehensive health and sleep information of the person below, please predict their overall sleep biological age and ten organ-specific ages: cardiovascular, neurological, metabolic, respiratory, skeletal, hepatic, renal, immune, urinary, and digestive. For each of the eleven age predictions, you must provide four distinct outputs:
1. Aging Physiological Chain: A concise string showing the pathway from specific data to an accelerated aging outcome. You MUST use '->' to connect steps. Focus on detrimental factors.
2. Youthful Physiological Chain: A concise string showing the pathway from specific data to a more youthful or resilient state. You MUST use '->' to connect steps. Focus on protective factors.
3. Chain of Thought (CoT): A narrative reasoning process that holistically weighs the evidence from both physiological chains to justify the specific numerical age you have predicted. This should explain why you chose that particular number by synthesizing all relevant evidence.
4. Predicted Age: The final numerical age prediction.

Age Estimation Agent Prompt (2 of 2): Output Schema

The format of your answer must be a single, valid JSON object, with no additional text outside the JSON structure. Please adhere strictly to the following format:

```json
{
  "overall sleep biological age aging physiological chain": "string",
  "overall sleep biological age youthful physiological chain": "string",
  "overall sleep biological age chain of thought": "string",
  "overall sleep biological age": "float",

  "cardiovascular age aging physiological chain": "string",
  "cardiovascular age youthful physiological chain": "string",
  "cardiovascular age chain of thought": "string",
  "cardiovascular age": "float",

  "neurological age aging physiological chain": "string",
  "neurological age youthful physiological chain": "string",
  "neurological age chain of thought": "string",
  "neurological age": "float",

  "respiratory age aging physiological chain": "string",
  "respiratory age youthful physiological chain": "string",
  "respiratory age chain of thought": "string",
  "respiratory age": "float",

  "metabolic age aging physiological chain": "string",
  "metabolic age youthful physiological chain": "string",
  "metabolic age chain of thought": "string",
  "metabolic age": "float",

  "skeletal age aging physiological chain": "string",
  "skeletal age youthful physiological chain": "string",
  "skeletal age chain of thought": "string",
  "skeletal age": "float",

  "hepatic age aging physiological chain": "string",
  "hepatic age youthful physiological chain": "string",
  "hepatic age chain of thought": "string",
  "hepatic age": "float",

  "renal age aging physiological chain": "string",
  "renal age youthful physiological chain": "string",
  "renal age chain of thought": "string",
  "renal age": "float",

  "immune age aging physiological chain": "string",
  "immune age youthful physiological chain": "string",
  "immune age chain of thought": "string",
  "immune age": "float",

  "urinary age aging physiological chain": "string",
  "urinary age youthful physiological chain": "string",
  "urinary age chain of thought": "string",
  "urinary age": "float",

  "digestive age aging physiological chain": "string",
  "digestive age youthful physiological chain": "string",
  "digestive age chain of thought": "string",
  "digestive age": "float"
}
```

Give you the person's health information as:
{input}
Based on the information above, your answer is:

Figure M.2: Complete Age Estimation Agent prompt. The first panel defines the estimation task, and the second specifies the complete eleven-age output schema.

Consistency Audit Agent Prompt

SYSTEM

You are a clinical safety gate for explanations accompanying eleven sleep-related biological-age estimates.

USER

Read every target block and apply two hard-consistency checks.

For `direction_checks`, inspect the adjacent links but use `GROSS_ERROR` only for an impossible order defined by sleep-measurement timing. Within one sleep episode, sleep-onset latency precedes post-onset wake measures, whereas total sleep time and sleep efficiency summarize the episode; a link that reverses this order is erroneous. General causal direction, risk pathways, marker–disease relationships, and summary states are `CLEAR`.

For `state_consistency`, use `GROSS_ERROR` only when the same named variable is both unknown and assigned a normal or abnormal state, or is explicitly assigned opposing states such as high and low. Different variables and coexisting risk and protective factors are not contradictions.

Evidence support beyond these checks, omitted alternatives, clinical completeness, and numerical-age magnitude are outside this audit.

Return one valid JSON object and no additional text. `direction_checks` must contain one record for each target, in this order: overall_sleep, cardiovascular, neurological, metabolic, respiratory, skeletal, hepatic, renal, immune, urinary, digestive.

```
{
  "direction_checks": [
    {
      "target": "age target",
      "link": "verbatim adjacent link",
      "verdict": "CLEAR | GROSS_ERROR",
      "reason": "concise reason for a gross error, otherwise null"
    }
  ],
  "state_consistency": {
    "verdict": "CLEAR | GROSS_ERROR",
    "target": "age target or null",
    "quote_1": "first verbatim statement or null",
    "quote_2": "second verbatim statement or null",
    "reason": "concise explanation or null"
  },
  "status": "PASS | CLINICIAN_REVIEW"
}
```

Set `status` to `CLINICIAN_REVIEW` if any check is `GROSS_ERROR`; otherwise set it to `PASS`.

AGE ESTIMATOR OUTPUT

{prediction}

Figure M.3: Complete Consistency Audit Agent prompt. The audit applies the two prespecified hard-consistency checks and returns a PASS or CLINICIAN REVIEW decision.

N REPRESENTATIVE AGENT OUTPUTS

The examples were selected without access to longitudinal outcomes. Together, they show a complete routed Evidence-ID plan, a complete eleven-age estimator output, and the structured decision returned by the consistency audit.

Evidence-ID Routing Plan (1 of 2): Baseline Profile


```
{
 "sections": [
  {
   "section": "Basic information",
   "groups": [
    {
     "group": "demographics",
     "evidence_ids": ["e_age", "e0011", "e0012", "e0013", "e0014"]
    }
   ]
  },
  {
   "section": "Lifestyle",
   "groups": [
    {
     "group": "smoking",
     "evidence_ids": ["e0015", "e0016", "e0024", "e0025"]
    },
    {
     "group": "alcohol_and_caffeine",
     "evidence_ids": ["e0017", "e0018", "e0019", "e0020", "e0021", "e0022", "e0023"]
    }
   ]
  },
  {
   "section": "Physical examination",
   "groups": [
    {
     "group": "anthropometry",
     "evidence_ids": ["e0026", "e0027", "e0028", "e0029", "e0030", "e0031"]
    },
    {
     "group": "vital_signs",
     "evidence_ids": ["e0032", "e0033"]
    },
    {
     "group": "pulmonary_function",
     "evidence_ids": ["e0034", "e0035"]
    },
    {
     "group": "laboratory_biomarkers",
     "evidence_ids": ["e0036", "e0037", "e0038"]
    }
   ]
  },
```

Evidence-ID Routing Plan (2 of 2): Sleep and Clinical History


```
{
    "section": "Sleep study measurements",
    "groups": [
     {
       "group": "sleep_quality",
       "evidence_ids": ["e0001", "e0039"]
     },
     {
       "group": "sleep_continuity",
       "evidence_ids": ["e0132", "e0133", "e0134", "e0135", "e0136", "e0137", "e0138"]
     },
     {
       "group": "sleep_architecture",
       "evidence_ids": ["e0139", "e0140", "e0141", "e0142", "e0143", "e0144", "e0145", "e0146"]
     },
     {
       "group": "fragmentation_arousal",
       "evidence_ids": ["e0153", "e0154", "e0155", "nx0001", "nx0002", "nx0003", "nx0004", "nx0005"]
     },
     {
       "group": "respiratory_events",
       "evidence_ids": ["e0166", "e0167", "e0168", "e0169", "e0170", "nx0006", "nx0007", "nx0008", "nx0009",
"nx0010", "nx0011", "nx0012", "nx0013"]
     },
     {
       "group": "oxygenation_burden",
       "evidence_ids": ["e0197", "e0198", "e0199", "e0200", "e0201", "e0202", "e0204", "e0205", "e0206",
"nx0014", "nx0015", "nx0016", "nx0017", "nx0018"]
     },
     {
       "group": "eeg_microstructure",
       "evidence_ids": ["e0236", "e0238", "e0240", "e0241", "e0242", "e0244", "e0248", "e0250", "e0257",
"e0258", "e0262", "e0263", "e0264", "e0268", "e0283"]
     }
    ]
   },
   {
    "section": "Family history",
    "groups": [
     {
       "group": "family_history",
       "evidence_ids": ["u_family_father", "u_family_mother"]
     }
    ]
   },
   {
    "section": "Medical history",
    "groups": [
     {
       "group": "medical_history",
       "evidence_ids": ["u_medical_history"]
     },
     {
       "group": "medications",
       "evidence_ids": ["u_medications"]
     }
    ]
   }
  ],
  "omitted_evidence": []
}
```


Figure N.1: Representative Evidence-ID routing plan. The two panels form one complete JSON object for a de-identified participant. The first records basic information, lifestyle, and physical-examination groups. The second records sleep-study measurements, family history, medical history, and medications.

Eleven-Age Estimation Output (1 of 3): Overall to Metabolic


```
{
  "overall_sleep_aging_physiological_chain": "Poor sleep quality -> Frequent obstructive sleep apnea events -> Increased oxidative stress and inflammation -> Accelerated cellular aging",
  "overall_sleep_youthful_physiological_chain": "Minimal alcohol consumption -> Lower risk of sleep disruptions -> Better sleep quality -> Enhanced cellular repair and resilience",
  "overall_sleep_chain_of_thought": "Considering the person's severe obstructive sleep apnea hypopnea syndrome, high blood pressure, and elevated inflammatory markers, their overall sleep biological age is likely higher than their chronological age. However, their minimal smoking history and alcohol consumption may have a slightly protective effect. Weighing these factors, I predict their overall sleep biological age to be around 85 years old.",
  "overall_sleep_biological_age_years": 85.0,

  "cardiovascular_aging_physiological_chain": "Hypertension -> Increased cardiac workload -> Cardiac remodeling and fibrosis -> Decreased cardiovascular function",
  "cardiovascular_youthful_physiological_chain": "Minimal smoking history -> Lower risk of cardiovascular disease -> Better cardiovascular health -> Enhanced cardiovascular resilience",
  "cardiovascular_chain_of_thought": "Given the person's hypertension, high systolic and diastolic blood pressure, and severe obstructive sleep apnea hypopnea syndrome, their cardiovascular age is likely higher than their chronological age. However, their minimal smoking history may have a slightly protective effect. Considering these factors, I predict their cardiovascular age to be around 90 years old.",
  "cardiovascular_biological_age_years": 90.0,

  "neurological_aging_physiological_chain": "Severe obstructive sleep apnea hypopnea syndrome -> Chronic intermittent hypoxia -> Neuroinflammation and oxidative stress -> Neurodegeneration",
  "neurological_youthful_physiological_chain": "Minimal alcohol consumption -> Lower risk of neurodegeneration -> Better neurological health -> Enhanced cognitive function",
  "neurological_chain_of_thought": "The person's severe obstructive sleep apnea hypopnea syndrome and high blood pressure may contribute to an increased neurological age. However, their minimal alcohol consumption may have a slightly protective effect. Weighing these factors, I predict their neurological age to be around 80 years old.",
  "neurological_biological_age_years": 80.0,

  "metabolic_aging_physiological_chain": "Hypertension -> Insulin resistance -> Metabolic dysregulation -> Increased risk of metabolic disorders",
  "metabolic_youthful_physiological_chain": "Minimal alcohol consumption -> Lower risk of metabolic disorders -> Better metabolic health -> Enhanced insulin sensitivity",
  "metabolic_chain_of_thought": "The person's hypertension, high blood pressure, and elevated glucose levels may contribute to an increased metabolic age. However, their minimal alcohol consumption may have a slightly protective effect. Considering these factors, I predict their metabolic age to be around 82 years old.",
  "metabolic_biological_age_years": 82.0,
```

Eleven-Age Estimation Output (2 of 3): Respiratory to Renal

"respiratory_aging_physiological_chain": "Severe obstructive sleep apnea hypopnea syndrome -> Chronic respiratory stress -> Inflammation and oxidative stress -> Decreased respiratory function",
"respiratory_youthful_physiological_chain": "Minimal smoking history -> Lower risk of respiratory disease -> Better respiratory health -> Enhanced lung function",
"respiratory_chain_of_thought": "Considering the person's severe obstructive sleep apnea hypopnea syndrome and high respiratory event index, their respiratory age is likely higher than their chronological age. However, their minimal smoking history may have a slightly protective effect. Weighing these factors, I predict their respiratory age to be around 88 years old.",
"respiratory_biological_age_years": 88.0,

"skeletal_aging_physiological_chain": "Gout -> Chronic inflammation -> Bone resorption and osteoporosis -> Decreased skeletal function",
"skeletal_youthful_physiological_chain": "Minimal smoking history -> Lower risk of osteoporosis -> Better skeletal health -> Enhanced bone density",
"skeletal_chain_of_thought": "Given the person's gout and high uric acid levels, their skeletal age may be higher than their chronological age. However, their minimal smoking history may have a slightly protective effect. Weighing these factors, I predict their skeletal age to be around 84 years old.",
"skeletal_biological_age_years": 84.0,

"hepatic_aging_physiological_chain": "Hypertension -> Increased liver workload -> Liver fibrosis and inflammation -> Decreased liver function",
"hepatic_youthful_physiological_chain": "Minimal alcohol consumption -> Lower risk of liver disease -> Better liver health -> Enhanced liver function",
"hepatic_chain_of_thought": "The person's hypertension and high blood pressure may contribute to an increased hepatic age. However, their minimal alcohol consumption may have a slightly protective effect. Considering these factors, I predict their hepatic age to be around 81 years old.",
"hepatic_biological_age_years": 81.0,

"renal_aging_physiological_chain": "Hypertension -> Increased renal workload -> Renal fibrosis and inflammation -> Decreased renal function",
"renal_youthful_physiological_chain": "Minimal smoking history -> Lower risk of renal disease -> Better renal health -> Enhanced renal function",
"renal_chain_of_thought": "Given the person's hypertension, high blood pressure, and elevated creatinine levels, their renal age is likely higher than their chronological age. However, their minimal smoking history may have a slightly protective effect. Weighing these factors, I predict their renal age to be around 89 years old.",
"renal_biological_age_years": 89.0,

Eleven-Age Estimation Output (3 of 3): Immune to Digestive

"immune_aging_physiological_chain": "Severe obstructive sleep apnea hypopnea syndrome -> Chronic inflammation -> Immune dysregulation -> Decreased immune function",
"immune_youthful_physiological_chain": "Minimal alcohol consumption -> Lower risk of immune disorders -> Better immune health -> Enhanced immune function",
"immune_chain_of_thought": "The person's severe obstructive sleep apnea hypopnea syndrome and high inflammatory markers may contribute to an increased immune age. However, their minimal alcohol consumption may have a slightly protective effect. Considering these factors, I predict their immune age to be around 83 years old.",
"immune_biological_age_years": 83.0,

"urinary_aging_physiological_chain": "Prostate hyperplasia -> Increased urinary frequency and urgency -> Urinary tract inflammation -> Decreased urinary function",
"urinary_youthful_physiological_chain": "Minimal smoking history -> Lower risk of urinary disorders -> Better urinary health -> Enhanced urinary function",
"urinary_chain_of_thought": "Given the person's prostate hyperplasia and high prostate-specific antigen levels, their urinary age may be higher than their chronological age. However, their minimal smoking history may have a slightly protective effect. Weighing these factors, I predict their urinary age to be around 86 years old.",
"urinary_biological_age_years": 86.0,

"digestive_aging_physiological_chain": "Gout -> Chronic inflammation -> Gut dysbiosis and inflammation -> Decreased digestive function",
"digestive_youthful_physiological_chain": "Minimal alcohol consumption -> Lower risk of digestive disorders -> Better digestive health -> Enhanced digestive function",
"digestive_chain_of_thought": "The person's gout and high uric acid levels may contribute to an increased digestive age. However, their minimal alcohol consumption may have a slightly protective effect. Considering these factors, I predict their digestive age to be around 80 years old.",
"digestive_biological_age_years": 80.0
}

Figure N.2: Representative complete eleven-age estimation output. The three panels form one JSON object containing the overall sleep age and ten organ-resolved ages. Each estimate is accompanied by aging and resilience chains, an explanatory synthesis, and the numerical age.

Consistency Audit Output (1 of 2): Direction Checks


```
{
  "direction_checks": [
    {
      "target": "overall_sleep",
      "link": "Poor sleep quality -> Frequent arousals",
      "verdict": "CLEAR",
      "reason": null
    },
    {
      "target": "cardiovascular",
      "link": "Hypertension -> Increased blood pressure",
      "verdict": "CLEAR",
      "reason": null
    },
    {
      "target": "neurological",
      "link": "Poor sleep quality -> Increased stress",
      "verdict": "CLEAR",
      "reason": null
    },
    {
      "target": "metabolic",
      "link": "Type 2 diabetes -> Insulin resistance",
      "verdict": "CLEAR",
      "reason": null
    },
    {
      "target": "respiratory",
      "link": "Frequent apnea events -> Increased respiratory stress",
      "verdict": "CLEAR",
      "reason": null
    },
    {
      "target": "skeletal",
      "link": "Unknown skeletal health -> Potential skeletal dysfunction",
      "verdict": "CLEAR",
      "reason": null
    },
```

Consistency Audit Output (2 of 2): Review Decision


```
{
    "target": "hepatic",
    "link": "Unknown liver function -> Potential liver dysfunction",
    "verdict": "CLEAR",
    "reason": null
  },
  {
    "target": "renal",
    "link": "Unknown kidney function -> Potential kidney dysfunction",
    "verdict": "CLEAR",
    "reason": null
  },
  {
    "target": "immune",
    "link": "Poor sleep quality -> Increased stress",
    "verdict": "CLEAR",
    "reason": null
  },
  {
    "target": "urinary",
    "link": "Unknown urinary health -> Potential urinary dysfunction",
    "verdict": "CLEAR",
    "reason": null
  },
  {
    "target": "digestive",
    "link": "Post-gastric cancer surgery -> Potential digestive dysfunction",
    "verdict": "CLEAR",
    "reason": null
  }
 ],
 "state_consistency": {
   "verdict": "GROSS_ERROR",
   "target": "hepatic",
   "quote_1": "Unknown liver function -> Potential liver dysfunction",
   "quote_2": "Normal liver function -> Reduced hepatic risk",
   "reason": "Same variable (liver function) is described as unknown and normal, creating contradictory states"
 },
 "status": "CLINICIAN_REVIEW"
}
```


Figure N.3: Representative Consistency Audit Agent output. Direction checks were clear, whereas the state-consistency check routed the output to CLINICIAN REVIEW because liver function was described as both unknown and normal. The audit records the inconsistency without altering the frozen age vector.

O EVIDENCE REGISTRY AND VALIDATION PROCEDURE

The evidence registry specifies the provenance, canonical label, unit, report section, clinical group, redundancy group, and missingness rule for every admissible variable. Sleep medicine experts curated these entries and routing constraints from the three cohort dictionaries before profile decoding. The registry contains definitions rather than participant values. Participant values enter only when individual evidence reports are rendered.

When a variable is unavailable in a cohort, its registry rule determines how it is represented in the report. Registry-guided validation confirms that every routed Evidence ID belongs to the assigned report section and that each rendered label, value, and unit matches the source field. Trace checks identify duplicate or conflicting entries before the report enters the Age Estimation Agent.