nature portfolio

Peer Review File

Sleep chart of biological aging clocks in middle and late life

Corresponding Author: Dr Junhao Wen

Any redactions in this file are there to maintain patient confidentiality, the confidentiality of unpublished data, or to remove third-party material

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Referee #1

(Remarks to the Author)
In this manuscript, led by Prof. Junhao Wen, Cliodhna O'Toole and colleagues across the MULTI study, the authors conduct one of the largest and most comprehensive studies of the relationship between sleep disturbances and health outcomes to date. Combining clinical, genetic, proteomic, metabolomic and MRI data across several datasets including the UK Biobank, Finngen, Psychiatric Genomics Consortium, TriNetX and the Multi-Ethic Study of Atherosclerosis, the study conducts a series of genome-wide association studies, genetic correlation analyses, causal inference studies, proteome- and metabolome-wide associations, survival analyses and structural equation modeling to reveal several important insights. First, these analyses highlight how the well-described U-shaped relationship between sleep disturbances and health outcomes (whereby both shorter sleep, defined as less than six hours per night, and longer sleep, defined a greater than eight hours per night, are associated with adverse outcomes) extends beyond the central nervous system, encompassing strong associations with biological age gaps across pulmonary, hepatic, immune, skin and endocrine biological age, in addition to brain. Second, the study demonstrates that this U-shaped relationship is characterized by alterations across multiple biological pathways impacting inflammation, tissue repair dysfunction, neurogenesis, metabolic homeostasis, coagulation, and innate defense pathways - implicating more systemic modulation of the sleep-health relationship than previously understood. Third, the study illustrates that sleep disturbances are genetically correlated with over 150 illnesses across multiple bodily systems - with shorter sleep (<6 hours) showing the most prominent and widespread associations. Fourth, the study investigates the more nuanced, indirect relationship between long sleep (>8 hours) and a narrower collection of health outcomes - paying particularly close attention to late-life depression (LLD). The authors highlight that both short sleep and long sleep contribute to LLD, but while the former likely influences the illness more directly via immune dysregulation and HPA axis activation, the latter contributes more indirectly via accelerated brain and adipose tissue aging - potentially reflecting physiological compensatory mechanisms or subclinical processes related to aging.

Overall, this is an impressive manuscript. It represents one of the first large-scale, systematic investigations of the relationship between sleep disturbances, biological age and related health outcomes, constructing a fine-grained multi-omic atlas of the complex interplay between sleep and health. The authors should be commended for the rigor with which they have conducted the analyses, the precise but accessible writing style, the integration of multiple datasets and analytical approaches, and the creation of an intuitive online browser facilitating interactive perusal of the results. The contents of the manuscript would be of high relevance to the general readership of Nature. Indeed, more general audiences may find the study equally as interesting given the cultural relevance of the subject matter (see, for example, the relatively recent novel, 'My Year of Rest and Relaxation').

Although my general impression of the manuscript is positive, I outlined several questions and recommendations for the authors below.

First, one of the study's major limitations - already acknowledged by the authors - is its cross-sectional design, limiting the ability to determine causality. Causality is of paramount importance in any investigation of the relationship between sleep and complex phenotypes like major depressive disorder - prompting the question of whether depressed individuals have disturbed sleep due to their depression, or whether these individuals are depressed due to disturbed sleep. In the absence of longitudinal data from clinical intervention studies, it is challenging to answer that question effectively. Nonetheless, the authors attempt to address the question indirectly via a series of Mendelian randomization (MR) analyses, which ultimately revealed "no evidence for reverse causality". This may be a contentious assertion, considering that a previous (albeit less strongly powered) study found evidence of a bidirectional relationship between sleep and depression. To ensure that the present MR analysis is robust and that the general readership more readily accepts the conclusion that no evidence of

reverse causality could be detected, I would recommend that the authors address the following:

(1) Highlight how five distinct MR methods were employed in the main text; this is currently buried in the supplement. It adds weight to the claim that no evidence for reverse causality was detected,
(2) Report per-SNP and conditional F-statistics to assess instrument bias (considering that the GWASs of short and long sleep only produced a small number of genome-wide significant SNPs, it is possible that the instruments are too weak to detect any reverse causality); alternatively or additionally, please report whether a formal MR power analysis was conducted,
(3) Consider integrating MR-PRESSO (Mendelian Randomization PIeiotropy RESidual Sum and Outlier) as a sixth method to assess horizontal pleiotropy and robustness of findings (not an essential integration, but if feasible, strongly recommended),
(4) Alternatively and/or additionally to the above, please conduct a non-linear Mendelian randomization approach like fractional polynomal MR (e.g., https://rdrr.io/github/jrs95/nlmr/man/fracpoly_mr.html). A U-shaped causal effect may be attenuated using linear MR methods. This is a more essential integration than the MR-PRESSO mentioned above.

Second, while the authors adjust for sex in all analyses and highlight several clear sex differences in the relationship between sleep and biological aging, I noticed that the U-shaped curve is at least visually apparent for males but not females across several biological age gaps including heart protein BAG, endocrine protein BAG and potentially heart MRI BAG. As an alternative to adjusting for sex status and interaction terms in the models, have the authors considered conducting sex-stratified analyses for these BAGs to determine if the U-shaped relationship is significant for males but not females?

Additional, given the wealth of hormonal blood biochemistry readily available in the UK Biobank, including oestradiol, SHBG, and testosterone levels, have the authors considered whether the sleep - health outcome slope is modulated by these variables?

Beyond the two issues outlined above, other pieces of minor feedback are listed below:

• Line 124: Define "GAMs" as 'generalized additive models' before shortening to its acronym
• Line 161: For clarity / easier interpretation, I would recommend updating the first sentence of this paragraph to, "Overall, across the nine significant protein, metabolomic and MRI-based BAGs, optimal sleep duration ranged from 6.5 to 7.8 hours for females and 6.4 to 7.7 hours for male."
• Lines 219-233: This paragraph is a little messy; it may be more intuitive if the disease codes are relegated to the supplementary materials and replaced in the main text with genetic correlation coefficient values.
• Line 246: Figure 2A is not particularly informative; if the authors decide to retain this part of the figure, I would recommend appending the genetic locus with the closest genes to the significant SNPs. Figure 2B shows no significant associations for long sleep - even at a nominal level.
• Line 276: Please clarify if "disturbed sleep duration" refers to long, short, or both.
• Line 290: In Figure 3 (as well as Figure 2C), I was expecting the brain ICD-10 codes to be closer to the visualized brain area on the body map. If possible, I would rearrange the ICD-10 codes on both images so that they more closely correspond with their respective depictions on the anatomical map on the right.
• Line 290: Figure 3 part B - I would recommend switching "short" and "long" in the axis labels to ensure they are aligned with the labels in the inner diagram. At the moment, "short (<6 hrs) aligns with "sleep_long" visually, which is counterintuitive for the reader.
• Lines 310-311: Be more specific about the LLD1 and LLD2 outcomes; they are currently described as "AI- and MRI-derived outcomes of late-life depression", which I first interpreted as two different measures, one being AI-derived and the other being MRI-derived. I would recommend changing this to either "MRI-derived" or "MRI-based, AI-derived".
• Lines 315-316: Outline what c2 and a1 refer to in the context of structural equation modeling; this will be helpful for a more general audience.
• Lines 408: It is worth noting that the UK Biobank's MRI study has a well known bias towards healthier, well educated people. Therefore, while the authors' assertion that reverse causality may 'pull' the right side of the U-shaped curve upward, it may also simply be the case that protein BAGs and metabolomic BAGs are derived from a more representative collection of participants encompassing a wider range of health states versus those who participated in the MRI sub-study.
• Line 665: It is worth noting that significant redundancies exist in the proteomic data; many proteins show high correlations with one another. As such, Bonferroni adjustment may be overly stringent compared with FDR adjustment or other approaches. This is worth at least noting in the text, if the authors decide to retain the more stringent analyses in the main study.
Line 665: Another point on the same analysis is that kidney functon(via markers like ALT/AST) impacts concentrations of many of these proteins, but was not adjusted for in the model. Although it might be extreme for the authors to completely re-conduct these analyses adjusting for kidney function, it might be worth checking whether the most significant findings remain significant after adjusting for these covariates.

Referee #2

(Remarks to the Author)
Overall assessment
This manuscript makes a timely and impactful contribution, with strong relevance for both the scientific community and the

public. The authors introduce a “Sleep Chart” that systematically maps nonlinear (U-shaped) associations between sleep duration and biological aging clocks across multiple organs and omics layers, with a strong effect for shorter than normal sleep. Leveraging the UK Biobank and several independent cohorts, the study provides compelling evidence that both short and long sleep duration accelerate biological aging and increase disease risk. The systemic approach, spanning imaging, proteomics, metabolomics, and genetics, represents an advance over prior brain-focused work, or blood proteomics-focused work.

Originality and significance

• Novelty: This work extends sleep–aging research beyond the brain to a multi-organ, multi-omics framework, incorporating imaging, proteomics, and metabolomics.
• Significance: It leverages biological aging clocks as systemic markers, bridging molecular biology, imaging, and clinical outcomes.
• Integration: The study unifies genetic correlation, survival analysis, and mediation modeling, highlighting both direct and indirect pathways linking sleep to disease.
• Public health implications: Sleep is modifiable, and the identification of organ- and sex-specific optimal durations makes this clinically and societally relevant.

Data and methodology

• Strengths:
o Rigorous and comprehensive methodology with convergent evidence across modalities.
o Replication in independent datasets enhances confidence.
o An interactive portal for data exploration improves transparency, and promotes engagement.
• Points for consideration:
o Optimal durations differ across modalities (e.g., ~7.7h proteomics vs. ~6.5h MRI); please discuss how this discrepancy relates to biological interpretation.
o Proteomic and metabolomic signatures are dynamic; consider discussing temporal variability and the value of longitudinal data.
o Clarify cross-modality alignment: are imaging, proteomic, and metabolomic brain-age gaps concordant, or pathways?

Appropriate use of statistics and treatment of uncertainties

• Statistical approaches are appropriate and clearly applied, including nonlinear modeling, mediation, and replication tests.
• Please specify more explicitly how the “optimal model” was defined (EDF, smoothing parameters, model fit).
• Age-stratified analyses appear non-significant; could the authors should briefly discuss possible reasons (sample size, age range).

Conclusions: robustness, validity, reliability

• The conclusions are generally robust and well-supported by the evidence.
• Replication across independent datasets lends credibility.
• The discussion could directly address the challenges of replicating UKBB-scale studies and the value of complementary smaller cohorts.

Suggestions for revision/Questions

1. Different optimal sleep durations: Provide a biological rationale for differing optima across organ clocks.
2. Circadian timing and sleep structure: Briefly note whether circadian misalignment or fragmentation might influence associations.
3. Cross-modality alignment: Clarify how organ-specific imaging vs. proteomics results relate.
4. Dynamics of omics data: Acknowledge variability and value of longitudinal measures.
5. Definitions:
o Define GTEx and PGC on first mention.
o Specify that F419, G473 are ICD codes.
o Correct typos (e.g., “staitsics”).
6. Repetition: Streamline duplicated description in Supplementary Note 4 and File 8.
7. Figure 4: Clarify whether it involves a second visit.
8. Replication challenges: Explicitly discuss limitations of “stitching” multiple datasets versus controlled cohorts.
9. ProtBAG/MetBAG interpretation: Clarify how blood-based clocks are linked to multiple organs (direct vs. MAGMA inference).
10. Optimal model criteria: Define explicitly.
11. Age-specific effects: Note possible reasons for null results.
12. Table 1: Clarify that MULTI demographics aggregate existing datasets.

References

• Rather comprehensive, but ensure appropriate credit is given to prior sleep–aging studies focused on the brain and blood proteomics.

Clarity and context

• The abstract well-structured, and perhaps could briefly emphasize the multi-organ novelty.
• The introduction provides strong context but should highlight how this study differentiates from prior brain-only or proteomics-only approaches.
• Conclusions are impactful and balanced, and could acknowledge replication challenges directly.

Recommendation

This is a landmark study. I recommend publication after minor clarifications. The points above are not fundamental flaws but may strengthen clarity, reproducibility, and interpretability.

Referee #3

(Remarks to the Author)
Summary
The authors report associations between self-reported sleep duration and a large number of outcomes and quantify these associations as 'biological age gaps (BAGS)'. The emphasis is on U-shaped associations, and this in particular for non-brain outcomes. Genetic analyses, Mendelian randomization approaches and mediation analyses suggest that 1) mechanisms underlying associations between short sleep and outcomes are different from those for long sleep; 2 The sleep duration dependent BAGS (and derived optimal sleep duration) differ between men and women, and between organs, etc. Overall, the data suggest that modifying sleep duration may have substantial 'health benefits and modifying sleep duration may be feasible. Overall ,this is an impressive set of analyses pulling many data and analyses together which makes it in principle a significant contribution.

Critique
As with so many other biobank-based studies, this remains a descriptive study. The importance of sleep for health is well recognised and documented and U-shaped associations have been reported repeatedly for sleep duration when based on self-report, but less often when based on objective sleep duration assessment. In fact, I am aware of at least one study in which both objective and sleep-reported sleep duration data were available, and the associations were strikingly dissimilar, with linear associations between sleep duration and outcomes for objective sleep duration, and U-shaped associations for self-reported sleep duration (see Fig 2 in PMID: 36892074). In general, dissociations between objective and subjective measures of sleep duration are more pronounced in specific populations or health conditions such as depression, insomnia, etc. Thus, the major and significant limitation of the study is the sleep data on which it is based.
Leaving the limitations of the major exposure variable aside, a few other aspects need to be mentioned. The age range of the participants in the UKBB and other cohorts is limited and most of the participants are older. Sleep duration and sleep need change dramatically over the lifespan. Thus, any statement about optimal sleep duration needs to be qualified with 'For people aged X to Y the optimal sleep duration is'. The authors also report that optimal sleep duration varies across organs, etc. but the statistical underpinning of the significance of these differences was not clear to me. This also holds for the sex differences (which are likely to be related to a sex related sleep need). Care needs to be taken when using the word 'optimal'. For example the first sentence of the Abstract now reads 'Optimal sleep plays a vital role in promoting healthy aging and enhancing longevity'. Obviously, this statement is necessarily true and contains no information.
From a naive sleep and health research perspective a first question to be asked to these data would be 'Is there a significant association between sleep duration and outcome'. The second question is ' what is the nature of the association'. Here the authors jump almost immediately to the second question, which is unfortunate. In addition, in several places in the MS the authors use the phrase 'beyond the brain'. I don't understand this emphasis on 'beyond the brain'. The field accepts that sleep has benefits for both the brain and body.
The general message and the presented evidence that sleep is a modifiable risk factor is attractive. However, and given the U-shaped associations so much emphasized by the authors, the question whether self-reported long sleep is a cause or a consequence of outcomes becomes rather important. If it is a cause, then long sleepers should reduce their time in bed. If it is a consequence, then the advice may be rather different. The authors do not seem to be able to come to a clear conclusion with respect to this question (see for example the Abstract but also elsewhere in the manuscript), which is a very important question in the sleep-health field.

Version 1:

Reviewer comments:

Referee #1

(Remarks to the Author)
The authors have thoroughly addressed all of my comments and comprehensively integrated my suggested analyses. I have no further comments and wish them the best of luck with the remainder of the submission process.

Referee #2

(Remarks to the Author)
I thank the authors for their comprehensive and thoughtful revisions. The manuscript is improved, and the majority of my prior concerns have been fully addressed. I have only a small number of remaining suggestions that focus on transparency, clarity, and reproducibility.
1. Supplementary display of the remaining 14 BAGs
Figure 1 is elegant and appropriately highlights the 9 significant BAGs, which should remain in the main figure. For transparency and to allow readers to examine the full modeling results, I suggest adding a supplementary figure showing the GAM curves for the remaining 14 BAGs.
A single multi-panel supplementary figure would reassure readers that all 23 BAGs were analyzed using the same modeling framework, the absence of significance in the remaining BAGs reflects the modeled results rather than omission, and the U-shaped pattern is not selectively shown.
2. Clarification regarding MR instrument strength (Line ~1255)

Although several SNP-level R2 values are small, this is entirely typical for MR instruments of complex behavioral and physiological traits. What matters for instrument strength is the aggregated F-statistic, not the individual R2 values.
I recommend adding a brief clarifying statement noting that the combined variance explained is R2=0.00113 , this corresponds to a robust F-statistic ≈ 38.7 (N = 274,330; k = 8), which exceeds the conventional F > 10 threshold defining adequate instruments in univariable Mendelian randomization. This is well done for the subsequent figure already.
3. Optional note distinguishing EDF from effect size
Some readers may misinterpret the effective degrees of freedom (EDF) in the GAM models as reflecting the strength or magnitude of the association. A brief sentence in the Methods or Supplementary Note explaining that EDF reflects spline flexibility (curve complexity), not effect size, would preempt this misunderstanding.
4. Optional Supplementary table with full GAM statistics
Although not required, a small supplementary table listing—for all 23 BAGs—the chosen family distribution, EDF, P-values, and relevant covariates would further enhance transparency and reproducibility. You already report these statistics for the significant BAGs; extending this to all outcomes would help readers who wish to examine the modeling details more closely.
5. Optional acknowledgment of MR linearity assumptions
Given that the primary phenotypic analyses emphasize U-shaped nonlinear associations, it may be helpful to add a sentence noting that MR methods assume linear genetic associations with the exposure. This is standard for MR but would help frame the interpretational scope of causal inference in contrast to the observed nonlinear sleep–phenotype relationships.
Overall assessment
These suggestions are minor and aimed solely at enhancing clarity and transparency. The scientific content is strong and comprehensive, and the revisions have been thorough and thoughtful, going beyond what was expected. I believe the manuscript will be further strengthened by addressing items (1) and (2), with the optional items providing additional clarity for interested readers.

Referee #3

(Remarks to the Author)
The Authors have been responsive to my comments. One of my major concern was that for the reader of the abstract is was not clear that all the findings are based on sleep-reported sleep duration which is very different from behaviour or physiology based assessment of sleep. From the Abstract it is now clear that all results are based on self-reports in a cross-sectional study. This is a limitation but the results are impressive and important.

Referee #4

(Remarks to the Author)
Summary:
This paper provides an in-depth analysis of the association of self-reported short and long sleep duration with multiple biological clocks (BAGs) using diverse samples. Further, the authors conduct GWAS to identify genetic variants for sleep duration; genetic correlations among sleep and other variables; MR (to explore disease-sleep duration causal associations); incident analyses (to identify role of sleep duration in mortality and MRI markers of LLD); and mediation (SEM) models to explore biological pathways linking sleep to disease. They report novel and strong U-shaped associations with BAGs (consistent with an extensive literature reporting similar associations with several chronic diseases). The reported differences in short and long sleep duration associations also vary by tissue and BAG source, as well as by sex. They point out that they observe more associations for short vs long sleep with incident disease, although both were associated with mortality. They report that mediation analysis demonstrated stronger direct effects between short sleep and MRI-LLD, while associations for long sleep were partially “mediated” through MRIBAGs. The paper proposes many intriguing biological causal pathways that differ for long and short sleep. Given the limitations of data and analyses, its strongest and most novel finding is that self-reported sleep duration is associated with multiple BAGs-with variation by tissue source/BAG data- in a non-linear pattern which varies by sex.

Originality:
The paper presents a comprehensive and highly original account of how multimodal and organ-specific biological aging estimates relate to self-reported sleep duration, showing the expected (from prior disease-specific associations) nonlinear pattern, with accelerated aging associated with both shorter and longer sleep duration. The use of multiple data modalities (MRI, proteomics, and metabolomics) and organ-specific measures reveals a detailed yet remarkably consistent U-shaped association, providing estimates of “optimal” self-reported sleep duration for each organ system (although with caveats noted below). The authors also employ a diverse set of analyses to examine the genetic underpinnings, causal effects, and predictive power of short and long sleep duration in relation to various health outcomes. While these latter analyses are complementary, several analyses (GWAS, incident disease assessments, MR, etc) have been reported before using the same data sources.

Methods, data quality, analyses and conclusions:
The very limited reference to methodological details in the Results section made it challenging to understand which specific sample (cohort) and statistical model (covariates included) were used for each set of analyses. Given the heterogeneity across samples and need to develop statistical models specific for each hypothesis, it was difficult to fully appreciate the strength of the findings. Providing a bit more on the methodological details in the Results would improve interpretation of the results.

Given the interest in this paper in causal inference, it would be helpful to see a justification for inclusion of variables in each

set of analyses. For example, smoking and alcohol both influence sleep and many of the outcomes of interest, but did not appear to be considered. Other important variables that are confounders or precision variables such as physical activity and socioeconomic status should be considered.

The authors nicely acknowledge the limitations of their methods in their concluding Limitations paragraph: the use of self-reported sleep data, heterogeneity from multiple cohorts, largely “cross-sectional” data and lack of longitudinal sleep and omics/MRI data; studying selected (healthy) samples; no consideration of other important sleep traits that vary with sleep duration. However, these significant limitations are largely not reflected elsewhere in the paper; rather the text often describes findings using strong causal language (influence, shape, impact, etc). The authors should consider more conservatively presenting their findings, providing a more balanced presentation.

The authors should consider whether labeling sleep durations as “optimal” may be misleading due to how data were derived (with specific numeric values likely to vary when using a different sleep question, sleep assessment modality, or different sample). A more cautious description may reduce the chance that the results are misinterpreted.

Given the importance of sleep duration measurement—and its variation in measurement across the cohorts examined- with varying influence on measurement error, it is important to be clearer on what was measured in each cohort. For example, the UKB field name was provided but the question used for sleep duration not provided. I think the paper reported data from a question that asked about sleep over a 24 hour period, which includes napping and the main sleep period. If so, the authors should be careful to interpret their data with this phenotype in mind, noting that napping and overnight sleep duration represent different concepts and health associations.

The authors acknowledge the limitations of the mediation analysis due to the sparse longitudinal data and single assessment of sleep and single assessment of the BAG mediator, and limitations of questionnaire and ICD outcome data. Providing further caution in describing those findings and acknowledging the difficulty in knowing the true temporal association between the exposures, mediators, and outcomes would help (while some acknowledgement of the latter is made, the results still are stated very strongly).

The use of ICD codes for insomnia and hypersomnia as proxies for short and long sleep duration, respectively, simplifies the important differences between these terms. Notably, insomnia is comprised of subtypes with normal and short sleep, and many individuals with short sleep do not meet criteria for insomnia (their short sleep reflects environmental or behavioral factors that curtail their time in bed). Clinical hypersomnia includes disorders of excessive sleepiness- although these sleep disorders often associate with long sleep duration, those sleeping > 8 hours per 24 hour period often will not meet criteria for disorders of hypersomnia. So while these diseases may represent more “extreme” phenotypes (as noted by the authors), they also are distinct from short vs long self-reported sleep duration.

Suggestions for improvement:
Please provide more details in the results section (including the footnotes in figures) to allow the reader to more easily appreciate what sample(s) were analyzed and what variables were adjusted for.
Describe the specific sleep duration questions used for analysis for each cohort; discuss more the implications of these definitions.

The datasets examined have information on sleep apnea, RLS, and insomnia. It would be very helpful to know if results varied (including sex-specific findings) with adjustment, stratification, or restriction based on these diseases.

Please provide a description of the UKBB subsamples corresponding to each data modality (MRI, proteomic, metabolomic) in the Methods 1: UK Biobank or Method 3: 23 Multi-organ aging clocks section for clarity.

It would be helpful to see a justification for the choice of variables used in each model, and minimally perform sensitivity analyses with a more comprehensive set of covariates.

The authors report sex differences in several BAGs, with some effects in opposite directions (i.e., accelerated aging in females based on brain MRI but the opposite pattern for the brain proteomic clock). It would be helpful if the authors could comment on whether these effects align with previous literature. Do these findings suggest that BAG estimation might benefit from sex-specific models? Could this also help explain differences in optimal sleep duration between males and females?

In the age-stratified analysis (Supplementary Figure 2), it appears that the optimal sleep duration differs between midlife and older adults, with older individuals showing shorter optimal sleep. The age-related pattern could be presented in parallel with the sex-related differences in optimal sleep duration. Has this pattern been formally tested- were age*sex interactions included? Please consider other reasons for sex differences, such as differences in reporting of sleep by males/females; differences in other sleep exposures that differ by sex (insomnia/RLS is more common in females and sleep apnea in males).

The BAG estimates across clocks were normalized for the analysis but providing information or a reference on the biological age prediction performance for each modality would be helpful. Among all clocks used, which one showed the best age prediction accuracy?

For the analysis of disease risk based on ICD codes, it would be important to clarify in the results section that this analysis was performed in a dataset independent of the main analytic sample (UKBB). In addition, including case–control sample

sizes (e.g., as additional labels in Figure 3a) would improve clarity. Please consider the concern that insomnia and hypersomnia are not proxies for short and long sleep as defined here.

Please provide the rationale for re-running the GWAS analyses for short and long sleep. Previous studies (PMID: 37770476 and PMID: 30846698), which included UKBB data, identified more genome-wide significant loci than reported here. Please clarify how this analysis differs from or extends those prior efforts and explain the differences in genetic signals reported.

Genetic correlations for sleep duration also have been reported in multiple prior reports- please compare to literature and justify the novelty here (PMID: 30846698, PMID: 37770476, PMID: 38388528, PMID: 34482950, PMID: 39856408, etc)

Decisions/limitations on data included: The use of specific sources of data were unclear. For example, regarding the discussion on objective vs. subjective sleep duration, both MESA and BLSA have objective sleep measures (and MESA has larger samples for sleep data than reported here.

Have the authors considered using an actigraphy-based sleep duration PRS (from PMID: 30846698) as a proxy for objective sleep duration? This could be estimated in the UKBB to test whether the observed U-shaped trajectory replicates.

On p. 11 lines 287-288, the hazard ratios for all-cause mortality associated with sleep duration do not match the results shown in Figure 3b (particularly for long sleep: 1.4 in the text vs. 1.1 in the figure). Please verify and correct this discrepancy.

To address the possibility that differences in the estimates for optimal sleep vary across different clocks could be attributable to sample differences, it would be helpful to conduct sensitivity analyses that restrict analyses to participants with overlapping MRI, proteomic, and metabolomic data. If the differences in optimal sleep duration persist in this subsample, it would suggest that the effect is not primarily driven by sample composition.

Minor comments:
Missing reference for MetGAB estimation (assuming this should be reference #14).

Please list which six brain disease GWAS summary statistics were obtained from PGC.

The abbreviations “DE” and “IDP” are defined too late in the text (DE first mentioned on p. 22 but explained on p. 27; IDP first mentioned on p. 18 but defined in the legend of Extended Data Figure 1). Please, defining them upon first use.
Typo in the “Ethnics” statement.

Version 2:

Reviewer comments:

Referee #4

(Remarks to the Author)
The authors have responded thoughtfully to my comments.

Referee #5

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Junhao (Hao) Wen, PhD
Department of Radiology

Author response letter, ID: NATURE-2025-05-11244A

We thank the editor for allowing us to revise our manuscript and the 3 reviewers for their thoughtful comments. In this round of revision, we carefully revised the manuscript in response to the comments.

Main updates for this revision include:

- Additional sensitivity analyses for the GAM non-linear modeling (including additional small-sample-sized actigraphy sleep duration data from BLSA).
- Additional sensitivity analyses to scrutinize the inverse causality (disease burden, or more specifically, major depressive disorder, to long/short sleep duration).
- Submitting our script for the GAM analysis and improving the SleepChart portal (https://labs-laboratory.com/sleepchart/) to improve reproducibility and transparency.

This response letter displays the reviewers' comments in black font, while our responses are in blue. In the main manuscript, the revised text is in yellow-colored text in this letter.

622 West 168th Street, New York, NY 10032

Referee #1 (Remarks to the Author):

In this manuscript, led by Prof. Junhao Wen, Cliodhna O'Toole and colleagues across the MULTI study, the authors conduct one of the largest and most comprehensive studies of the relationship between sleep disturbances and health outcomes to date. Combining clinical, genetic, proteomic, metabolomic and MRI data across several datasets including the UK Biobank, Finngen, Psychiatric Genomics Consortium, TriNetX and the Multi-Ethic Study of Atherosclerosis, the study conducts a series of genome-wide association studies, genetic correlation analyses, causal inference studies, proteome- and metabolome-wide associations, survival analyses and structural equation modeling to reveal several important insights. First, these analyses highlight how the well-described U-shaped relationship between sleep disturbances and health outcomes (whereby both shorter sleep, defined as less than six hours per night, and longer sleep, defined a greater than eight hours per night, are associated with adverse outcomes) extends beyond the central nervous system, encompassing strong associations with biological age gaps across pulmonary, hepatic, immune, skin and endocrine biological age, in addition to brain. Second, the study demonstrates that this U-shaped relationship is characterized by alterations across multiple biological pathways impacting inflammation, tissue repair dysfunction, neurogenesis, metabolic homeostasis, coagulation, and innate defense pathways - implicating more systemic modulation of the sleep-health relationship than previously understood. Third, the study illustrates that sleep disturbances are genetically correlated with over 150 illnesses across multiple bodily systems - with shorter sleep (<6 hours) showing the most prominent and widespread associations. Fourth, the study investigates the more nuanced, indirect relationship between long sleep (>8 hours) and a narrower collection of health outcomes - paying particularly close attention to late-life depression (LLD). The authors highlight that both short sleep and long sleep contribute to LLD, but while the former likely influences the illness more directly via immune dysregulation and HPA axis activation, the latter contributes more indirectly via accelerated brain and adipose tissue aging - potentially reflecting physiological compensatory mechanisms or subclinical processes related to aging.

We thank you for your precise summary of our study and your enthusiasm for our work.

Overall, this is an impressive manuscript. It represents one of the first large-scale, systematic investigations of the relationship between sleep disturbances, biological age and related health outcomes, constructing a fine-grained multi-omic atlas of the complex interplay between sleep and health. The authors should be commended for the rigor with which they have conducted the analyses, the precise but accessible writing style, the integration of multiple datasets and analytical approaches, and the creation of an intuitive online browser facilitating interactive perusal of the results. The contents of the manuscript would be of high relevance to the general readership of Nature. Indeed, more general audiences may find the study equally as interesting given the cultural relevance of the subject matter (see, for example, the relatively recent novel, 'My Year of Rest and Relaxation').

We appreciate your encouraging feedback and your innovative insight. I'm particularly intrigued by the novel you suggested, and we agree that the sleep–depression interaction has major implications for public health. In the revised manuscript, we also cited this novel in the results section for the mediation analysis for LLD (**Line: 507**): "The finding also parallels a cultural illustration from Moshfegh's *My Year of Rest and Relaxation*, where an unnamed narrator methodically increases prescription medication to achieve a year-long sleep."

Junhao (Hao) Wen, PhD
Department of Radiology

Although my general impression of the manuscript is positive, I outlined several questions and recommendations for the authors below.
Thank you! We carefully addressed your comments below!

First, one of the study's major limitations - already acknowledged by the authors - is its cross-sectional design, limiting the ability to determine causality. Causality is of paramount importance in any investigation of the relationship between sleep and complex phenotypes like major depressive disorder - prompting the question of whether depressed individuals have disturbed sleep due to their depression, or whether these individuals are depressed due to disturbed sleep. In the absence of longitudinal data from clinical intervention studies, it is challenging to answer that question effectively. Nonetheless, the authors attempt to address the question indirectly via a series of Mendelian randomization (MR) analyses, which ultimately revealed "no evidence for reverse causality". This may be a contentious assertion, considering that a previous (albeit less strongly powered) study found evidence of a bidirectional relationship between sleep and depression.
We totally agree with you! This is also the reason why we tend to be conservative regarding this directional effect between sleep and disease burden (in particular from disease burden to sleep). We believe that we have made this inverse causal relationship (MDD to sleep duration as an example, as we have specifically looked at late-life depression) **non-definitive** throughout our manuscript, even though our MR results report negative/null findings (no significant causal relationship from MDD to long/short sleep duration, passing the multiple comparisons based on the total number of diseases tested).

- For example, in the Abstract, we emphasize (**Line: 68**): "Although our Mendelian randomization does not show strong causal effects from disease to sleep disturbances, it does not fully rule out the possibility that sleep disturbances may, in part, reflect underlying disease burden.".
- In the Results section (**Line: 341**): "These analyses did not support a widespread causal effect of disease on sleep disturbances, reinforcing the interpretation of sleep disturbances as potential risk factors. We nevertheless acknowledge the possibility of bidirectional effects."
- Regarding the non-linearity of sleep duration, in the revised manuscript, we have now added sentences to highlight this (see more detailed response below for your following comment on the fractional polynomial MR method) (**Line: 343**): "For example, Pasman et al.[35] found bidirectional causality between major depressive disorder and insomnia, but not sleep duration. In addition, some previous studies modeled sleep duration as a continuous trait using linear Mendelian randomization, which does not capture potential non-linear relationships that methods such as fractional polynomial Mendelian randomization can address[36].

Here, I provide a more detailed review of the literature investigating the causal relationship between sleep duration and MDD, which are all candidate-based (pre-selecting certain psychiatric traits) MR analyses and not hypothesis-free. Overall, Mendelian randomization studies have reported a potential causal effect from sleep duration (exposure) → MDD (outcome) (many of them treat sleep duration as a continuous exposure variable, ignoring the potential non-linearity, e.g.,

https://www.sciencedirect.com/science/article/pii/S0165032724012801?via%3Dihub; https://pmc.ncbi.nlm.nih.gov/articles/PMC11215123/; https://academic.oup.com/pmj/article-abstract/101/1199/845/8002072;); however, the reverse direction, MDD as the exposure, has not, to our knowledge, been clearly demonstrated.

- A recent Nature Mental Health study (https://www.nature.com/articles/s44220-025-00471-x#Sec8) found: "For the circadian traits, we found support for a positive bidirectional effect of MDD on insomnia, such that genetic liability to MDD increased insomnia risk, and vice versa, while there was **limited** evidence for a similar effect on sleep duration or chronotype from either direction.")
- Another Translational Psychiatry paper (https://www.nature.com/articles/s41398-023-02622-z) also reported a similar conclusion using longitudinal data: "One standard deviation increase in PGS for short-sleep was associated with 14% higher odds of depression onset (95% CI = 1.03–1.25, p = 0.008). **However, PGS for sleep duration (OR = 0.92, 95% CI = 0.84–1.00, p = 0.053) and long-sleep (OR = 0.97, 95% CI = 0.89–1.06, p = 0.544) were not associated with depression onset during follow-up**."
- Another study (https://pubmed.ncbi.nlm.nih.gov/35465862/) showed that: "Genetically predicted sleep duration was also nominally associated with the risk of bipolar disorder (BD). Conversely, PTSD and MDD were associated with an increased risk of insomnia (OR = 1.06, 95% CI 1.03-1.10, p = 7.85 × 10-4 for PTSD; OR = 1.37, 95% CI 1.14-1.64; p = 0.001 for MDD). A **suggestive** inverse association of ADHD and MDD with sleep duration was also observed."
- An earlier NC paper (https://www.nature.com/articles/s41467-023-41249-y) also investigated the causal relationship between short/long sleep duration vs MDD, and concluded: "MR analysis investigating the causal influence of short sleep on MDD supported a positive causal association between short sleep and increased risk of MDD, using the inverse variance weighted method (β = 0.19 (0.02) p = 1.5x10−19, Fig. 6, Supplementary data 33). Conversely, MR on the impact of MDD on short sleep did not support a causal association (β = 0.01 (0.03), p = 0.69). MR investigating the causal influence of long sleep on MDD reveals a positive causal association between long sleep and increased risk of MDD, using the inverse variance weighted method (β = 0.14 (0.03), p = 1.64 × 10−5). This was a bidirectional effect, with an analysis of the impact of MDD on long sleep also revealing evidence of a causal association (β = 0.14 (0.04), p = 7.6 × 10−5).". However, they used a lenient P-value threshold for selecting the instrumental variables for MDD GWAS summary statistics: "The genetic instruments for all traits were defined as the independent variants that reached a significance threshold of $\boldsymbol{p < 1 \times 10^{-5}}$. Independent associations were identified by LD clumping with $r^2 = 0.6$ and a window of 250 kb. To avoid sample overlap, we used PGC data excluding UK Biobank samples." We have tried to reproduce the results by downloading the same PGC MDD GWAS, but after we used the genome-wide P-value threshold ($5x10^{-8}$) and performed clumping, no SNPs were left for valid IVs. We then downloaded the most recent FinnGen MDD GWAS (Version R12). The updated results are in the revised manuscript (**Line: 523**): "While a previous study by Austin-Zimmerman[67] reported a bidirectional relationship between long sleep duration and major depressive disorder, their analysis relied on a lenient P-value threshold (P-value<$1x10^{-5}$) for selecting instrumental variables, which

may cause weak instrument bias. We attempted to replicate their findings using the same PGC exposure GWAS by Wray et al.[68] (45,396 cases, removing UKBB samples) using a genome-wide P-value threshold (P-value<$5x10^{-8}$). After LD clumping, no independent genome-wide significant instruments remained. To further investigate this, we downloaded the most recent major depressive disorder (MDD) GWAS from FinnGen (59,333 cases). Overall, although the IVW (P = 0.003) estimator suggested a positive causal effect of genetic liability, the significant MR-Egger intercept (P=0.047), MR-PRESSO global pleiotropy (P=0.002), and high heterogeneity (e.g., P=0.0009) indicate pleiotropy-induced biases. Pleiotropy-aware methods showed attenuated effects and were sensitive to pleiotropy assumptions (balanced vs. directional pleiotropy), including MR-Egger for null/negative (P=0.24), MRMix for near-null (P=0.41), and MR-RAPS for nominally positive (P=0.019). Therefore, while suggestive, the evidence for a causal influence of MDD on long sleep duration remains inconclusive and should be interpreted cautiously (**Extended Data Figure 7**)." In addition, in this study, they performed a meta-analysis for sleep duration by using both the UK Biobank and MVP datasets, and have different bins to define short and long sleep duration, compared to our study.

In summary, prior "candidate-approach" studies (focusing on specific mental disorders) have reported mixed and sometimes contradictory findings regarding the direction and magnitude of this association. In our updated MDD MR analyses, we observed a potential signal with a causal effect of MDD on long sleep duration; **however, this did not remain significant after multiple testing correction and showed evidence of global pleiotropy**. Therefore, although we cannot entirely rule out the possibility of a directional causal influence of MDD on sleep duration, this relationship should be interpreted with caution. We sense that our claim in the Abstract is appropriate based on these additional results (**Line: 68**): "Although our Mendelian randomization does not show strong causal effects from disease to sleep disturbances, it does not fully rule out the possibility that sleep disturbances may, in part, reflect underlying disease burden.".

Also, we would appreciate it if you could direct us to the study you mentioned, if we did not include it here.

To ensure that the present MR analysis is robust and that the general readership more readily accepts the conclusion that no evidence of reverse causality could be detected, I would recommend that the authors address the following:

We thank you very much for your suggestions to further ensure the rigor of our analyses.

(1) Highlight how five distinct MR methods were employed in the main text; this is currently buried in the supplement. It adds weight to the claim that no evidence for reverse causality was detected,

Thank you! We agree with you. We have now added a method section for MR in **Method 10** (**Line: 835**). Our Mendelian randomization assessed the causality from DE → Sleep duration for in total of 525 disease endpoints (179 effective exposure variables after ensuring we had at least 7 IVs). Only one significant signal passed the Bonferroni correction (E4_OBESITYCAL from FinnGen for obesity). In addition, we made it explicit that we employed 5 different MR estimators and presented all results in **Supplementary File 8** (**Line: 338**): "To test this, we

conducted Mendelian randomization analyses using 5 different estimators (**Supplementary Note 4** and **File 8**) to test the reverse causal pathway, from 525 disease endpoints using FinnGen and Psychiatric Genomics Consortium (PGC) data, to the 2 binary sleep traits."

In the revised manuscript, we have specifically looked at the results for depression, and the two depression-related traits (F5_DEPRESSION_DYSTHYMIA and ANTIDEPRESSANTS as in **Supplementary File 8**) did not achieve even nominal significance (0.05), for example:

ciView: Data Plots

df

	outcome	exposure	method	nsnp	b	se	pval	lo_ci	up_ci	or	or_lci95	or_uci95
1065	long_vs_normal	F5_DEPRESSION_DYSTHYMIA	MR Egger	8	-0.84234	0.80255	0.33434	-2.41533	0.73065	0.43070	0.08934	2.07644
1066	long_vs_normal	F5_DEPRESSION_DYSTHYMIA	Weighted ...	8	0.13207	0.11172	0.23713	-0.08689	0.35104	1.14119	0.91677	1.42054
1067	long_vs_normal	F5_DEPRESSION_DYSTHYMIA	Inverse va...	8	0.14844	0.09483	0.11752	-0.03744	0.33431	1.16002	0.96326	1.39698
1068	long_vs_normal	F5_DEPRESSION_DYSTHYMIA	Simple mo...	8	0.02403	0.19827	0.90695	-0.36459	0.41264	1.02432	0.69448	1.51081
1069	long_vs_normal	F5_DEPRESSION_DYSTHYMIA	Weighted ...	8	0.00063	0.19539	0.99751	-0.38233	0.38359	1.00063	0.68227	1.46754

We added more depth sensitivity checks for this specific MR results (**Extended Data Figure 6** for scatter plot, leave-one-SNP-out analysis, funnel plot, and single-SNP analysis), on top of what you suggested below (e.g., F-statistics and MR-PRESSO).

(2) Report per-SNP and conditional F-statistics to assess instrument bias (considering that the GWASs of short and long sleep only produced a small number of genome-wide significant SNPs, it is possible that the instruments are too weak to detect any reverse causality);

We thank you for this valuable suggestion. Before providing these additional sensitivity analyses, we want to emphasize that the instrument variables were selected based on the exposure variable. In this context, it is the GWAS summary statistics from F5_DEPRESSION_DYSTHYMIA (not sleep duration, which showed only a few genome-wide SNPs as you mentioned) from FinnGen, and had 8 SNPs after clumping that pass genome-wide significance.

To address your comment, we evaluated instrument strength by reporting per-SNP F-statistics. Our eight LD-clumped SNPs for *F5_DEPRESSION_DYSTHYMIA* each exceeded the conventional weak-instrument threshold (per-SNP $F = (\beta/\mathrm{SE})^2$; min $F = 30.70$, median $F = 36.88$, mean $F = 38.62$; proportion $F < 10 = 0$). The SNPs jointly explained $R^2_{\text{total}} = 0.001126$(~0.11%) of exposure variance in $N = 274{,}330$, yielding an overall first-stage $F \approx 38.66$. Of note, the effective sample size of this GWAS is comparable to the most recent GWAS of MDD from PGC (around 40k cases). These values indicate acceptable instrument strength. Note that conditional (Sanderson–Windmeijer) F-statistics are for multivariable MR and are not applicable to our univariable analyses.

alternatively or additionally, please report whether a formal MR power analysis was conducted,

In the initial analysis, we did not perform a dedicated MR power calculation in the original submission. Instead, we pre-screened the 525 exposures to ensure adequate instrument strength, requiring at least 7 independent genome-wide significant SNPs. As you suggested, in this revision, we provided evidence of strong instruments (per-SNP F-statistics generally >10; mean F ≈ 38.6 in MDD → long-sleep analysis). This complements our heterogeneity and pleiotropy tests and addresses the concern about weak instruments.

(3) Consider integrating MR-PRESSO (Mendelian Randomization Pleiotropy RESidual Sum and Outlier) as a sixth method to assess horizontal pleiotropy and robustness of findings (not an essential integration, but if feasible, strongly recommended),

Junhao (Hao) Wen, PhD
Department of Radiology

We thank you for this great suggestion! We performed additional analysis using the MR-PRESSO estimator and confirmed the absence of causality from depression → long-sleep analysis. Results are updated in **Extended Data Figure 6f**.

(4) Alternatively and/or additionally to the above, please conduct a non-linear Mendelian randomization approach like fractional polynomal MR (e.g., https://rdrr.io/github/jrs95/nlmr/man/fracpoly_mr.html). A U-shaped causal effect may be attenuated using linear MR methods. This is a more essential integration than the MR-PRESSO mentioned above.

This is an excellent suggestion! I think this non-linear MR method should already have been applied to previous studies where they tested the causal direction: **continuous** sleep duration (exposure) → MDD (outcome) (many of them treat sleep duration as a continuous exposure variable, ignoring the potential non-linearity, e.g., https://www.sciencedirect.com/science/article/pii/S0165032724012801?via%3Dihub; https://pmc.ncbi.nlm.nih.gov/articles/PMC11215123/; https://academic.oup.com/pmj/article-abstract/101/1199/845/8002072;).

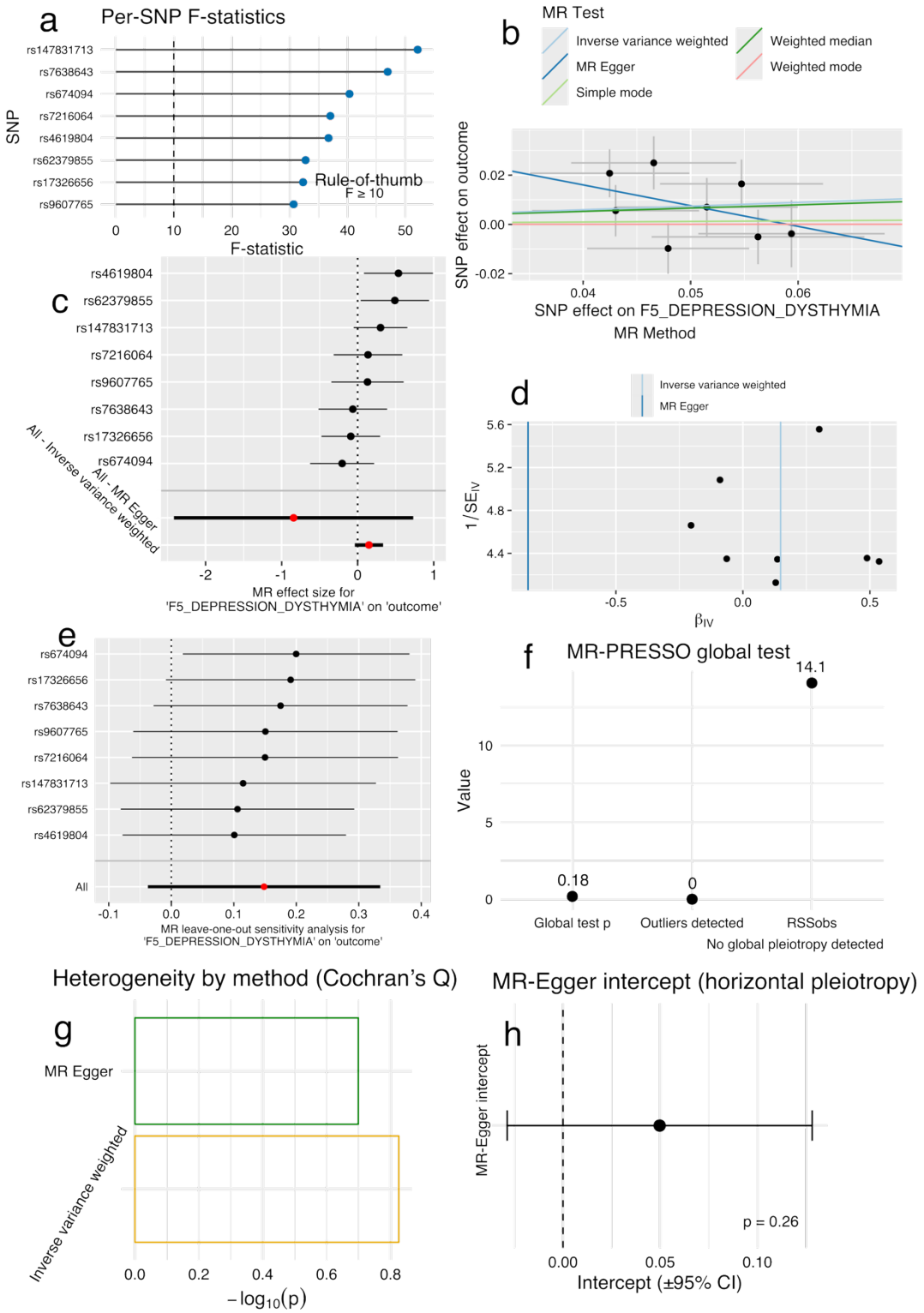
However, this approach is designed to evaluate non-linear **dose–response** relationships for **continuous exposures** by stratifying on an "IV-free" exposure. In our current analysis, the exposure is binary (MDD case–control), and the outcomes are binary categories of sleep duration (short vs. normal; long vs. normal). Because there is no continuous exposure to stratify, fractional-polynomial MR may not be directly applicable to our MDD→sleep setting.

In the revision, we clarify this point and note that prior MR studies treating sleep duration as a continuous trait should have considered non-linear MR (**Line: 344**): "In addition, some previous studies modeled sleep duration as a continuous trait using linear Mendelian randomization, which does not capture potential non-linear relationships that methods such as fractional polynomial Mendelian randomization can address[36]."

In the revised manuscript, we provided all additional sensitivity check analyses in **Extended Data Figure 6**. Additionally, in our own experiments or in the literature, MDD is not causally linked to a short sleep duration pattern.

622 West 168th Street, New York, NY 10032

Extended Data Figure 6: Sensitivity analysis for the causal relationship from FinnGen F5_DEPRESSION_DYSTHYMIA to long sleep duration

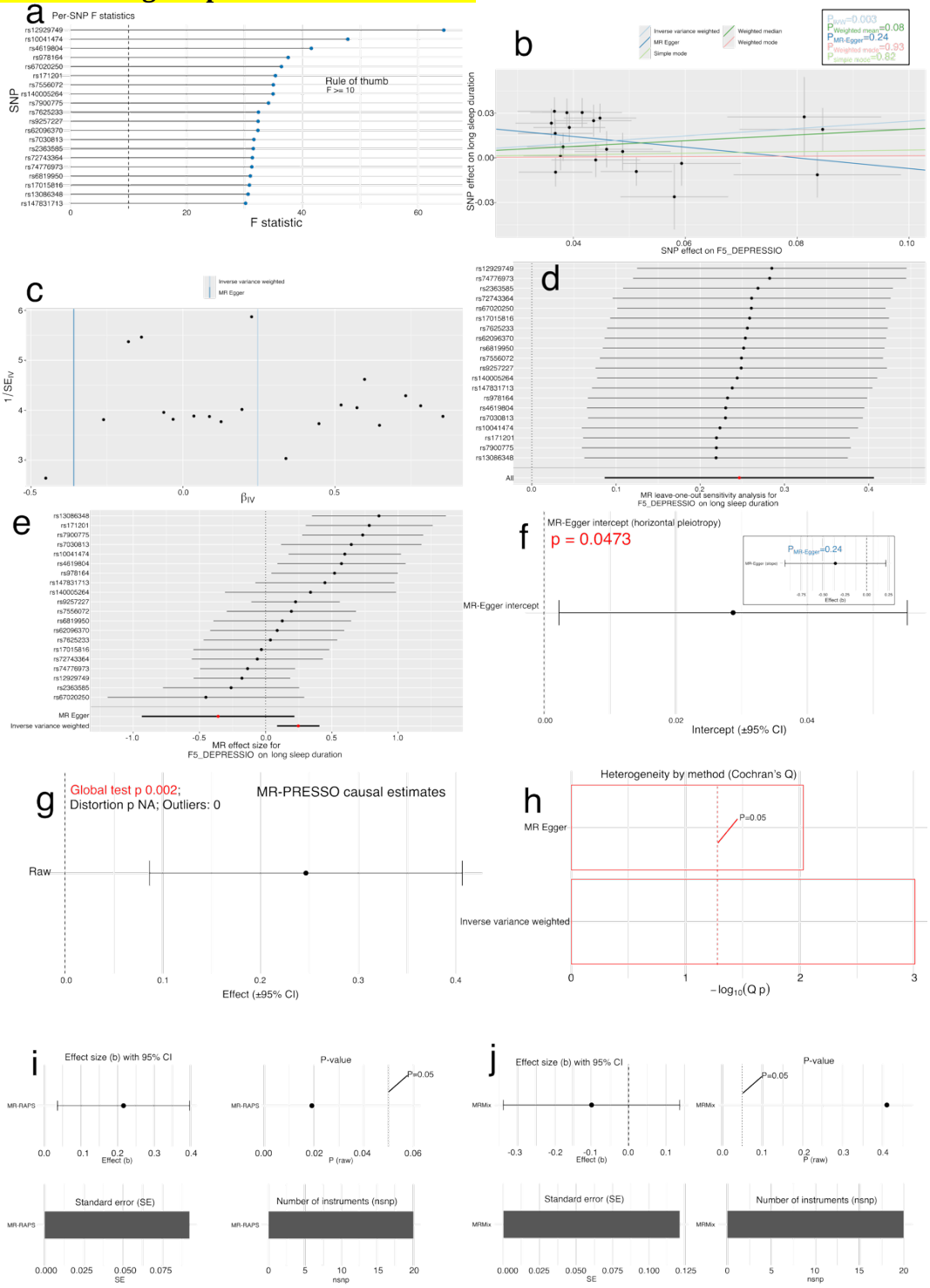


a) Lollipop plot of first-stage F-statistics for the 8 genome-wide significant depression instruments after clumping, ordered from weakest to strongest. In this analysis, all instruments exceed the suggestive threshold for F=10 (min F = 30.70; median F = 36.88; mean F = 38.62).

The combined variance explained is R^2-total = 0.00113, giving an approximate overall F ≈ 38.66 (N ≈ 274,330; k = 8), indicating adequate instrument strength for univariable Mendelian randomization. **b**) Scatter plot for the Mendelian randomization effect sizes of the SNP-depression association (x-axis, log OR) and the SNP-long sleep duration associations (y-axis, log OR) with standard error bars. The slopes of the 5 lines correspond to the causal effect sizes estimated by the five MR estimators, respectively. For all five methods, no statistical significance was achieved (P-value>0.05). **c**) Forest plot for the single-SNP Mendelian randomization results. Each dot represents the effect (log OR), and the error bar displays the 95% CI using only one SNP; the red line shows the MR effect using all SNPs together for IVW and Egger estimators. **d**) Funnel plot for the relationship between the causal effect of depression on long sleep duration. Each dot represents causal effect sizes estimated using each SNP as a separate instrument against the inverse of the standard error of the causal estimate. **e**) Leave-one-SNP-out analysis of depression on long sleep duration. Each dot represents the causal effect (log OR), and the error bar displays the 95% CI by excluding that SNP from the analysis. The red line depicts the IVW estimator using all SNPs. **f**) MR-PRESSO confirms the absence of causality and a significant causal signal (P=0.16). Dots summarize the three key MR-PRESSO outputs: the global test p-value, the number of outlier instruments detected, and the observed residual sum of squares (RSSobs). Using eight SNPs, the global test was not significant (P-value=0.18), no outliers were identified (0), and RSSobs was 14.1, indicating no evidence of horizontal pleiotropy. MR-PRESSO was run with 1,000 simulations at α=0.05; because no outliers were detected, no outlier-corrected estimate was produced. **g**) Heterogeneity (Cochran's Q) test indicated no strong between-instrument heterogeneity: MR-Egger Q=8.56 (df=6, p=0.200) and IVW Q=10.76 (df=7, P-value=0.149) using 8 SNPs. These P-values >0.05 suggest the causal estimates are not driven by substantial heterogeneity. **h**) The MR-Egger intercept was 0.0497 (SE=0.0400; P=0.260), indicating no evidence that directional (unbalanced) pleiotropy biases the causal estimate in this analysis.

In addition, we also downloaded the most recent MDD exposure GWAS summary statistics from the most recent version of FinnGen and added additional discussion. The updated results are in the revised manuscript (**Line: 524**): "While a previous study by Austin-Zimmerman[67] reported a bidirectional relationship between long sleep duration and major depressive disorder, their analysis relied on a lenient P-value threshold (P-value<$1x10^{-5}$) for selecting instrumental variables, which may cause weak instrument bias. We attempted to replicate their findings using the same PGC exposure GWAS by Wray et al.[68] (45,396 cases, removing UKBB samples) using a genome-wide P-value threshold (P-value<$5x10^{-8}$). After LD clumping, no independent genome-wide significant instruments remained. To further investigate this, we downloaded the most recent major depressive disorder (MDD) GWAS from FinnGen (59,333 cases). Overall, although the IVW (P = 0.003) estimator suggested a positive causal effect of genetic liability, the significant MR-Egger intercept (P=0.047), MR-PRESSO global pleiotropy (P=0.002), and high heterogeneity (e.g., P=0.0009) indicate pleiotropy-induced biases. Pleiotropy-aware methods showed attenuated effects and were sensitive to pleiotropy assumptions (balanced vs. directional pleiotropy), including MR-Egger for null/negative (P=0.24), MRMix for near-null (P=0.41), and MR-RAPS for nominally positive (P=0.019). Therefore, while suggestive, the evidence for a causal influence of MDD on long sleep duration remains inconclusive and should be interpreted cautiously (**Extended Data Figure 7**)."

Extended Data Figure 7: Sensitivity analysis for the causal relationship from the most recent MDD to long sleep duration in FinnGen



Mendelian randomization of major depressive disorder (MDD) on long sleep duration using the FinnGen (R12) GWAS. We downloaded the R12 FinnGen MDD GWAS with an expanded patient sample (*N*=59,333) to ensure adequate statistical power. **a)** Lollipop plot of first-stage F-statistics for the eight genome-wide significant MDD instruments after LD

clumping, ordered from weakest to strongest. All instruments exceeded the conventional F > 10 threshold, suggesting sufficient instrument strength. **b)** Scatter plot of SNP-specific associations between MDD (x-axis, log OR) and long sleep duration (y-axis, log OR). The slopes of the five regression lines correspond to causal effect estimates from the inverse-variance weighted (IVW), weighted median, MR-Egger, simple-mode, and weighted-mode estimators. Among these, only the IVW method reached nominal significance (P=0.003). The MR-Egger slope was negative, contrasting with the positive slopes from all other estimators. **c)** Funnel plot showing the relationship between individual SNP causal effects and their precision, revealing heterogeneity across variants. **d)** Leave-one-SNP-out analysis of the MDD → long-sleep association; each point and 95 % CI represent the IVW estimate obtained when excluding a single SNP. **e)** Forest plot of single-SNP causal effects with corresponding 95 % confidence intervals. **f)** MR-Egger intercept plot showing a significant positive intercept (0.03 ± 0.01; P = 0.047), indicating directional pleiotropy that may bias the overall causal estimate. **g)** MR-PRESSO global test also detected horizontal pleiotropy (P = 0.002) but no outlier variants; consequently, the distortion test was not applicable. **h)** Cochran's Q heterogeneity statistics for both IVW (P=0.0009) and MR-Egger (P=0.009) methods further confirmed substantial heterogeneity among the instruments. **i**) MR-RAPS models balanced (mean-zero) horizontal pleiotropy via an over-dispersion term and down-weights outliers using a robust loss, while profiling out weak-instrument bias. MR-RAPS point estimate remains positive but greatly attenuated versus IVW (P=0.019), indicating a signal that persists under balanced-pleiotropy/weak-IV adjustments, though with increased uncertainty. Interpretation should note that MR-RAPS does not correct for directional pleiotropy; results are therefore contingent on the balanced-pleiotropy assumption. **j**) MRMix estimates the causal effect by modeling SNP residuals as a two-component mixture, a spike for valid IVs (no direct effect) and a normal component with variance σ^2 for invalid/pleiotropic IVs. By probabilistically allocating pleiotropic SNPs to the high-variance component, MRMix limits their influence on the slope and is robust to a large fraction of invalid instruments when pleiotropic effects are mean-zero on average. Here, the estimate is near-null/slightly negative (P=0.41), with a CI spanning zero, providing no pleiotropy-robust evidence for a causal effect under the mixture model. To summarize, although IVW suggested a positive causal effect of genetic liability to MDD on long sleep (P = 0.003), multiple heterogeneity/pleiotropy diagnostics argue this is likely biased: the MR-Egger intercept is significant, the MR-PRESSO global test is positive, and Cochran's Q indicates substantial heterogeneity. Pleiotropy-robust estimators do not corroborate the IVW signal: MR-Egger (slope) is null/negative (β=−0.36, P = 0.24), MRMix is near-null (β = −0.10, P = 0.41), and MR-RAPS, while largely attenuated and nominally positive (β=0.22, P = 0.019), remains sensitive to pleiotropy assumptions. Taken together, the data do not provide pleiotropy-robust evidence for a causal effect; any apparent association should be viewed as suggestive at best and interpreted with caution.

Second, while the authors adjust for sex in all analyses and highlight several clear sex differences in the relationship between sleep and biological aging, I noticed that the U-shaped curve is at least visually apparent for males but not females across several biological age gaps including heart protein BAG, endocrine protein BAG and potentially heart MRI BAG. As an alternative to adjusting for sex status and interaction terms in the models, have the authors considered

conducting sex-stratified analyses for these BAGs to determine if the U-shaped relationship is significant for males but not females?

This is an excellent observation and a great suggestion for sex-stratified analyses. Before the additional experiment, we want to acknowledge the sex-specific patterns in our results. Of note, in **Fig. 1**, the BAGs with * symbol and color-icons denote significant results after Bonferroni correction (0.05/23). Heart MRIBAG and heart ProtBAG were not significant (potentially due to the P-value threshold cutoff sensitivity).

In this revision, we added the results for the sex-stratified results (**Supplementary Figure 3**).

Junhao (Hao) Wen, PhD
Department of Radiology

TEXT REDACTED

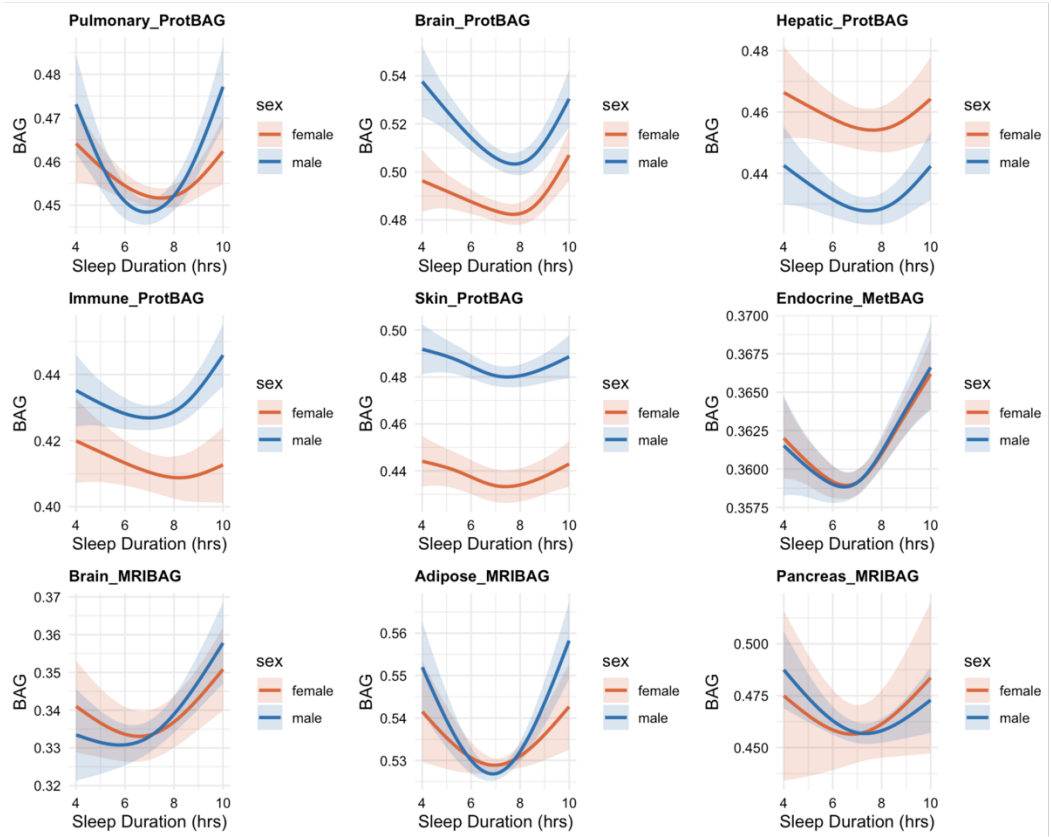
FIGURE REDACTED

We split the cohort into females and males and re-fit the same GAM specification selected in the main analysis (fixed optimal k and family), testing the sleep term within each sex. Significance was assessed with Bonferroni correction ($\alpha = 0.05/23$); significant panels are marked with icons and asterisks. Overall, U-shaped associations were broadly reproducible across sexes, with some BAGs showing stronger effects in males (e.g., Endocrine_ProtBAG) and others in females (e.g., Pancreas_MRIBAG, Kidney_MRIBAG). These sex differences may reflect biology, such as hormone milieu (estrogen/androgen signaling), sex-specific sleep architecture and circadian regulation, organ-specific metabolic and inflammatory pathways, and life-stage effects (e.g., menopause), in addition to possible sample-size or residual confounding differences between strata.

Additional, given the wealth of hormonal blood biochemistry readily available in the UK Biobank, including oestradiol, SHBG, and testosterone levels, have the authors considered whether the sleep - health outcome slope is modulated by these variables?
Thank you for this great suggestion! We agree that sex hormones could plausibly modulate the sleep–health associations. In the UK Biobank, testosterone (Field 30850), SHBG (30830), and estradiol (30800) are available in our data bucket. In the revision, we add each hormone and fit the interaction term by testing whether slopes for sleep duration differ by hormone level (i.e., hormone * sleep duration).
Results are updated in **Supplementary Figure 4-6**:

622 West 168th Street, New York, NY 10032

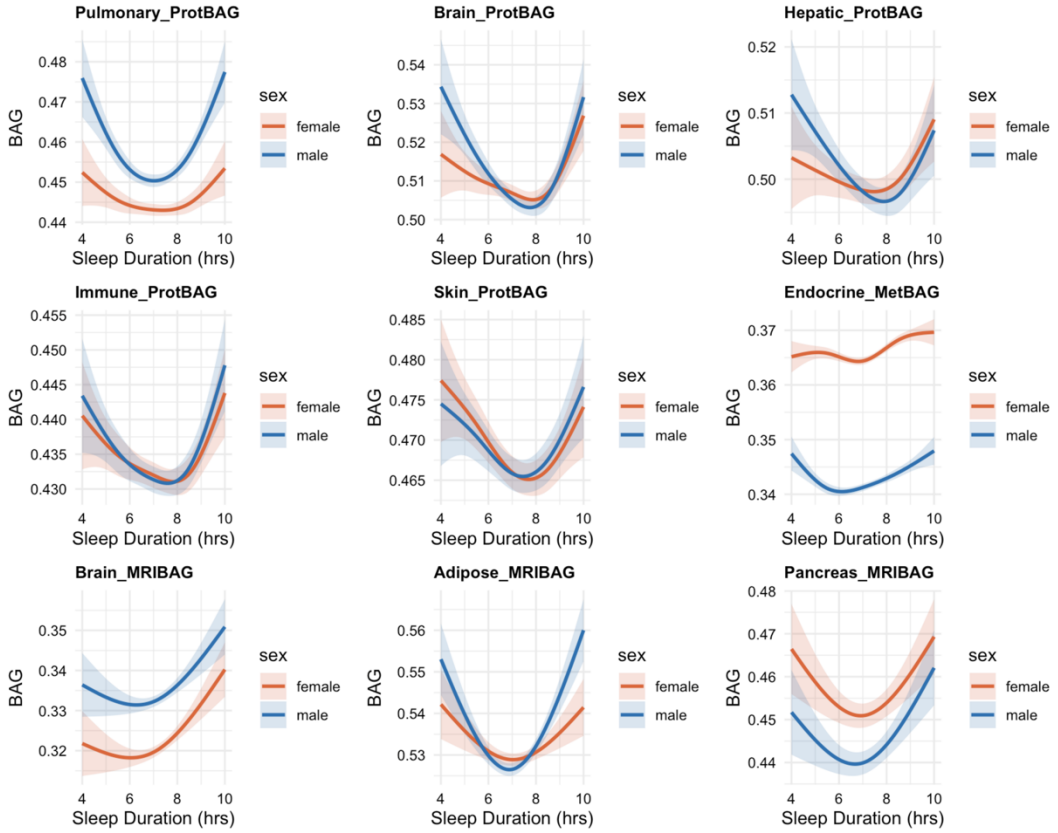
Supplementary Figure 4: Sensitivity check analysis by testing the interaction of sleep duration and testosterone



We refit the GAMs to include testosterone as a covariate and an interaction term with sleep duration (sleep duration × testosterone). None of the interaction effects reached significance (all $P > 0.05$), and while the 9 U-shaped associations remained, their curvature or inflection points shifted slightly in a few BAGs (e.g., endocrine MetBAG).

Junhao (Hao) Wen, PhD
Department of Radiology

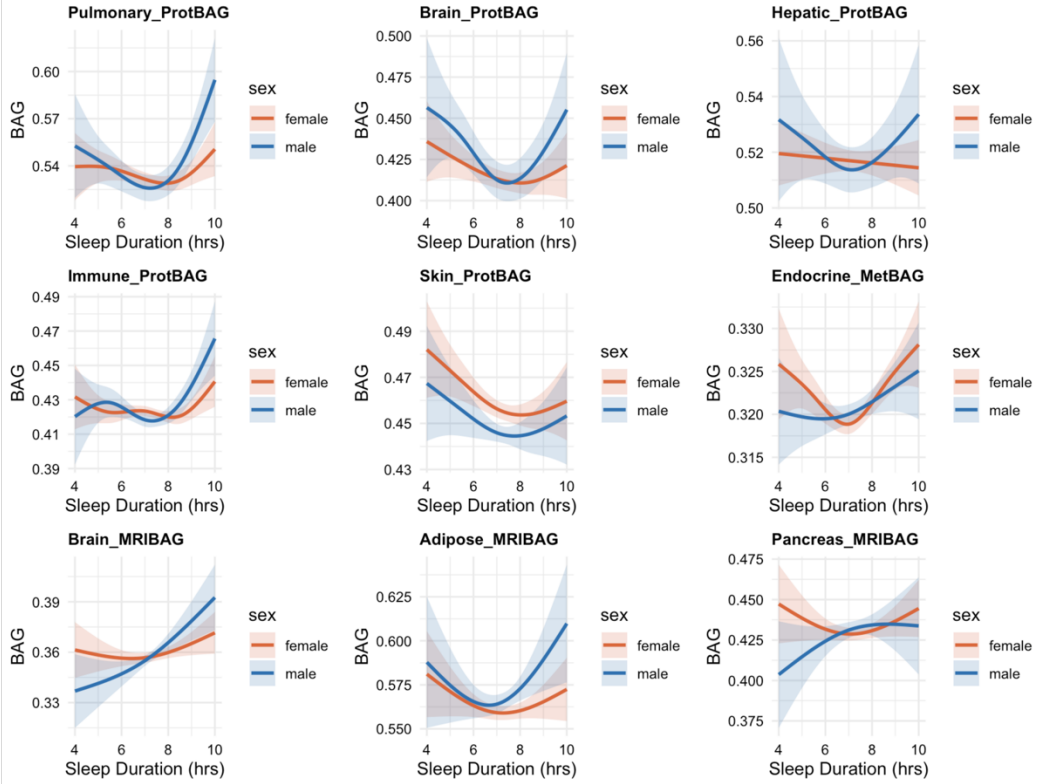
Supplementary Figure 5: Sensitivity check analysis by testing the interaction of sleep duration and SHBG



We refit the GAMs to include SHBG as a covariate and an interaction term with sleep duration (sleep duration × SHBG). None of the interaction effects reached significance (all $P > 0.05$), and while the 9 U-shaped associations remained, their curvature or inflection points shifted slightly in a few BAGs (e.g., endocrine MetBAG).

622 West 168th Street, New York, NY 10032

Supplementary Figure 6: Sensitivity check analysis by testing the interaction of sleep duration and oestradiol



We refit the GAMs to include estradiol as a covariate and an interaction term with sleep duration (sleep duration × estradiol). Adding estradiol altered the fitted curves and, for the hepatic ProtBAG, the sleep × estradiol interaction was nominally significant (P=0.0196). That means the association between sleep duration and hepatic BAG appears to depend on estradiol level, e.g., the U-shape's curvature/inflection may differ at low vs. high estradiol.

Beyond the two issues outlined above, other pieces of minor feedback are listed below:

• Line 124: Define "GAMs" as 'generalized additive models' before shortening to its acronym
Thank you! We have defined GAM at its first occurrence at **Line: 109**.

• Line 161: For clarity / easier interpretation, I would recommend updating the first sentence of this paragraph to, "Overall, across the nine significant protein, metabolomic and MRI-based BAGs, optimal sleep duration ranged from 6.5 to 7.8 hours for females and 6.4 to 7.7 hours for male."
Thank you for the suggestion to improve the readability of our manuscript. We have now revised the sentence (**Line: 164**): "Overall, across the nine significant proteomic, metabolomic, and MRI-based BAGs, optimal sleep duration ranged from 6.5 to 7.8 hours for females and 6.4 to 7.7 hours for males."

• Lines 219-233: This paragraph is a little messy; it may be more intuitive if the disease codes are relegated to the supplementary materials and replaced in the main text with genetic correlation coefficient values.

Junhao (Hao) Wen, PhD
Department of Radiology

Thank you for this great suggestion! It has been updated in the revised manuscript (**Line: 221**):
"Genetic correlation analyses revealed 153 positive associations between abnormal sleep duration patterns and systemic diseases across multiple organ systems (P<0.05/527), dominated by short sleep duration (**Fig. 2c**). For short sleep duration, we observed robust correlations with cardiovascular diseases, including ischemic heart disease (g_c=0.19±0.04; P=2.52x10^{-6}), heart failure (g_c=0.31±0.06; P=5.80x10^{-6}), and coronary atherosclerosis (g_c=0.18±0.04; P=1.51x10^{-5}), metabolic disorders, such as type 2 diabetes (g_c=0.18±0.04; P=5.33x10^{-5}) and insulin-requiring diabetes (g_c=0.17±0.04; P=4.45x10^{-5}), and musculoskeletal conditions, like low back pain (g_c=0.40±0.06; P=1.13x10^{-12}), osteoarthritis (g_c=0.19±0.05; P=7.07x10^{-5}) and soft tissue disorders (g_c=0.50±0.06; P=2.04x10^{-16}). Neurological and psychiatric associations included migraine (g_c=0.25±0.06; P=1.91x10^{-5}), depression (g_c=0.37±0.04; P=1.57x10^{-18}), anxiety (g_c=0.32±0.06; P=1.83x10^{-7}), substance use disorders (g_c=0.37±0.05; P=3.73x10^{-15}), and suicidality (g_c=0.34±0.06; P=4.28x10^{-8}), indicating broad involvement of the central nervous system. Pulmonary and infectious conditions, such as asthma (g_c=0.22±0.05; P=2.50x10^{-5}), bronchitis (g_c=0.42±0.12; P=5.00x10^{-4}), and chronic obstructive pulmonary disease (g_c=0.28±0.05; P=2.74x10^{-7}), were also genetically correlated, as were gastrointestinal and hepatic diseases, including reflux (g_c=0.34±0.06; P=6.09x10^{-8}), diverticulosis (g_c=0.23±0.04; P=1.14x10^{-6}), and irritable bowel syndrome (g_c=0.32±0.09; P=3.00x10^{-4}). In contrast, long sleep duration showed a more focused genetic correlation profile, predominantly involving brain-related phenotypes, such as major depressive disorder (g_c=0.29±0.04; P=2.57x10^{-11}), schizophrenia (g_c=0.28±0.03; P=3.47x10^{-16}), bipolar disorder (g_c=0.21±0.03; P=1.09x10^{-7}), alcohol dependence (g_c=0.23±0.05; P=3.26x10^{-5}), ADHD (g_c=0.28±0.04; P=2.24x10^{-12}), and migraine (g_c=0.28±0.07; P=7.09x10^{-5}), suggesting potential compensatory or indirect neuropsychiatric mechanisms (further elucidated in **Fig. 4**). **Supplementary File 4** present detailed statistics for our genetic correlation analysis."

• Line 246: Figure 2A is not particularly informative; if the authors decide to retain this part of the figure, I would recommend appending the genetic locus with the closest genes to the significant SNPs. Figure 2B shows no significant associations for long sleep - even at a nominal level.

Thank you! I agree that the cytogenetic region is not specific! In the revised manuscript for **Fig. 2A**, we added the top lead SNP, its nearest gene, on top of the cytogenetic region. We have also updated the results to correct a typo when we plotted **Fig. 2a** in the initial manuscript. This correction only changes the visualization in **Fig. 2A** but does not affect any of the analyses or conclusions for downstream post-GWAS (i.e., genetic correlation). The updated results are now reflected in **Supplementary Table 4** and the original GWAS summary statistics (Data Availability: "All GWAS summary statistics are also available at Synapse (https://www.synapse.org/Synapse:syn64923248/wiki/630992[98] and https://www.synapse.org/Synapse:syn68602065/wiki/633329[99])."). Regarding **Fig. 2B**, to avoid confusion, we clarify that this panel presents MAGMA gene-property (tissue enrichment) results derived from the GWAS summary statistics – no single-variant GWAS hits. Consistent with the main text, short sleep (vs. normal) shows broader tissue enrichment, whereas long sleep (vs. normal) does not yield tissue enrichments passing FDR.

• Line 276: Please clarify if "disturbed sleep duration" refers to long, short, or both.

622 West 168th Street, New York, NY 10032

Thank you! We have revised this accordingly (**Line: 280**): "Finally, a cluster of digestive disorders, including gastritis and duodenitis (K297), gastroesophageal reflux disease (K219), and functional intestinal disorders (K590), was significantly predicted by both long and short sleep duration."

• Line 290: In Figure 3 (as well as Figure 2C), I was expecting the brain ICD-10 codes to be closer to the visualized brain area on the body map. If possible, I would rearrange the ICD-10 codes on both images so that they more closely correspond with their respective depictions on the anatomical map on the right.
Thank you for this excellent suggestion to enhance the clarity and readability of our manuscript. In the revised version of **Figure 3** and **Figure. 2C**, we have implemented your recommendation by arranging the diseases according to their corresponding organ systems, placing brain-related diseases at the top, followed by those related to the heart, lungs, and other organs, consistent with the human anatomy illustration on the right. Additionally, we have revised **Extended Data Figure 5** accordingly, following your suggestion.

• Line 290: Figure 3 part B - I would recommend switching "short" and "long" in the axis labels to ensure they are aligned with the labels in the inner diagram. At the moment, "short (<6 hrs) aligns with "sleep_long" visually, which is counterintuitive for the reader.
Thank you for this suggestion! This has been corrected accordingly in the revised manuscript.

• Lines 310-311: Be more specific about the LLD1 and LLD2 outcomes; they are currently described as "AI- and MRI-derived outcomes of late-life depression", which I first interpreted as two different measures, one being AI-derived and the other being MRI-derived. I would recommend changing this to either "MRI-derived" or "MRI-based, AI-derived".
Thank you! We have revised this accordingly (**Line: 315**): "To evaluate this hypothesis, we performed structural equation modeling (SEM) using sleep duration patterns (short or long) as exposures, the 7 MRIBAGs as mediators, and two MRI-derived subtypes of late-life depression (LLD1 and LLD2[8]) as outcomes."

• Lines 315-316: Outline what c2 and a1 refer to in the context of structural equation modeling; this will be helpful for a more general audience.
Thank you! In the revised manuscript, we have made this specifically in **Method 9** (**Line: 319**): "In the pathway linking short sleep to LLD1, characterized by preserved subcortical brain volume, we observed strong direct effects (c2 ranging from -0.33 to -0.24) across 6 organ systems; refer to **Method 9** for the definition of c1, c2, and a1 coefficients."

• Lines 408: It is worth noting that the UK Biobank's MRI study has a well known bias towards healthier, well educated people. Therefore, while the authors' assertion that reverse causality may 'pull' the right side of the U-shaped curve upward, it may also simply be the case that protein BAGs and metabolomic BAGs are derived from a more representative collection of participants encompassing a wider range of health states versus those who participated in the MRI sub-study.
Thank you for this insight! We have added this in the Discussion section (**Line: 414**): "In addition, UKBB's MRI sub-study over-represents healthier and more highly educated participants; this may attenuate or distort the right limb of the observed U-shape independent of reverse causality. In contrast, the proteomic and metabolomic BAGs are derived from broader,

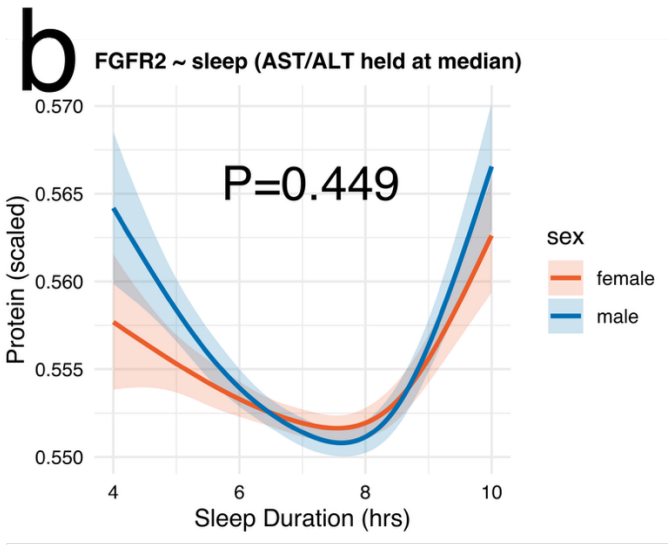
potentially more representative samples spanning wider health states, which could partly explain their stronger or more symmetric associations."

• Line 665: It is worth noting that significant redundancies exist in the proteomic data; many proteins show high correlations with one another. As such, Bonferroni adjustment may be overly stringent compared with the FDR adjustment or other approaches. This is worth at least noting in the text, if the authors decide to retain the more stringent analyses in the main study.
We fully agree with this point. Proteomic features are highly correlated, so Bonferroni can be conservative relative to FDR or "effective number of tests" approaches. Multiple testing corrections were applied using Bonferroni adjustment (P<0.05/2923) for ProWAS to avoid false positive signals, considering the large number of proteins tested. We have added this discussion in the revised manuscript (**Line: 712**): "Given the substantial correlation structure among proteomic measures, we acknowledge this choice is conservative; we retain Bonferroni correction in the main text to minimize false positives."

Line 665: Another point on the same analysis is that kidney functon (via markers like ALT/AST) impacts concentrations of many of these proteins, but was not adjusted for in the model. Although it might be extreme for the authors to completely re-conduct these analyses adjusting for kidney function, it might be worth checking whether the most significant findings remain significant after adjusting for these covariates.
We thank you for this additional insight. In the revised manuscript, we have performed additional sensitivity analysis by including ALT/AST as an additional covariate and tested its interaction with the example protein tested (FGFR2) (**Supplementary Figure 10**).

Supplementary Figure 10: Sensitivity check of the ProWAS for protein FGFR2 of sleep duration mediated by the AST/ALT ratio

[PANEL REDACTED]



a) **[TEXT REDACTED]** **b**) Results after including AST/ALT as an additional covariate and its interaction with sleep duration. The interaction term was not statistically significant (P-value>0.05), although the U-shaped association pattern showed a slight change.

As shown in the figure above, we explicitly tested whether the AST/ALT * FGFR2 interaction term and found no significant signal (P=0.449), although this slightly changed the U-shaped patterns between females and males.

Thank you for the time and care you put into reviewing our work – your thoughtful suggestions have substantially improved the manuscript's rigor and readability.

Junhao (Hao) Wen, PhD
Department of Radiology

Referee #2 (Remarks to the Author):

Overall assessment
This manuscript makes a timely and impactful contribution, with strong relevance for both the scientific community and the public. The authors introduce a "Sleep Chart" that systematically maps nonlinear (U-shaped) associations between sleep duration and biological aging clocks across multiple organs and omics layers, with a strong effect for shorter than normal sleep. Leveraging the UK Biobank and several independent cohorts, the study provides compelling evidence that both short and long sleep duration accelerate biological aging and increase disease risk. The systemic approach, spanning imaging, proteomics, metabolomics, and genetics, represents an advance over prior brain-focused work, or blood proteomics-focused work.
Originality and significance
Thank you for your enthusiasm and precise summary of our manuscript! We have carefully addressed your individual comments listed below.

• Novelty: This work extends sleep–aging research beyond the brain to a multi-organ, multi-omics framework, incorporating imaging, proteomics, and metabolomics.
Thank you!

• Significance: It leverages biological aging clocks as systemic markers, bridging molecular biology, imaging, and clinical outcomes.
Thank you!

• Integration: The study unifies genetic correlation, survival analysis, and mediation modeling, highlighting both direct and indirect pathways linking sleep to disease.
Thank you!

• Public health implications: Sleep is modifiable, and the identification of organ- and sex-specific optimal durations makes this clinically and societally relevant.
Thank you!

Data and methodology
• Strengths:
o Rigorous and comprehensive methodology with convergent evidence across modalities.
Thank you!

o Replication in independent datasets enhances confidence.
Thank you! Although we have indeed included independent studies, another study at the same scale as the UK Biobank is required to more confidently replicate these results in future studies (**Line: 547**): "First, although our findings revealed U-shaped associations between sleep duration and various phenotypes, these patterns require external validation in independent cohorts, potentially on the same scale as UKBB."

o An interactive portal for data exploration improves transparency and promotes engagement.
Thank you! We have improved the portal during the revision.

622 West 168th Street, New York, NY 10032

• Points for consideration:
o Optimal durations differ across modalities (e.g., ~7.7h proteomics vs. ~6.5h MRI); please discuss how this discrepancy relates to biological interpretation.
Thank you! This is an excellent suggestion. In the Discussion section, we have now added sentences to discuss this in more detail (Reviewer #1 also commented on this and gave a great suggestion) (**Line: 406**): "For instance, brain ProtBAG showed the lowest aging burden around 7.7 hours of sleep, whereas brain MRIBAG reached its inflection point closer to 6.5 hours. These findings suggest that molecular aging in the brain may require longer restorative sleep than its structural counterpart, potentially reflecting the different timescales and mechanisms by which proteomic versus imaging phenotypes capture sleep-related damage or resilience[46,48]. Alternatively, this can be due to a reverse causality, where longer sleep reflects brain aging rather than causes it, pulling the right side of the U-shaped curve upward, and making the apparent optimal sleep duration appear shorter for the brain MRIBAG (e.g., 6.5 hours). In addition, UKBB's MRI sub-study over-represents healthier and more highly educated participants; this may attenuate or distort the right limb of the observed U-shape independent of reverse causality. In contrast, the proteomic and metabolomic BAGs are derived from broader, potentially more representative samples spanning wider health states, which could partly explain their stronger or more symmetric associations. To summarize, the differing optimal sleep durations across organ-specific clocks likely reflect the heterogeneous physiological demands and recovery processes of each organ system. For example, organs such as the brain may be more sensitive to sleep deprivation and circadian disruption, whereas metabolic or peripheral organs (e.g., liver, pancreas) may exhibit delayed or compensatory responses, resulting in distinct sleep–aging optima."

o Proteomic and metabolomic signatures are dynamic; consider discussing temporal variability and the value of longitudinal data.
Thank you for this great suggestion. In the revised manuscript, we have added this in the Limitation section (**Line: 557**): "Furthermore, proteomic and metabolomic signals fluctuate with time, illness, medication, and diet, so single snapshots can misclassify biology. Longitudinal sampling may yield more reliable estimates and separate transient noise from persistent, risk-relevant biology".

o Clarify cross-modality alignment: are imaging, proteomic, and metabolomic brain-age gaps concordant, or pathways?
Thank you! This is an excellent suggestion. In the Discussion section, we have now added sentences to discuss this in more detail (Reviewer #1 also commented on this and gave a great suggestion) (**Line: 406**): "For instance, brain ProtBAG showed the lowest aging burden around 7.7 hours of sleep, whereas brain MRIBAG reached its inflection point closer to 6.5 hours. These findings suggest that molecular aging in the brain may require longer restorative sleep than its structural counterpart, potentially reflecting the different timescales and mechanisms by which proteomic versus imaging phenotypes capture sleep-related damage or resilience[46,48]. Alternatively, this can be due to a reverse causality, where longer sleep reflects brain aging rather than causes it, pulling the right side of the U-shaped curve upward, and making the apparent optimal sleep duration appear shorter for the brain MRIBAG (e.g., 6.5 hours). In addition, UKBB's MRI sub-study over-represents healthier and more highly educated

participants; this may attenuate or distort the right limb of the observed U-shape independent of reverse causality. In contrast, the proteomic and metabolomic BAGs are derived from broader, potentially more representative samples spanning wider health states, which could partly explain their stronger or more symmetric associations. To summarize, the differing optimal sleep durations across organ-specific clocks likely reflect the heterogeneous physiological demands and recovery processes of each organ system. For example, organs such as the brain may be more sensitive to sleep deprivation and circadian disruption, whereas metabolic or peripheral organs (e.g., liver, pancreas) may exhibit delayed or compensatory responses, resulting in distinct sleep–aging optima."

At the pathway level, using the brain ProtBAG ProWAS as an example (**Extended Data Figure 2**), our protein-set enrichment analysis identifies biological pathways linked to brain ProtBAG-related proteins, particularly those associated with RTN4R, FGFR2, ADAM22, PTPRZ1, EDIL3, MEPE, and TAFA5. Enriched pathways include as below:

- Synaptic signaling and axonal guidance (RTN4R, ADAM22, PTPRZ1): RTN4R (Nogo receptor) and ADAM22 are crucial for synaptic plasticity and axonal growth inhibition. Dysregulation of these pathways limits axonal repair and dendritic spine maintenance – processes tightly linked to neurodegeneration and gray matter atrophy, particularly in the hippocampus and prefrontal cortex (as shown in **Extended Data Figure 1b**). PTPRZ1 modulates oligodendrocyte differentiation and myelination; decreased function has been associated with white matter integrity loss and accelerated cortical thinning with aging (**Extended Data Figure 1c**).
- Growth factor and cell signaling (FGFR2, EDIL3): FGFR2 (Fibroblast Growth Factor Receptor 2) participates in neurodevelopment, angiogenesis, and neurogenesis. Altered FGFR2 signaling contributes to impaired neuronal maintenance and reduced cortical volume in aging and Alzheimer's disease. EDIL3, an extracellular matrix glycoprotein, influences vascular stability and blood–brain barrier integrity, both implicated in microvascular-driven atrophy in older adults.
- Extracellular matrix and neuroinflammation (MEPE, TAFA5): MEPE (Matrix Extracellular Phosphoglycoprotein) and TAFA5 are involved in ECM remodeling and immune signaling. Aberrant ECM and glial responses can lead to neuroinflammatory cascades and neuronal loss, mechanisms underlying diffuse brain atrophy patterns observed in biological aging.

As an illustration, together, these pathways represent axon guidance, ECM remodeling, neurotrophic signaling, and vascular homeostasis: four key molecular systems that sustain brain structure. Dysregulation of these processes is commonly observed in aging-related atrophy, Alzheimer's disease, and cerebrovascular pathologies, aligning well with the brain ProtBAG's molecular signature of biological aging.

Appropriate use of statistics and treatment of uncertainties
• Statistical approaches are appropriate and clearly applied, including nonlinear modeling, mediation, and replication tests.
Thank you!

• Please specify more explicitly how the "optimal model" was defined (EDF, smoothing parameters, model fit).

Thank you for this comment to improve the readability and reproducibility of our work. In the revised manuscript, we have provided more details on this (**Line: 658**): "Model selection was conducted by systematically evaluating combinations of smoothing dimensions (k = 3, 5, 10, 15, 20) and distribution families (Gaussian, t, and Gamma). For each candidate model, the estimated EDF and smoothing parameters were optimized internally via penalized maximum likelihood estimation. The optimal model was defined as the one yielding the lowest Akaike Information Criterion, indicating the best balance between model fit and complexity, while ensuring that the EDF did not approach the upper limit of the specified k, thereby avoiding overfitting."

• Age-stratified analyses appear non-significant; could the authors should briefly discuss possible reasons (sample size, age range).
Thank you for this insightful observation. As shown in **Supplementary Figure 2**, the age-stratified analyses largely replicated the main findings, with only non-significant associations observed in the 60–80-year age group for the skin ProtBAG and brain MRIBAG, and in the 40–60-year group for the pancreas MRIBAG. The absence of significant signals in these age strata likely reflects reduced sample sizes and narrower age distributions within each subgroup, which limit statistical power and the detectable variance in biological age measures. We have clarified this in the figure caption as follows: "Age-stratified analyses largely reproduced the main associations, with only non-significant effects observed in the 60–80-year group for skin ProtBAG and brain MRIBAG, and in the 40–60-year group for pancreas MRIBAG. Reduced statistical power and a narrower age range within strata may explain the lack of significant findings in these age groups"

Conclusions: robustness, validity, reliability
• The conclusions are generally robust and well-supported by the evidence.
Thank you!

• Replication across independent datasets lends credibility.
Thank you!

• The discussion could directly address the challenges of replicating UKBB-scale studies and the value of complementary smaller cohorts.
Thank you! We totally agree with this. In the Limitation section, we have emphasized this as a limitation (**Line: 547**): "First, although our findings revealed U-shaped associations between sleep duration and various phenotypes, these patterns require external validation in independent cohorts, potentially on the same scale as UKBB."

Suggestions for revision/Questions
1. Different optimal sleep durations: Provide a biological rationale for differing optima across organ clocks.
Thank you! In the revised manuscript, we have now added this (**Line: 418**): "To summarize, the differing optimal sleep durations across organ-specific clocks likely reflect the heterogeneous physiological demands and recovery processes of each organ system. For example, organs such as the brain may be more sensitive to sleep deprivation and circadian disruption, whereas metabolic or peripheral organs (e.g., liver, pancreas) may exhibit delayed or compensatory

responses, resulting in distinct sleep–aging optima." After we added more details for the brain MRIBAG and brain ProtBAG (**Line: 406**): "For instance, brain ProtBAG showed the lowest aging burden around 7.7 hours of sleep, whereas brain MRIBAG reached its inflection point closer to 6.5 hours. These findings suggest that molecular aging in the brain may require longer restorative sleep than its structural counterpart, potentially reflecting the different timescales and mechanisms by which proteomic versus imaging phenotypes capture sleep-related damage or resilience[46,48]. Alternatively, this can be due to a reverse causality, where longer sleep reflects brain aging rather than causes it, pulling the right side of the U-shaped curve upward, and making the apparent optimal sleep duration appear shorter for the brain MRIBAG (e.g., 6.5 hours). In addition, UKBB's MRI sub-study over-represents healthier and more highly educated participants; this may attenuate or distort the right limb of the observed U-shape independent of reverse causality. In contrast, the proteomic and metabolomic BAGs are derived from broader, potentially more representative samples spanning wider health states, which could partly explain their stronger or more symmetric associations."

2. Circadian timing and sleep structure: Briefly note whether circadian misalignment or fragmentation might influence associations.
Thank you for this excellent comment. In the revised manuscript, we have added this as a limitation: (**Line: 564**; after we discussed the different optimal sleep durations for the brain MRIBAG and ProtBAG) "In addition, circadian misalignment and sleep fragmentation were not directly assessed, which may influence the observed associations between sleep duration and organ-specific biological aging."

3. Cross-modality alignment: Clarify how organ-specific imaging vs. proteomics results relate.
Thank you for this great suggestion! In the revised manuscript, we addressed this in more detail using the brain MRIBAG and brain ProtBAG as an example (**Line: 406**): "For instance, brain ProtBAG showed the lowest aging burden around 7.7 hours of sleep, whereas brain MRIBAG reached its inflection point closer to 6.5 hours. These findings suggest that molecular aging in the brain may require longer restorative sleep than its structural counterpart, potentially reflecting the different timescales and mechanisms by which proteomic versus imaging phenotypes capture sleep-related damage or resilience[46,48]. Alternatively, this can be due to a reverse causality, where longer sleep reflects brain aging rather than causes it, pulling the right side of the U-shaped curve upward, and making the apparent optimal sleep duration appear shorter for the brain MRIBAG (e.g., 6.5 hours). In addition, UKBB's MRI sub-study over-represents healthier and more highly educated participants; this may attenuate or distort the right limb of the observed U-shape independent of reverse causality. In contrast, the proteomic and metabolomic BAGs are derived from broader, potentially more representative samples spanning wider health states, which could partly explain their stronger or more symmetric associations. To summarize, the differing optimal sleep durations across organ-specific clocks likely reflect the heterogeneous physiological demands and recovery processes of each organ system. For example, organs such as the brain may be more sensitive to sleep deprivation and circadian disruption, whereas metabolic or peripheral organs (e.g., liver, pancreas) may exhibit delayed or compensatory responses, resulting in distinct sleep–aging optima."

4. Dynamics of omics data: Acknowledge variability and value of longitudinal measures.

Thank you! In the revised manuscript, we have added this in the Limitation section (**Line: 554**): "Third, the cross-sectional design of this study limits our ability to determine causality or the direction of effect; although our current analysis positions sleep disturbance as a modifiable risk factor, longitudinal follow-up is needed to clarify whether sleep disturbance is a modifiable risk factor or a consequence of disease burden. Furthermore, proteomic and metabolomic signals fluctuate with time, illness, medication, and diet, so single snapshots can misclassify biology. Longitudinal sampling may yield more reliable estimates and separate transient noise from persistent, risk-relevant biology."

5. Definitions:
o Define GTEx and PGC on first mention.
Thank you! We have addressed the two points accordingly!

o Specify that F419, G473 are ICD codes.
This is a great suggestion to increase the readability! In the revised manuscript, we have explicitly specified this is an ICD code at its first occurrence (**Line: 270**): "Within the brain-related disorders, short sleep was significantly associated with depressive episodes (F329), anxiety disorders (ICD code: F419), and primary insomnia (G473),"

o Correct typos (e.g., "staitsics").
Corrected! Thank you!

6. Repetition: Streamline duplicated description in Supplementary Note 4 and File 8.
We have double-checked the two sections and ensured no duplication. Thank you!

7. Figure 4: Clarify whether it involves a second visit.
Thank you! In the revised manuscript, we have now specified in the figure's caption (**Line: 356**): "We tested whether 7 MRIBAGs mediate the effect of abnormal sleep duration on two AI-derived subtypes of late-life depression (LLD1 and LLD2[8], using only the first visit of MRI)."

8. Replication challenges: Explicitly discuss limitations of "stitching" multiple datasets versus controlled cohorts.
Thank you! We have now added this in the Limitation section (**Line: 547**): "First, although our findings revealed U-shaped associations between sleep duration and various phenotypes, these patterns require external validation in independent cohorts, potentially on the same scale as UKBB, and additional consideration of harmonization of resources from different studies and protocols."

9. ProtBAG/MetBAG interpretation: Clarify how blood-based clocks are linked to multiple organs (direct vs. MAGMA inference).
Thank you! We are sorry that we did not make this clear in the initial submission.

- The proteome and metabolome-based aging clocks were trained based on what the field calls organ-enriched proteins, which were initially proposed by Oh et al., 2024, Nature (https://www.nature.com/articles/s41586-023-06802-1): "*We used the Gene Tissue Expression Atlas (GTEx) human tissue bulk RNA-seq database18 to identify organ-enriched genes and plasma proteins (**Extended Data Fig. 1**). Tissue gene*

expression data were normalized using the DESeq2 (ref. 67) R package. We define organ-enriched genes in accordance with the definition proposed by the Human Protein Atlas: a gene is enriched if it is expressed at least four times higher in a single organ compared to any other organ. Within GTEx, we grouped tissues of the same organ together, such that a gene's expression level for a given organ was the maximum gene expression value among its subtissues. For example, all GTEx brain regions were considered subtissues of the brain organ. We define the immune organ, which is not a GTEx tissue, as expression in the blood and the spleen tissues. Organ-enriched genes were mapped to the 4,979 plasma proteins quantified in the v.4 SomaScan assay." In our previous Nature Aging paper (https://www.nature.com/articles/s43587-025-00928-9) to derive the 11 ProtBAG, we used a similar but more comprehensive approach because the HPA platform uses a much broader set of resources beyond GTEx, as well as proteomics data on top of RNA data.

- **MAGMA-based inference:** We used MAGMA gene property analysis to assess tissue- and organ-specific enrichment by integrating GWAS summary statistics with GTEx gene expression profiles. In this framework, MAGMA first maps SNP-level associations to genes and then tests whether genes with higher expression in a given tissue show stronger GWAS signals, thereby identifying tissues most relevant to the genetic architecture of the trait. Notably, this analysis relies on GWAS-derived genetic effects rather than protein-level associations.

10. Optimal model criteria: Define explicitly.

This is a great suggestion! In the revised manuscript, we have provided more detail in **Method 2** (**Line: 658**): "Model selection was conducted by systematically evaluating combinations of smoothing dimensions (k = 3, 5, 10, 15, 20) and distribution families (Gaussian, t, and Gamma). For each candidate model, the estimated EDF and smoothing parameters were optimized internally via penalized maximum likelihood estimation. The optimal model was defined as the one yielding the lowest Akaike Information Criterion, indicating the best balance between model fit and complexity, while ensuring that the EDF did not approach the upper limit of the specified k, thereby avoiding overfitting." We also submitted the R script in this revision and will share this publicly on Zenodo.

11. Age-specific effects: Note possible reasons for null results.

Thank you for this insightful observation. As shown in **Supplementary Figure 2**, the age-stratified analyses largely replicated the main findings, with only non-significant associations observed in the 60–80-year age group for skin ProtBAG and brain MRIBAG, and in the 40–60-year group for pancreas MRIBAG. The absence of significant signals in these age strata likely reflects reduced sample sizes and narrower age distributions within each subgroup, which limit statistical power and the detectable variance in biological age measures. We have clarified this in the figure caption as follows: "Age-stratified analyses largely reproduced the main associations, with only non-significant effects observed in the 60–80-year group for skin ProtBAG and brain MRIBAG, and in the 40–60-year group for pancreas MRIBAG. Reduced statistical power and a narrower age range within strata may explain the lack of significant findings in these age groups."

12. Table 1: Clarify that MULTI demographics aggregate existing datasets.
Thank you! We have now made this clear that MULTI is to consolidate and harmonize existing datasets in **Method 1** (**Line 574**): "The MULTI Consortium is an ongoing initiative to integrate and consolidate existing multi-organ data and multi-omics data, including imaging, genetics, and proteomics."

References
• Rather comprehensive, but ensure appropriate credit is given to prior sleep–aging studies focused on the brain and blood proteomics.
We believe that the citation for brain-focused sleep study is quite comprehensive. In the revised manuscript, we added the following citations on this topic:

- Vaquer-Alicea, A., Yu, J., Liu, H. & Lucey, B. P. Plasma and cerebrospinal fluid proteomic signatures of acutely sleep-deprived humans: an exploratory study. Sleep Adv 4, zpad047 (2023).
- Argentieri, M. A. et al. Proteomic aging clock predicts mortality and risk of common age-related diseases in diverse populations. Nat Med 30, 2450–2460 (2024).

We would appreciate it if you could provide more relevant references.

Clarity and context
• The abstract well-structured, and perhaps could briefly emphasize the multi-organ novelty.
Thank you for this suggestion! We have revised the Abstract accordingly!

• The introduction provides strong context but should highlight how this study differentiates from prior brain-only or proteomics-only approaches.
In the Introduction, we have emphasize that our study extends from the brain-only approach to a multi-organ systemic approach to answer several questions (**Line: 93**): "Prior studies have demonstrated a nonlinear, U-shaped relationship between sleep duration and several phenotype-based aging clocks[17,18,19,20], suggesting that both short (e.g., <6 h) and long (e.g., >8 h) sleep may accelerate neurobiological aging. However, it remains unclear whether this relationship generalizes beyond the brain to body systems (i.e., multi-organ[21,22,23,24]) and omics layers (i.e., multi-omics[12,22]), and is similar in males and females. Do similar U-shaped patterns emerge in the structural, functional, and molecular hallmarks of aging in organs and tissues beyond the brain? Are these U-shaped patterns and the optimal sleep duration time consistent across sex and organs?"

• Conclusions are impactful and balanced, and could acknowledge replication challenges directly.
Thank you!

Recommendation
This is a landmark study. I recommend publication after minor clarifications. The points above are not fundamental flaws but may strengthen clarity, reproducibility, and interpretability.
We sincerely thank the reviewer for their enthusiastic assessment of our work and their thoughtful comments. We are grateful for the recognition of our study's significance and have carefully addressed all suggested clarifications to further enhance the clarity, reproducibility, and

Junhao (Hao) Wen, PhD
Department of Radiology

interpretability of the manuscript. In addition, we would be grateful if you can provide some reference for proteomics-based work on this topic!

622 West 168th Street, New York, NY 10032

Referee #3 (Remarks to the Author):

Summary
The authors report associations between self-reported sleep duration and a large number of outcomes and quantify these associations as 'biological age gaps (BAGS)'. The emphasis is on U-shaped associations, and this in particular for non-brain outcomes. Genetic analyses, Mendelian randomization approaches and mediation analyses suggest that 1) mechanisms underlying associations between short sleep and outcomes are different from those for long sleep; 2 The sleep duration dependent BAGS (and derived optimal sleep duration) differ between men and women, and between organs, etc. Overall, the data suggest that modifying sleep duration may have substantial 'health benefits and modifying sleep duration may be feasible. Overall ,this is an impressive set of analyses pulling many data and analyses together which makes it in principle a significant contribution.
Thank you for your positive assessment of our manuscript. We have carefully addressed all your comments in the revised version.

Critique
As with so many other biobank-based studies, this remains a descriptive study.
We totally agree with you regarding this!

The importance of sleep for health is well recognised and documented and U-shaped associations have been reported repeatedly for sleep duration when based on self-report, but less often when based on objective sleep duration assessment.
This is true! Our manuscript's Introduction starts from the brain-focused study by linking self-reported sleep duration and biological aging clock (**Line: 93**): "Prior studies have demonstrated a nonlinear, U-shaped relationship between sleep duration and several phenotype-based aging clocks[17,18,19,20], suggesting that both short (e.g., <6 h) and long (e.g., >8 h) sleep may accelerate neurobiological aging. However, it remains unclear whether this relationship generalizes beyond the brain to body systems (i.e., multi-organ[21,22,23,24]) and omics layers (i.e., multi-omics[12,22]), and is similar in males and females. Do similar U-shaped patterns emerge in the structural, functional, and molecular hallmarks of aging in organs and tissues beyond the brain? Are these U-shaped patterns and the optimal sleep duration time consistent across sex and organs?"
However, as we explained below, using additional small-sample-sized BLSA dataset, we have shown that self-reported sleep duration and actigraphy-derived sleep duration only show a modest correction, and in nature, the two measurements capture different respect of sleep biology. Therefore, we don't expect that the exact same patterns can be entirely replicated in objective sleep duration like from actigraphy or PSG (see explanation below).

In fact, I am aware of at least one study in which both objective and sleep-reported sleep duration data were available, and the associations were strikingly dissimilar,
Yes! Your observation is correct – there are previous papers (e.g., PMCID: PMC5771411; PMCID: PMC4782764) showing that sleep duration from self-reported questionnaire vs. actigraphy or PSG-derived only show small to medium correlations (**Line: 106**): "Self-reported measures are less objective than actigraphy or polysomnography but capture different yet complementary aspects of sleep biology, with only moderate correlations between

modalities[25,26]. The large sample size (~500,000) allows robust identification of nonlinear associations using generalized additive models (**GAM; Method 2**)."

with linear associations between sleep duration and outcomes for objective sleep duration, and U-shaped associations for self-reported sleep duration (see Fig 2 in PMID: 36892074). In general, dissociations between objective and subjective measures of sleep duration are more pronounced in specific populations or health conditions such as depression, insomnia, etc. Thus, the major and significant limitation of the study is the sleep data on which it is based.

Thank you, a lot, for your comments regarding the dissociation between the self-reported and objective sleep durations from actigraphy or PSG.

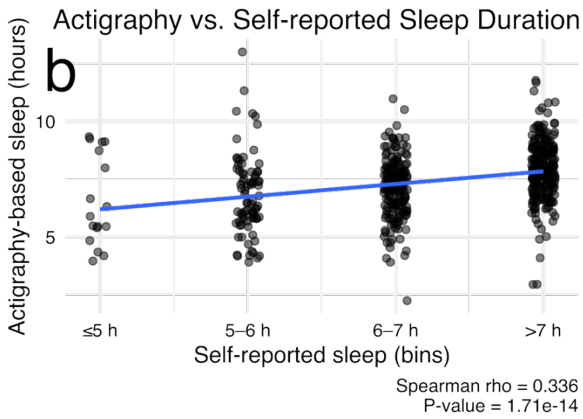
We would like to clarify the following points. First, self-reported and actigraphy-based sleep durations capture distinct aspects of sleep: self-reported measures reflect perceived or habitual sleep duration influenced by cognitive and subjective factors, while actigraphy provides an objective estimate of sleep–wake behavior based on movement data. As shown in prior studies (e.g., PMCID: PMC5771411; PMCID: PMC4782764), correlations between these modalities are typically modest, indicating that they assess complementary, rather than interchangeable, dimensions of sleep biology. Therefore, we do not expect the U-shaped association observed in self-reported sleep duration to be fully replicated using actigraphy-derived measures. Second, we are familiar with the paper the reviewer mentioned. To clarify, that study focused on modeling **mortality risk** and sleep duration (see their **Fig. 2**), whereas our analysis examines the association between sleep duration and **biological aging clocks**.

We have mentioned this in our submission (**Line: 106**): "Self-reported measures are less objective than actigraphy or polysomnography but capture different yet complementary aspects of sleep biology, with only moderate correlations between modalities[25,26]. The large sample size (~500,000) allows robust identification of nonlinear associations using generalized additive models (**GAM; Method 2**)." We also mentioned this as a limitation for future investigation (**Line: 550**): "Second, the reliance on self-reported, questionnaire-based sleep duration measurement may introduce recall bias or misclassification, and future studies incorporating objective measures, such as actigraphy or polysomnography, are essential to better understand underlying mechanisms."

In this revision, to better address your question, we have also included another NIA-supported study, although have very limited sample size to achieve enough power to detect significant signals (BLSA; *N*=385), which has actigraphy-based sleep duration, as well as self-reported sleep duration (but with only bins of sleep duration time available: ≤5 h, 5–6 h, 6–7 h, >7 h, Refused to answer, and Don't know – this differs from the questionnaire-based sleep duration from UK Biobank). Due to the limited sample size, the nature of the self-reported sleep duration, and the demographic differences (BLSA participants are much older than UK Biobank, shown in **Supplementary Table 1**, and are overall healthy participants without cognitive decline), no significant signal was detected. Here is the updated figure in the revised manuscript:

Supplementary Figure 9: External validation of the U-shaped patterns based on BLSA actigraphy-derived sleep duration

[PANEL REDACTED]



a) [TEXT REDACTED]**b**) Correlation between actigraphy-based and self-reported sleep duration, indicating only a modest concordance between objective and subjective measures, echoing previous studies[4,5]. The U-shaped association between sleep duration and biological age gap observed with self-reported sleep likely reflects perceived sleep adequacy and long-term behavioral patterns, which integrate both physiological and psychosocial dimensions of sleep health. In contrast, actigraphy-based sleep provides a more objective, momentary measure of sleep–wake activity, capturing nightly variability but less influenced by subjective factors such as fatigue, sleep satisfaction, or chronic insomnia symptoms. Self-reports tend to overestimate sleep duration and incorporate cognitive and emotional perceptions of restfulness, which are closely tied to overall well-being and biological aging processes. Conversely, actigraphy may detect shorter but fragmented sleep periods that do not necessarily correspond to self-perceived sleep deficiency. As a result, self-reported sleep duration may better reflect long-term sleep behaviors and their cumulative physiological impact, whereas actigraphy captures short-term sleep patterns that might not yet manifest as measurable biological aging differences.

As we can see, the U-shaped pattern is not observed, and the correlation between self-reported and actigraphy-based sleep duration is medium, consistent with the literature, which has been stated in our manuscript (**Line: 550**): "Self-reported measures are less objective than actigraphy or polysomnography but capture different yet complementary aspects of sleep biology, with only moderate correlations between modalities[25,26]. The large sample size (~500,000) allows robust identification of nonlinear associations using generalized additive models (**GAM; Method 2**)."

As we suggested in the figure caption in **Supplementary Figure 9**: "The U-shaped association between sleep duration and biological age gap observed with self-reported sleep likely reflects perceived sleep adequacy and long-term behavioral patterns, which integrate both

physiological and psychosocial dimensions of sleep health. In contrast, actigraphy-based sleep provides a more objective, momentary measure of sleep–wake activity, capturing nightly variability but less influenced by subjective factors such as fatigue, sleep satisfaction, or chronic insomnia symptoms. Self-reports tend to overestimate sleep duration and incorporate cognitive and emotional perceptions of restfulness, which are closely tied to overall well-being and biological aging processes. Conversely, actigraphy may detect shorter but fragmented sleep periods that do not necessarily correspond to self-perceived sleep deficiency. As a result, self-reported sleep duration may better reflect long-term sleep behaviors and their cumulative physiological impact, whereas actigraphy captures short-term sleep patterns that might not yet manifest as measurable biological aging differences."

We hope these additional experiments and discussion can sufficiently address your insightful comment!

Leaving the limitations of the major exposure variable aside, a few other aspects need to be mentioned.

Thank you for this thoughtful comment. In this revision, we have further emphasized these limitations to strengthen the discussion.

The age range of the participants in the UKBB and other cohorts is limited and most of the participants are older. Sleep duration and sleep need change dramatically over the lifespan. Thus, any statement about optimal sleep duration needs to be qualified with 'For people aged X to Y the optimal sleep duration is'.

Thank you! We totally agree with you! We have mentioned this as a limitation in our initial submission, and expect future studies/biobanks to cover the entire lifespan, but we believe this needs a community-wide effort to collect another rich set of biomedical databases like the UK Biobank (**Line: 566**): "Sixth, future research needs to extend the Sleep chart across the lifespan to better capture dynamic patterns of sleep duration beyond adults."

In this revision, regarding your comment that most of the participants are older, I think the UK Biobank represents a general population for adulthood, ranging from 37-84 years old in our study's population, which we used to derive the major results of the Sleep Chart. In the revised manuscript, we explicitly specify this age range in the Abstract (**Line: 62**): "First, a systemic, U-shaped pattern shows that both short (<6 hours) and long (>8 hours) sleep duration are linked to elevated biological age gaps (BAGs) across 9 brain and body systems and 3 omics types, with optimal sleep time varying by organ and sex ([6.4-7.8] hours) in the UK Biobank (ages 37–84 years)." We also changed the paper's title to "**Sleep chart of biological aging clocks in middle and late life**". Finally, we also made it clear in the Abstract that sleep duration is from self-reported measure: "This study proposes a Sleep Chart to assess the relationship between self-reported sleep duration and 23 biological aging clocks across 17 organ systems or tissues and 3 omics data types (imaging, proteomics, and metabolomics)."

The authors also report that optimal sleep duration varies across organs, etc. but the statistical underpinning of the significance of these differences was not clear to me. This also holds for the sex differences (which are likely to be related to a sex related sleep need).

Thank you for this important point, and we are sorry that we did not make this clear in the first submission.

In **Method 2**, we have now made it clear what the GAM tests are and how the significant signal was determined (**Fig. 1**'s caption) (**Line: 658**): "Model selection was conducted by systematically evaluating combinations of smoothing dimensions (k = 3, 5, 10, 15, 20) and distribution families (Gaussian, t, and Gamma). For each candidate model, the estimated EDF and smoothing parameters were optimized internally via penalized maximum likelihood estimation. The optimal model was defined as the one yielding the lowest Akaike Information Criterion, indicating the best balance between model fit and complexity, while ensuring that the EDF did not approach the upper limit of the specified k, thereby avoiding overfitting. We tested *i*) the main effect (sleep duration; P_1), *ii*) sex difference in population mean (P_2), and *iii*) sex-sleep interaction terms (P_3) on each BAG. The solid curves in **Fig. 1** depict estimated BAG, while shaded bands represent the 95% confidence interval (CI)."

As stated in the method, our GAM directly tests the sex term and sex * sleep duration interaction term in our models. As also suggested by other referees, in the revised manuscript, we performed an additional sex-stratified analyses:

TEXT REDACTED

FIGURE REDACTED

We split the cohort into females and males and re-fit the same GAM specification selected in the main analysis (fixed optimal k and family), testing the sleep term within each sex. Significance was assessed with Bonferroni correction ($\alpha = 0.05/23$); significant panels are marked with icons and asterisks. Overall, U-shaped associations were broadly reproducible across sexes, with some BAGs showing stronger effects in males (e.g., Endocrine_ProtBAG) and others in females (e.g., Pancreas_MRIBAG, Kidney_MRIBAG). These sex differences may reflect biology, such as hormone milieu (estrogen/androgen signaling), sex-specific sleep architecture and circadian regulation, organ-specific metabolic and inflammatory pathways, and life-stage effects (e.g., menopause), in addition to possible sample-size or residual confounding differences between strata.

Junhao (Hao) Wen, PhD
Department of Radiology

Care needs to be taken when using the word 'optimal'. For example the first sentence of the Abstract now reads 'Optimal sleep plays a vital role in promoting healthy aging and enhancing longevity'. Obviously, this statement is necessarily true and contains no information.

We thank the reviewer for this insightful comment. We agree that the use of the term "optimal" in the opening sentence could be interpreted as self-evident or imprecise. In the revised manuscript, we have clarified the statement to emphasize that our study empirically identifies sleep durations associated with the most favorable biological aging profiles, rather than implying an a priori "optimal" value. Accordingly, we have reworded and emphasized the optimal sleep duration is an empirical finding (**Line: 99**): "Are the observed U-shaped associations and empirically derived optimal sleep durations consistent across sexes and organ systems?".

We would appreciate it if you could suggest better wording regarding this!

From a naive sleep and health research perspective a first question to be asked to these data would be 'Is there a significant association between sleep duration and outcome'.

Thank you for your important insights. This is exactly what we answered and tested in **Fig. 1**. Our GAM models statistically test the significance of these associations in a data-driven fashion, as stated in the comment above.

The second question is ' what is the nature of the association'. Here, the authors jump almost immediately to the second question, which is unfortunate.

Thank you for the thoughtful suggestion. We may be misunderstanding the request and question here, as we believe **Fig. 2** (genetics) and **Extended Data Figs. 1–3** (neuroimaging for neurobiology; proteomics and metabolomics for molecular mechanisms and pathways) already address this point at least in part.

In the revised manuscript, we have made this more clear (**Line: 93**): "Prior studies have demonstrated a nonlinear, U-shaped relationship between sleep duration and several phenotype-based aging clocks[17,18,19,20], suggesting that both short (e.g., <6 h) and long (e.g., >8 h) sleep may accelerate neurobiological aging. However, it remains unclear whether this relationship generalizes beyond the brain to body systems (i.e., multi-organ[21,22,23,24]) and omics layers (i.e., multi-omics[12,22]), and is similar in males and females. Do similar U-shaped patterns emerge in the structural, functional, and molecular hallmarks of aging in organs and tissues beyond the brain? Are the observed U-shaped associations and empirically derived optimal sleep durations consistent across sexes and organ systems?"

We would appreciate it if you could be more specific if you are suggesting specific comments to improve the organization of our manuscript.

In addition, in several places in the MS the authors use the phrase 'beyond the brain'. I don't understand this emphasis on 'beyond the brain'. The field accepts that sleep has benefits for both the brain and body.

Thank you for this important point. Our primary motivation was to investigate sleep in a holistic, multi-organ framework, recognizing that sleep exerts systemic effects across both the brain and body. As also noted by Reviewer #2, who encouraged us to emphasize this multi-organ perspective, we consider this integrative approach to be one of the key novelties and strengths of our manuscript. This has been emphasized in the Introduction (**Line: 93**): "Prior studies have demonstrated a nonlinear, U-shaped relationship between sleep duration and several phenotype-based aging clocks[17,18,19,20], suggesting that both short (e.g., <6 h) and long (e.g., >8 h) sleep may

accelerate neurobiological aging. However, it remains unclear whether this relationship generalizes beyond the brain to body systems (i.e., multi-organ[21,22,23,24]) and omics layers (i.e., multi-omics[12,22]), and is similar in males and females. Do similar U-shaped patterns emerge in the structural, functional, and molecular hallmarks of aging in organs and tissues beyond the brain? Are the observed U-shaped associations and empirically derived optimal sleep durations consistent across sexes and organ systems?"

We would appreciate it if you could suggest better wording regarding this!

The general message and the presented evidence that sleep is a modifiable risk factor is attractive. However, and given the U-shaped associations so much emphasized by the authors, the question whether self-reported long sleep is a cause or a consequence of outcomes becomes rather important. If it is a cause, then long sleepers should reduce their time in bed. If it is a consequence, then the advice may be rather different. The authors do not seem to be able to come to a clear conclusion with respect to this question (see for example the Abstract but also elsewhere in the manuscript), which is a very important question in the sleep-health field.

Thank you for raising this key issue. We intentionally adopt a cautious stance regarding the directionality of sleep–disease relationships. In our cohort (main individual-level data from UK Biobank), we used structural equation modeling leveraging temporal ordering (sleep duration at the baseline visit; MRI measures at the subsequent visit), and found evidence consistent with an effect of sleep on downstream outcomes (LLD). We also performed Mendelian randomization analyses to probe reverse causation (disease → sleep disturbance) and did not find support for inverse effects (after multiple comparisons). That said, MR relies on strong assumptions (e.g., instrument validity, presence of pleiotropy, etc.), and our null results from MR may reflect the violation of underlying assumptions (although we provided extensive sensitivity check results), or the limited power of available GWAS rather than the true absence of effect. This is the reason why in the Abstract, we stated this with a conserved tone (**Line: 66**): "Finally, short and long sleep duration are associated with late-life depression via distinct pathways: long sleep may contribute indirectly through biological aging processes, while short sleep shows a more direct link. Although our Mendelian randomization does not show strong causal effects from disease to sleep disturbances, it does not fully rule out the possibility that sleep disturbances may, in part, reflect underlying disease burden."

Most importantly, we agree that the gold-standard approaches to adjudicate causality are longitudinal modeling with repeated measures (very promising with the advance of UK Biobank) or randomized/clinical interventions that manipulate sleep. We have revised the Abstract and Discussion to reflect this caution, emphasizing that while our findings more likely support sleep as a modifiable risk factor, definitive causal inference requires adequately powered longitudinal studies and trials (**Line: 553**): "Third, the cross-sectional design of this study limits our ability to determine causality or the direction of effect; although our current analysis positions sleep disturbance as a modifiable risk factor, longitudinal follow-up is needed to clarify whether sleep disturbance is a modifiable risk factor or a consequence of disease burden."

We sincerely thank you for the time and effort you dedicated to reviewing our work. Your constructive feedback has been invaluable in improving the manuscript, and we hope our revisions have adequately addressed your comments!

Junhao (Hao) Wen, PhD
Department of Radiology

Author response letter, ID: NATURE-2025-05-11244B

We thank the editor for allowing us to revise our manuscript and the 4 reviewers for their thoughtful comments. In this round of revision, we carefully addressed all comments. This response letter displays the reviewers' comments in black font, while our responses are in blue. In the main manuscript, the revised part is highlighted in yellow, as well as in this letter.

Major changes for this round of revision:

- We have carefully addressed Reviewer #4's new comments.
- We have carefully addressed additional editorial comments by reorganizing the manuscript, improving statistical rigor, and transparency.
- We have finished the full development of the SleepChart portal (https://labs-laboratory.com/sleepchart/), where the community can download our GWAS summary statistics and explore our experiment results.
- We also made the original scripts and related source files available at Zenodo (https://zenodo.org/records/17409426).
- We have carefully proofread the manuscript for grammar, typos, statistics, and analyses.

622 West 168th Street, New York, NY 10032

Referee #1 (Remarks to the Author):

The authors have thoroughly addressed all of my comments and comprehensively integrated my suggested analyses. I have no further comments and wish them the best of luck with the remainder of the submission process.

We thank you for your thoughtful comments to improve our study!

Junhao (Hao) Wen, PhD
Department of Radiology

Referee #2 (Remarks to the Author):
I thank the authors for their comprehensive and thoughtful revisions. The manuscript is improved, and the majority of my prior concerns have been fully addressed.
We thank you for your thoughtful comments to improve our study!

I have only a small number of remaining suggestions that focus on transparency, clarity, and reproducibility.
We further fully addressed your remaining suggestions, which are very critical to improving our manuscript!
1. Supplementary display of the remaining 14 BAGs
Figure 1 is elegant and appropriately highlights the 9 significant BAGs, which should remain in the main figure. For transparency and to allow readers to examine the full modeling results, I suggest adding a supplementary figure showing the GAM curves for the remaining 14 BAGs. A single multi-panel supplementary figure would reassure readers that all 23 BAGs were analyzed using the same modeling framework, the absence of significance in the remaining BAGs reflects the modeled results rather than omission, and the U-shaped pattern is not selectively shown.
We thank you very much for this suggestion. We want to ensure we correctly understand your suggestion: "*I suggest adding a supplementary figure showing the GAM curves for the remaining 14 BAGs*". We are not sure whether you are asking:

- i) adding the vertical dash lines (i.e., "optimal" sleep hours) for non-significant results
- ii) adding the GAM modeling results for non-significant results for some of our sensitivity analyses (which we only showed the 9 significant BAG-sleep signals in the original Supplementary Figures).

To fully address your comment, in the revised manuscript, we directly provided revised results for both aspects:

- (**Line: 147**): "Detailed statistics, including the P value, sample size, chosen family distribution, and EDF, are presented in **Supplementary Table 2** and **Supplementary File 1.**" The full statistics for **Fig. 1** are too big to put in **Supplementary Table 2**, so we put them in **Supplementary File 1**. We also added the vertical "optimal" sleep duration for all 23 BAGs for transparency in **Fig. 1**.
- (**Line: 150**): "Several sensitivity analyses were performed. **Supplementary Figure 1** displays Quantile-Quantile (QQ) plots and residual vs. fitted (RVF) plots for the best model of each BAG. **Supplementary Figure 2-16** and **Supplementary Table 3** present a full set of analyses to scrutinize the robustness of the U-shaped pattern, including a potential replication of the U-shaped relationship in two independent datasets[28,29], though the sample sizes (*N*=385 and 573) are much smaller and the populations are much older than UKBB.**"**
- **Supplementary File 1** and **Supplementary Table 3** contain the statistics for all main GAM modelling and the sensitivity check analyses.

2. Clarification regarding MR instrument strength (Line ~1255)
Although several SNP-level R2 values are small, this is entirely typical for MR instruments of complex behavioral and physiological traits. What matters for instrument strength is the aggregated F-statistic, not the individual R2 values.

622 West 168th Street, New York, NY 10032

I recommend adding a brief clarifying statement noting that the combined variance explained is R2=0.00113 , this corresponds to a robust F-statistic ≈ 38.7 (N = 274,330; k = 8), which exceeds the conventional F > 10 threshold defining adequate instruments in univariable Mendelian randomization. This is well done for the subsequent figure already.

We thank you for your initial and current comments to scrutinize the instrumental strength of the MR IVs.

Following your current comment, we now explicitly state that the combined variance explained (R^2_total = 0.00113) corresponds to an overall F-statistic of 38.66 (N = 274,330; k = 8), exceeding the conventional F > 10 threshold for adequate instruments in univariable Mendelian randomization in **Extended Data Figure 6**: "**a**) Lollipop plot of F-statistics for the 8 genome-wide significant depression instruments after preprocessing, considering linkage disequilibrium, ordered from weakest to strongest. All instruments exceed the commonly used rule-of-thumb threshold of F=10[109] (min F=30.70; median F=36.88; mean F=38.62). The combined variance explained is R^2-total = 0.00113, which corresponds to an approximate overall F-statistic of 38.66 (N=274,330; k=8), exceeding the conventional F>10 threshold that defines adequately strong instruments in univariable Mendelian randomization."

109. Staiger, D. & Stock, J. H. Instrumental Variables Regression with Weak Instruments. Econometrica 65, 557–586 (1997).

3. Optional note distinguishing EDF from effect size

Some readers may misinterpret the effective degrees of freedom (EDF) in the GAM models as reflecting the strength or magnitude of the association. A brief sentence in the Methods or Supplementary Note explaining that EDF reflects spline flexibility (curve complexity), not effect size, would preempt this misunderstanding.

We thank you for this suggestion to avoid potential confusion. To address this, in the revised manuscript where the EDF was mentioned for the first time (**Line: 125**), we explicitly stated this: "We assessed the nonlinearity of the sleep-BAG association using the effective degree of freedom (EDF) as a proxy, which reflects the complexity of the smoothing term, rather than the effect size of the overall model fit to the outcome"

4. Optional Supplementary table with full GAM statistics

Although not required, a small supplementary table listing—for all 23 BAGs—the chosen family distribution, EDF, P-values, and relevant covariates would further enhance transparency and reproducibility. You already report these statistics for the significant BAGs; extending this to all outcomes would help readers who wish to examine the modeling details more closely.

We appreciate your suggestions to improve research transparency and reproducibility, which has been a core principle of our work.

In the original submission, we provided the statistics for all 23 BAGs, not just the 9 significant BAGs' results in **Supplementary Table 2**, but not the full statistics for all covariates, nor the sample size *N*.

To fully address this, we now provide all statistics for these covariates in **Supplementary File 1** in the revised manuscript (as the table becomes large and cannot be easily put inside the Supplementary Materials). In addition, we added the sample size for each BAG-sleep association in **Supplementary Table 2**.

Junhao (Hao) Wen, PhD
Department of Radiology

5. Optional acknowledgment of MR linearity assumptions
Given that the primary phenotypic analyses emphasize U-shaped nonlinear associations, it may be helpful to add a sentence noting that MR methods assume linear genetic associations with the exposure. This is standard for MR but would help frame the interpretational scope of causal inference in contrast to the observed nonlinear sleep–phenotype relationships.
We sincerely appreciate that you made this point explicit, and we fully agree with your thoughts. It is precisely why we modelled sleep duration as a binary exposure (long vs. normal and short vs. normal), rather than as a continuous trait. We acknowledge that the MR methods themselves do not account for potential non-linearity in the exposure–outcome relationship, and we agree this limitation should be clearly stated for transparency.

To fully address this, we have added sentences in the Method section for our MR analyses (**Line: 877**): "Of note, our Mendelian randomization analyses rely on the standard assumption of linear genetic effects on sleep duration, providing an average causal effect per unit increment in the exposure. As such, we conducted the sleep GWAS using binary traits (long sleep vs. normal sleep and short sleep vs. normal sleep), rather than treating sleep duration as a continuous variable. Nevertheless, the MR estimates do not directly represent the nonlinear U-shaped sleep-BAG relationships observed in our phenotypic analyses and should be interpreted as complementary to, rather than a direct mirror of, the observational sleep–BAG associations."

Overall assessment
These suggestions are minor and aimed solely at enhancing clarity and transparency. The scientific content is strong and comprehensive, and the revisions have been thorough and thoughtful, going beyond what was expected. I believe the manuscript will be further strengthened by addressing items (1) and (2), with the optional items providing additional clarity for interested readers.
Thank you for your thoughtful and critical review of our manuscript. Your comments have been instrumental in sharpening both the methodological/technical aspects and the clinical implications of our work!

622 West 168th Street, New York, NY 10032

Referee #3 (Remarks to the Author):

The Authors have been responsive to my comments. One of my major concern was that for the reader of the abstract is was not clear that all the findings are based on sleep-reported sleep duration which is very different from behaviour or physiology based assessment of sleep. From the Abstract it is now clear that all results are based on self-reports in a cross-sectional study. This is a limitation but the results are impressive and important.

We sincerely thank you for your constructive comments highlighting several important limitations of our study. These points are crucial for balancing the potential societal impact of our findings (in order not to overclaim until future longitudinal data or interventional studies are available) with the constraints of the available data, and for guiding future work to deepen our understanding of the intriguing relationship between sleep and biological aging!

Junhao (Hao) Wen, PhD
Department of Radiology

Referee #4 (Remarks to the Author):

Summary:

This paper provides an in-depth analysis of the association of self-reported short and long sleep duration with multiple biological clocks (BAGs) using diverse samples. Further, the authors conduct GWAS to identify genetic variants for sleep duration; genetic correlations among sleep and other variables; MR (to explore disease-sleep duration causal associations); incident analyses (to identify role of sleep duration in mortality and MRI markers of LLD); and mediation (SEM) models to explore biological pathways linking sleep to disease. They report novel and strong U-shaped associations with BAGs (consistent with an extensive literature reporting similar associations with several chronic diseases). The reported differences in short and long sleep duration associations also vary by tissue and BAG source, as well as by sex. They point out that they observe more associations for short vs long sleep with incident disease, although both were associated with mortality. They report that mediation analysis demonstrated stronger direct effects between short sleep and MRI-LLD, while associations for long sleep were partially "mediated" through MRIBAGs. The paper proposes many intriguing biological causal pathways that differ for long and short sleep. Given the limitations of data and analyses, its strongest and most novel finding is that self-reported sleep duration is associated with multiple BAGs-with variation by tissue source/BAG data- in a non-linear pattern which varies by sex.

We thank you for your precise summary and careful reading of our manuscript. We will carefully address each of your comments below to further strengthen the rigor of our analyses and appropriately acknowledge the relevant limitations.

Originality:

The paper presents a comprehensive and highly original account of how multimodal and organ-specific biological aging estimates relate to self-reported sleep duration, showing the expected (from prior disease-specific associations) nonlinear pattern, with accelerated aging associated with both shorter and longer sleep duration. The use of multiple data modalities (MRI, proteomics, and metabolomics) and organ-specific measures reveals a detailed yet remarkably consistent U-shaped association, providing estimates of "optimal" self-reported sleep duration for each organ system (although with caveats noted below). The authors also employ a diverse set of analyses to examine the genetic underpinnings, causal effects, and predictive power of short and long sleep duration in relation to various health outcomes. While these latter analyses are complementary, several analyses (GWAS, incident disease assessments, MR, etc) have been reported before using the same data sources.

Thank you for these comments! We agree that several related analyses have previously been conducted using the same underlying resources (primarily the UK Biobank), but we included additional studies (although with rather smaller sample sizes). Maybe more innovatively, we performed these analyses from a multi-organ perspective. To the best of our knowledge, our study is the first to characterize the U-shaped relationship between sleep duration and biological aging across multiple organs and omics layers. In the Introduction, we already cite prior work linking biological aging, either brain age from MRI or traditional/conventional non-organ-specific clocks based on phenotypic data, to self-reported sleep duration, but none of these studies have examined this question through a unified multi-organ, multi-omics framework.

Regarding the genetic analyses of sleep traits, including the important studies you highlight below (we already cited some of them), we now, in the revised manuscript, provide a

more detailed comparison, explanation, and rationale relative to prior GWAS and genetic correlation work (see **Supplementary Note 3**, which we also addressed in another comment you listed below). In our opinion, we believe our analyses slightly differ from several previous sleep GWASs. First, our genetic analyses were guided by the hypothesis that sleep-duration disruptions reflect multi-organ processes that influence brain–body interconnections. Second, we modeled sleep duration as a binary trait (short vs. normal; long vs. normal) to better accommodate known non-linear relationships that may be obscured when sleep is treated as continuous. Third, publicly available GWAS summary statistics often lack key fields—such as allele frequencies and effect-size measures (e.g., beta or OR)—that are needed for genetic correlation and Mendelian randomization analyses to properly account for linkage disequilibrium. We address these points in more detail in response to your specific comments below. Additional experiments by leveraging pioneering sleep duration GWAS from the literature are summarized in **Supplementary Note 3**.

Methods, data quality, analyses and conclusions:
The very limited reference to methodological details in the Results section made it challenging to understand which specific sample (**cohort**) and statistical model (**covariates included**) were used for each set of analyses. Given the heterogeneity across samples and need to develop statistical models specific for each hypothesis, it was difficult to fully appreciate the strength of the findings. Providing a bit more on the methodological details in the Results would improve interpretation of the results.

We thank you for the suggestion and totally agree that this will improve the transparency and reproducibility of our work.

In the revised manuscript, we have added further detail for **the cohort for each analysis** in the **Methods** section.

Example changes for these details are below:

- **Method 1**: We have added the definition of the sleep duration variables from each study, and also the description and characteristics of each subpopulation for the multimodal data from UKBB.
- **Method 2**: We believe that for the GAM modeling, we provided enough details (e.g., covariates included, as well as a full set of sensitivity analysis suggested by you and other reviewers), along with the source code for the GAM modeling and the BAG training, to ensure reproducibility.
- **Method 3**: We added data splitting and cross-validation procedures to derive 23 BAGs, which. All these details are also in our previous BAG studies (cited in the paper in the Abstract), which improve reproducibility and transparency.
- **Method 4-6**: We believe that we provided enough details, regarding covariates included, P-value threshold for multiple comparisons, etc. We specified that these analyses used data from UKBB.
- **Method 7**: For all genetic analyses, we believe that we i) specified individual genetic data are from UKBB and summary-level data are from FinnGen and PGC (e.g., for genetic correlation and MR), and ii) provided enough details (e.g., covariates, p-value threshold) and references (the original population genetics statistical method) for each analysis.
- **Method 8**: For survival analyses, we gave details of the models and covariates included, and how we defined case and non-case participants.

Junhao (Hao) Wen, PhD
Department of Radiology

- **Method 9**: For structural equation modeling, we have specified analysis is done in UKBB, provided details of the assumptions and methods for multiple comparisons, etc. I don't think there is a commonly-cited method paper for SEM, but we cited our previous study where we performed a similar analysis using SEM (https://www.nature.com/articles/s41591-023-02296-6).
- **Method 10**: For Mendelian randomization, we detailed our preprocessing pipeline for selecting exposure and outcome variables, choosing instrumental variables, harmonizing GWAS summary statistics, and conducting sensitivity analyses, and we included a cautionary note regarding interpretation given the strong statistical assumptions underlying MR. Exemplary sensitivity check analyses are presented in **Extended Data Figure 6-7**, which is also suggested by other reviewers.

In addition, to make sure the analyses are reproducible, we also listed all the packages and software used in our paper, either from our own group or from open-sourced software (see **Code Availability** section).

Regarding the **covariates included** in all analyses, we believe that we have already provided sufficient information, as illustrated below:

- **Method 2: GAM models (Line: 639)**: "We adjusted for key demographic and physiological covariates (i.e., age, sex, weight, height, waist circumference, BMI, assessment center, diastole, systolic blood pressure, time differences for data collection (for MRIBAGs), and disease status)." Additional sensitivity analyses were presented in **Supplementary Figures 1-16**.
- **Method 4: IDP-wide associations**: "These models included age, sex, body mass index, height, weight, waist circumference, assessment center, blood pressure, and disease status as covariates. Additionally, organ-specific covariates were incorporated, such as brain positioning in the scanner (lateral, transverse, and longitudinal), head motion, and intracranial volume for the brain IDPs. Outlier values in the IDP outcome variables, defined as ±4 standard deviations from the mean, were removed to minimize the influence of extreme values."
- **Method 5: Proteome-wide associations**: The GAM was adjusted for common covariates, including age, sex, weight, height, waist circumference, BMI, assessment center, disease status, diastolic and systolic blood pressure, protein batch number, limit of detection, and the first 40 genetic principal components.
- **Method 6: Metabolome-wide associations**: The GAM controlled common covariates, including age, sex, weight, height, waist circumference, BMI, assessment center, disease status, diastolic and systolic blood pressure, and the first 40 genetic principal components.
- **Method 7: Genetic analyses**: Our GWAS adjusted common variates, including age, disease status, age-squared, sex, interactions of age with sex, BMI, waist circumference, standing height, weight, systolic/diastolic blood pressure, and the first 40 genetic principal components.

We intentionally made the Result section concise, as suggested by the reviewers in the previous round, for readability, and to follow the editorial guidance for word constraints. We'd appreciate it if you can be more specific, and we will address them with more details if you think some details are missing in the current Method section.

Given the interest in this paper in causal inference, it would be helpful to see a justification for inclusion of variables in each set of analyses. For example, smoking and alcohol both influence sleep and many of the outcomes of interest, but did not appear to be considered. Other important variables that are confounders or precision variables such as physical activity and socioeconomic status should be considered.

We thank you for your suggestion. As we showed in the above-mentioned comment, we are transparent for all covariates included in the analyses, and we also already performed additional sensitivity check analyses in the last revision provided by the other 3 reviewers (e.g., some sex hormones). Our initial strategies for choosing the included covariates are based on *i*) previous literature and ii) availability of data to cover the entire full sample size of the 500K UKBB populations to ensure fair statistical power, because self-reported sleep duration is relatively subjective and noisy.

First, to address your specific comments, in the revised manuscript, we added additional sensitivity analyses for the GAM modeling to include:

- smoking status (Field 20116): This variable covers 500k full population.

TEXT REDACTED

FIGURE REDACTED

Generalized additive models were fit in the UK Biobank after adding smoking status (Field ID: 20116) as an additional covariate in the original GAM model. Dashed vertical lines indicate the sex-specific sleep duration corresponding to the minimum predicted BAG (“optimal” sleep). Colored icons correspond to the 9 significant BAGs identified in **Fig. 1**. After accounting for smoking status in the GAM, the U-shaped relationship pattern between sleep and BAG was robust, although the curve and the "optimal" sleep duration showed slight variations.

Junhao (Hao) Wen, PhD
Department of Radiology

- alcohol consumption (Field 20117): This variable should cover the entire UKBB population, but somehow, in our local data, we only had 46,191 participants who covered our brain MRI population. I think this was because we forgot to refresh our data basket for this variable. Requesting data refreshment from UK Biobank data will be time-consuming (more than months), so in this revision, we only did the sensitivity check analysis for the brain MRIBAG.

TEXT REDACTED

FIGURE REDACTED

Generalized additive models were fit in the UK Biobank after adding alcohol consumption status (Field ID: 20117) as an additional covariate in the original GAM model for the brain MRIBAG (our data only covers the MRIBAG population for this variable). Dashed vertical lines indicate the sex-specific sleep duration corresponding to the minimum predicted BAG ("optimal" sleep).

- Townsend deprivation index (Field 22189): This variable covers 500k full population.

TEXT REDACTED

622 West 168th Street, New York, NY 10032

FIGURE REDACTED

Generalized additive models were fit in the UK Biobank after adding the Townsend deprivation index (TDI; Field ID: 22189) as an additional covariate in the original GAM model. Dashed vertical lines indicate the sex-specific sleep duration corresponding to the minimum predicted BAG (“optimal” sleep). Colored icons correspond to the 9 significant BAGs identified in **Fig. 1**. After accounting for TDI in the GAM, the U-shaped relationship pattern between sleep and BAG was robust, although the curve and the "optimal" sleep duration showed slight variations (e.g., endocrine MetBAG).

- Computer time use (Field 1080): For physical activity variables, based on the variables available at our local data, we chose this variable that covers all 500k population. We also tried to use the variables from this category ID (https://biobank.ndph.ox.ac.uk/ukb/label.cgi?id=100108), but it seems that these variables do not include the full UK Biobank population (500k).

TEXT REDACTED

Junhao (Hao) Wen, PhD
Department of Radiology

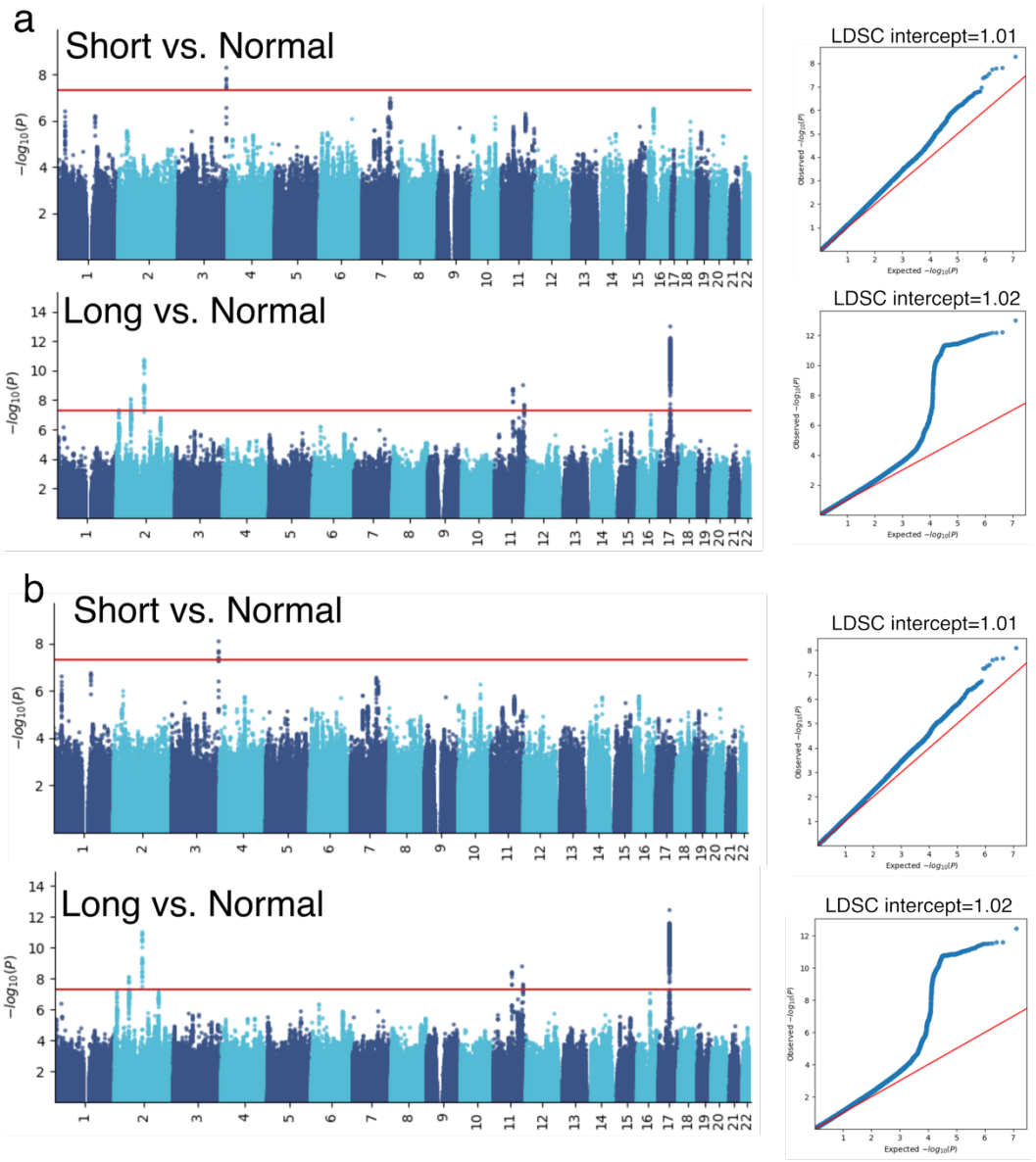
FIGURE REDACTED

Generalized additive models were fit in the UK Biobank after adding the computer use time (Field ID: 1080) as an additional covariate in the original GAM model. Dashed vertical lines indicate the sex-specific sleep duration corresponding to the minimum predicted BAG ("optimal" sleep). Colored icons correspond to the 9 significant BAGs identified in **Fig. 1**. After accounting for computer use time in the GAM, the U-shaped relationship pattern between sleep and BAG was robust, although the curve and the "optimal" sleep duration showed slight variations.

To address your comment specific to the causal signal ("Given the interest in this paper in causal inference"), we reperformed the GWAS by including smoking status, Townsend deprivation index, and computer use time as additional covariates, and performed the two-sample Mendelian randomization. Updated results for Manhattan and QQ plots are in **Supplementary Figure 18**:

Supplementary Figure 18: QQ and Manhattan plots of the GWAS results

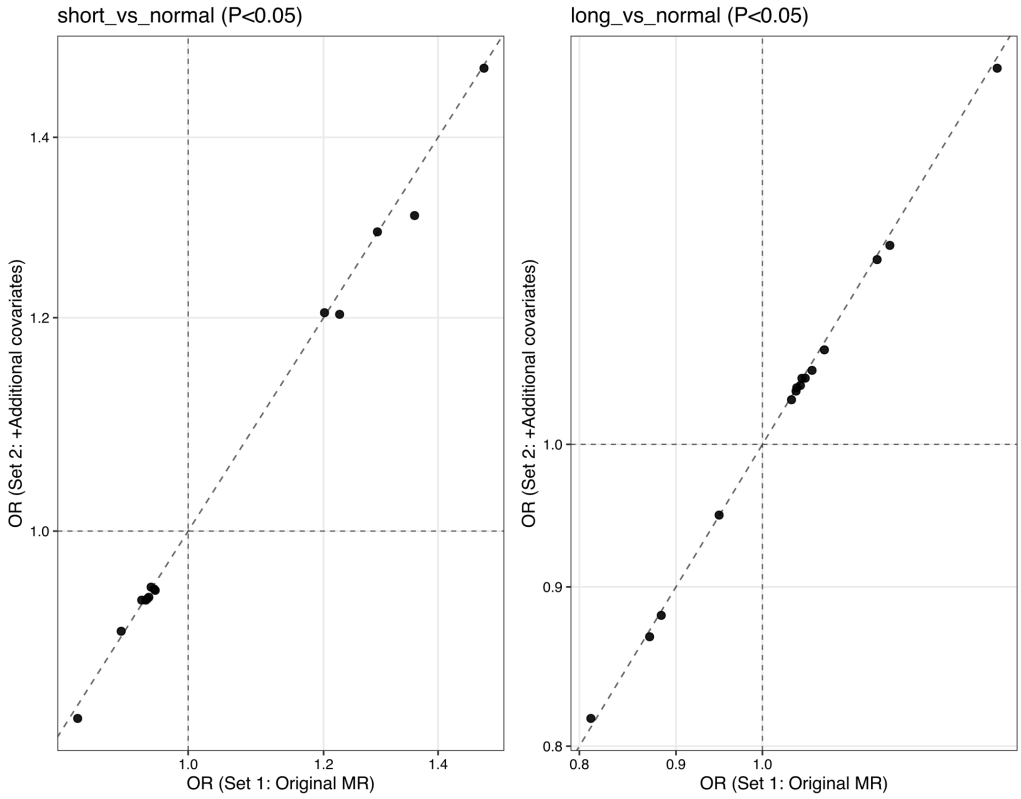
622 West 168th Street, New York, NY 10032



a) We present Manhattan and QQ plots for the two GWAS results: short vs. normal sleep duration and long versus normal sleep duration. The LDSC intercepts were close to 1, suggesting minimal bias from population structure. **b**) Manhattan and QQ plots for the two GWASs by adding additional covariates, including smoke status, Townsend deprivation index, and computer use time, in the original GWASs.

Using the updated GWAS that included additional covariates, we repeated the MR analyses from DE to sleep. Notably, in the original MR analysis, diabetes was the only exposure showing evidence of a causal association with sleep duration, and we did not observe broadly significant findings. To avoid giving readers the mistaken impression of a significant DE–sleep causal relationship, we present this MR comparison only in the response letter. Overall, nominally significant effects ($P < 0.05$) were highly correlated between the two result sets.

Junhao (Hao) Wen, PhD
Department of Radiology



Set 1: original MR using the original GWAS summary statistics. Set 2: new MR using the GWAS summary statistics by adding the 3 additional covariates.

In summary, we agree that unmeasured confounding may still influence the observed sleep–BAG associations. However, based on the extensive sensitivity analyses conducted across two rounds of revision in response to reviewers' comments (**Supplementary Figure 2-16**), we believe the observed U-shaped pattern is robust.

The authors nicely acknowledge the limitations of their methods in their concluding Limitations paragraph: the use of self-reported sleep data, heterogeneity from multiple cohorts, largely "cross-sectional" data and lack of longitudinal sleep and omics/MRI data; studying selected (healthy) samples; no consideration of other important sleep traits that vary with sleep duration. However, these significant limitations are largely not reflected elsewhere in the paper; rather the text often describes findings using strong causal language (influence, shape, impact, etc). The authors should consider more conservatively presenting their findings, providing a more balanced presentation.

We thank you very much for these suggestions.

We believe that, overall, we have been transparent in clearly acknowledging the limitations of our work – both in terms of the data (particularly the cross-sectional nature of the cohorts) and the methods (including MR and SEM). We are careful not to overstate any causal relationships, especially those observed between sleep and disease in our analyses. For example, this is why we explicitly emphasized in the Abstract (**Line: 65**): "Although our Mendelian randomization does not show strong causal effects from disease to sleep disturbances, it does not fully rule out the possibility that sleep disturbances may, in part, reflect underlying disease burden." This tone has been set in several places throughout the paper.

To further address your comment, we have down-toned the wording specifically for "influence, shape, impact, affect" (replacement to "associate, relate to, link, etc.) throughout the paper (yellow-colored text in the revised manuscript).

The authors should consider whether labeling sleep durations as "optimal" may be misleading due to how data were derived (with specific numeric values likely to vary when using a different sleep question, sleep assessment modality, or different sample). A more cautious description may reduce the chance that the results are misinterpreted.

We really appreciate your comment on this!

In the revised manuscript, we made these changes:

- (**Line: 58**): "The lowest BAGs are achieved between 6.4 and 7.8 hours of sleep duration, and vary by organ and sex in the UK Biobank (ages 37–84 years)."
- (**Line: 96**): "Are the observed U-shaped associations and empirically derived sample minimum values of the BAG-sleep relationships consistent across different sexes and organ systems?"
- (**Line: 134**): "We then estimated the sample minimum of the BAG-sleep curves based on the peak of the smoothed spline curves for the brain ProtBAG."

More broadly, we use "sample minimum values for the BAG" to replace the optimal sleep time throughout the paper.

Given the importance of sleep duration measurement—and its variation in measurement across the cohorts examined- with varying influence on measurement error, it is important to be clearer on what was measured in each cohort. For example, the UKB field name was provided but the question used for sleep duration not provided. I think the paper reported data from a question that asked about sleep over a 24 hour period, which includes napping and the main sleep period. If so, the authors should be careful to interpret their data with this phenotype in mind, noting that napping and overnight sleep duration represent different concepts and health associations.

We thank you for your comment. Yes, in the UK Biobank, we used the questionnaire-based sleep duration variable (https://biobank.ndph.ox.ac.uk/ukb/field.cgi?id=1160). In the revised manuscript, we have provided the details (**Line: 543**): "For the primary variable of interest in UKBB, sleep duration (Data-Field: 1160) was assessed using an ACE touchscreen questionnaire asking: "About how many hours sleep do you get in every 24 hours? (please include naps)." Participants entered a numeric value, which underwent basic quality control: responses of less than 1 hour or more than 23 hours were rejected, and values below 3 hours or above 12 hours triggered a confirmation prompt. If participants clicked the Help button, they were instructed that, if their sleep duration varied substantially, they should report the average number of hours slept in a 24-hour day over the past 4 weeks. For this variable, the value -1 indicates "Do not know" and the value -3 indicates "Prefer not to answer, which were excluded in the current work. For the multi-organ IDPs, we used multi-organ MRI data from 7 organ systems and tissues (Category ID: 100003), including the brain, heart, liver, pancreas, spleen, adipose, and kidney. Specifically, the MUSE atlas-derived brain IDPs from the T1-weighted MRI[11] were used for the brain MRIBAG generation. We also used neural networks to analyze the raw cardiac MRI images in our previous study[21] and returned them to the UKBB to extract heart-specific IDPs (Category ID: 157), which were used to derive the heart MRIBAG. For the other organs' IDPs, we used the pre-derived features from the UKBB (Category ID: 105). For the plasma proteomics data, we downloaded the original data (Category ID: 1838), which were analyzed and made

available to the community by the UK Biobank Pharma Proteomics Project (UKB-PPP[73]). The initial quality control procedures were described in the original work[74]; we conducted additional quality-check steps as outlined in **Method 5**. We also imputed missing normalized protein expression (NPX) values and defined the organ-specific proteins using the HPA platform (https://www.proteinatlas.org/), as detailed in our previous work. For the plasma metabolomics data, we downloaded the original data (Category ID: 220), which were analyzed and made available to the community by Nightingale Health Plc in the UKBB. Additional quality check analyses were performed as detailed in **Method 6**. **Supplementary Table 1** provides a detailed list of subsample characteristics corresponding to each data pattern."

We have also revised the explanation in the paper, no longer limiting sleep duration uniquely to the number of hours per night.

Similarly, for the BLSA study, we also provided details regarding the sleep duration variable (**Line: 605**): "For self-reported sleep duration, sleep duration is assessed using a standardized questionnaire item asking: "On average, in the past month, how many hours of sleep did you get each night?" Participants select from ordered categorical response options reflecting typical nightly sleep duration: more than 7 hours, more than 6 up to 7 hours, more than 5 up to 6 hours, or 5 hours or fewer. This measure captures habitual sleep duration over the prior month and represents participants' perceived average nightly sleep. We also included participants who underwent wrist actigraphy. They were instructed to wear an Actiwatch-2 wrist actigraphy (Philips-Respironics, Bend, OR) on their nondominant wrist for seven consecutive 24-hour periods. The devices continuously recorded activity counts and ambient light levels, and participants were asked to press the event marker each time they intended to go to sleep ("lights out") and when they got up to start their day[81]."

For the MESA study (**Line: 622**): "Self-reported sleep duration is obtained from the Exam 5 Sleep Questionnaire, in which participants report their habitual bedtimes and wake times separately for weekdays and weekends. The questionnaire includes items asking for bedtime and waketime in 24-hour format, from which MESA derives weekday sleep duration and weekend sleep duration expressed in hours or hours-and-minutes."

The authors acknowledge the limitations of the mediation analysis due to the sparse longitudinal data and single assessment of sleep and single assessment of the BAG mediator, and limitations of questionnaire and ICD outcome data. Providing further caution in describing those findings and acknowledging the difficulty in knowing the true temporal association between the exposures, mediators, and outcomes would help (while some acknowledgement of the latter is made, the results still are stated very strongly).

We thank you very much for your suggestion. We believe that we have stated in several places the limitations of our analyses regarding data (i.e., cross-sectional) and methods (i.e., underlying statistical assumptions), and did not strongly claim causality (e.g., in the Abstract).

In the Abstract (**Line: 65**): "Although our Mendelian randomization does not show strong causal effects from disease to sleep disturbances, it does not fully rule out the possibility that sleep disturbances may, in part, reflect underlying disease burden."

In the Result section (**Line: 292**): "Our mediation analysis treated sleep duration as a modifiable risk factor, based on the temporal sequence of data collection: sleep data were recorded at baseline, while brain MRI occurred during follow-up. As detailed in **Supplementary Table 6**, we conducted sensitivity analyses by alternately specifying brain MRIBAG and

LLD1/2 as mediators and outcomes. However, due to the limitations of the data's time-ordering, we cannot fully exclude the possibility that sleep disturbances result from underlying disease burden. To test this, we conducted Mendelian randomization analyses using 5 different estimators (**Supplementary Note 5** and **Supplementary File 9**) to test the reverse causal pathway, from 525 DEs using FinnGen and Psychiatric Genomics Consortium (PGC) data, to the 2 binary sleep traits. These analyses did not support a widespread causal effect of disease on sleep disturbances, reinforcing the interpretation of sleep disturbances as potential risk factors. We nevertheless acknowledge the possibility of bidirectional effects. For example, Pasman et al.[39] found bidirectional causality between major depressive disorder and insomnia, but not sleep duration. In addition, some previous studies modeled sleep duration as a continuous trait using linear Mendelian randomization, which does not capture potential non-linear relationships that methods such as fractional polynomial Mendelian randomization can address[40]. To probe this further, we conducted additional sensitivity analyses focused on depression-related endpoints in our MR results. Specifically, F5_DEPRESSION_DYSTHYMIA (number of cases: 48,222) in relation to long sleep duration was not statistically significant (P-value>0.05). The sensitivity analyses results are presented in **Extended Data Figure 6**, including the per-SNP F statistics, horizontal pleiotropy and heterogeneity test based on MR-Egger, and horizontal pleiotropy from MR-PRESSO and fractional polynomial MR."

In the Limitation section (**Line: 502**): "Third, the cross-sectional design of this study limits our ability to determine causality or the direction of effect; although our current analysis positions sleep disturbance as a modifiable risk factor, longitudinal follow-up is needed to clarify whether sleep disturbance is a modifiable risk factor or a consequence of disease burden."

We would appreciate it if you could specify additional changes if you think necessary.

The use of ICD codes for insomnia and hypersomnia as proxies for short and long sleep duration, respectively, simplifies the important differences between these terms. Notably, insomnia is comprised of subtypes with normal and short sleep, and many individuals with short sleep do not meet criteria for insomnia (their short sleep reflects environmental or behavioral factors that curtail their time in bed). Clinical hypersomnia includes disorders of excessive sleepiness- although these sleep disorders often associate with long sleep duration, those sleeping > 8 hours per 24 hour period often will not meet criteria for disorders of hypersomnia. So while these diseases may represent more "extreme" phenotypes (as noted by the authors), they also are distinct from short vs long self-reported sleep duration.

We thank you for this insight, and we agree with you on this.

In the revised manuscript, we avoided using the "independent validation" for the TriNetX data (**Line: 262**): "**Supplementary Note 4**, **Supplementary File 7**, and **Extended Data Figure 5** present results of examining the associations between two sleep disorders (insomnia and hypersomnia) and the DEs, using the TriNetX dataset, identified in **Fig. 3a**, as well as all-cause mortality.", as well as in the Method section (**Line: 602**): "We used this resource to assess associations between insomnia and hypersomnia and systemic disease outcomes across organ systems identified in UKBB (**Fig. 3**)."

We also avoid using the word "Sleep proxy" in the revised manuscript.

Suggestions for improvement:

Please provide more details in the results section (including the footnotes in figures) to allow the

reader to more easily appreciate what sample(s) were analyzed and what variables were adjusted for.

We appreciate this suggestion to enhance reproducibility and transparency.

In response, in the revised manuscript, we have specified the dataset used for each set of analyses in the Results section. Details on all covariates are already fully described in the Methods section; repeating them in the Results would make that section overly dense and redundant, potentially reducing its readability.

- **GAM results (Line: 124)**: "Using GAMs with the UKBB data, we assessed the nonlinear, U-shaped relationship between sleep duration and 23 multi-organ, multi-omics BAGs ($P<0.05/23$; **Fig. 1**)."
- **GWAS (Line: 177)**: "In UKBB, our genome-wide association study (GWAS; **Method 7**) identified 8 genetic loci ($P<5x10^{-8}$) associated with abnormal sleep duration patterns (**Fig. 2a**)."
- **Genetic correlation (Line: 204)**: "Using summary-level data from our GWAS and FinnGen and PGC, genetic correlation analyses revealed 153 positive associations between abnormal sleep duration patterns and systemic diseases across multiple organ systems ($P<0.05/527$), dominated by short sleep duration (**Fig. 2c**)."
- **Survival analysis (Line: 234)**: "Leveraging ICD code-based clinical diagnosis and echoing the genetic correlation results, we further assessed the relationships between abnormal sleep duration patterns and the time to incident disease prediction, as well as all-cause mortality, using a survival analysis framework in UKBB (**Method 8**)."
- **SEM (Line: 275)**: "To evaluate this hypothesis in UKBB, we performed structural equation modeling (SEM) using sleep duration patterns (short or long) as exposures, the 7 MRIBAGs as mediators, and two MRI-derived subtypes of late-life depression (LLD1 and LLD2[8]) as outcomes."
- For all other analyses using data from other studies, we explicitly specified the study and details. This information was also added in the **Method** section. All covariates are also stated for each set of analysis in the **Method** section.

Describe the specific sleep duration questions used for analysis for each cohort; discuss more the implications of these definitions.

Thank you for this excellent suggestion to improve the transparency of our study.

As described in our response to your earlier comment, we have added detailed information in the revised manuscript on how sleep duration, both self-reported and actigraphy-based, was collected in the UK Biobank, BLSA, and MESA (in **Method 1**). Overall, our primary analyses are based on the UK Biobank, while MESA and BLSA were included to assess potential replication of the U-shaped pattern. As anticipated, the sample sizes in these two cohorts are limited, and there are also differences in how questionnaire-based sleep duration was assessed across studies. We anticipated that further biobank-level data will be available to replicate the results in the UK Biobank. Some ongoing efforts include our local Columbia Biobank data (https://www.vagelos.columbia.edu/research/researchers/core-and-shared-facilities/new-instruments-and-facilities/bridge-biobanking-facility/columbia-university-biobank), in which sleep data will also be planned to be included for data collection.

The datasets examined have information on sleep apnea, RLS, and insomnia. It would be very

helpful to know if results varied (including sex-specific findings) with adjustment, stratification, or restriction based on these diseases.

We thank you a lot for this great insight. In our GAM modeling approach, we indeed include disease status as a covariate (based on ICD code or clinical record in the UKB) (**Line: 639**): "We adjusted for key demographic and physiological covariates (i.e., age, sex, weight, height, waist circumference, BMI, assessment center, diastole, systolic blood pressure, time differences for data collection (for MRIBAGs), and disease status)."

One caveat for UKB is that it is a population-based cohort rather than a clinically ascertained sleep-disorder cohort; consequently, the prevalence, severity, and sample sizes (*N*) of sleep apnea (G473), RLS (G258 under Other specified extrapyramidal and movement disorders), and insomnia (G470+F510) are relatively low compared with clinical studies, and case definitions rely on ICD codes and/or health-record ascertainment, which may introduce underdiagnosis and heterogeneous phenotyping. Therefore, stratifying or restricting the sample in these diseased populations for the GAM modeling will be too underpowered to detect this U-shaped pattern (if there is indeed a U-shaped pattern).

To address your comment here, instead, we re-performed our main GAM modeling by excluding the four sleep-related disorders (merging the remaining 491,427 people with the BAG populations, and each BAG's population varies based on the modality). Updated results are in **Supplementary Figure 6**:

TEXT REDACTED

FIGURE REDACTED

Generalized additive models were fit in the UK Biobank after removing individuals with any of the following sleep-related diagnoses (sleep apnea, insomnia, or restless legs syndrome). For each organ-specific BAG, we refit the same model as in **Fig. 1**. Dashed vertical lines indicate the sex-specific sleep duration corresponding to the minimum predicted BAG ("optimal" sleep).

Colored icons correspond to the 9 significant BAGs identified in **Fig. 1**. After excluding participants with three types of sleep disorders, the U-shaped pattern persisted, although the estimated curve and the "optimal" sleep duration changed slightly.

Please provide a description of the UKBB subsamples corresponding to each data modality (MRI, proteomic, metabolomic) in the Methods 1: UK Biobank or

We thank you for this suggestion!

In this revision, we further give more details for transparency (**Line: 551**): "For the multi-organ IDPs, we used multi-organ MRI data from 7 organ systems and tissues (Category ID: 100003), including the brain, heart, liver, pancreas, spleen, adipose, and kidney. Specifically, the MUSE atlas-derived brain IDPs from the T1-weighted MRI[11] were used for the brain MRIBAG generation. We also used neural networks to analyze the raw cardiac MRI images in our previous study[21] and returned them to the UKBB to extract heart-specific IDPs (Category ID: 157), which were used to derive the heart MRIBAG. For the other organs' IDPs, we used the pre-derived features from the UKBB (Category ID: 105). For the plasma proteomics data, we downloaded the original data (Category ID: 1838), which were analyzed and made available to the community by the UK Biobank Pharma Proteomics Project (UKB-PPP[74]). The initial quality control procedures were described in the original work[75]; we conducted additional quality-check steps as outlined in **Method 5**. We also imputed missing normalized protein expression (NPX) values and defined the organ-specific proteins using the HPA platform (https://www.proteinatlas.org/), as detailed in our previous work[12]. For the plasma metabolomics data, we downloaded the original data (Category ID: 220), which were analyzed and made available to the community by Nightingale Health Plc in the UKBB. Additional quality check analyses were performed as detailed in **Method 6**. **Supplementary Table 1** provides a detailed list of subsample characteristics corresponding to each data pattern."

Method 3: 23 Multi-organ aging clocks section for clarity.

As we mentioned in our response to your previous comment, we have provided more detailed information regarding the data partitioning and cross-validation used in the machine learning models of building our biological age clock (**Line: 670**): "In our previous studies[1,2,3], we described in detail the population definition and cross-validation procedures used for model training, which we summarize here. Applying a coherent machine learning framework, we assessed the performance of age-prediction models. Hyperparameter tuning was performed through nested, repeated holdout cross-validation with 50 repetitions (80% training/validation and 20% testing). Specifically, we performed a grid search for fine-tuning model-specific hyperparameters. The within-distribution, holdout test dataset was held out to unbiasedly evaluate model performance (different from the 20% test dataset from the cross-validation).

To rigorously train the age-prediction models, we first defined participants without any pathologies based on ICD code and clinical history as CN. We further split the CN into the following datasets:

- CN within-distribution, holdout test dataset: 500 participants were randomly drawn from the CN population. Within-distribution, holdout test datasets are ideal for objectively evaluating machine learning model performance, especially in studies with large sample sizes, such as the UK Biobank.

- CN training/validation dataset: 80% of the remaining CN population was used for the inner loop 10-fold cross-validation for hyperparameter selection.
- CN cross-validated test dataset: 20% of the remaining CN population was used for the outer-loop 50 repetitions.
- Patient dataset: All patients who have at least one ICD-10-based diagnosis or clinical history.

The CN training/validation/test datasets were used for model development and were employed with a nested cross-validation procedure for all machine learning models (LASSO regression and support vector regressor, elastic net, and neural network), whereas the within-distribution, holdout test set provided an unbiased assessment of model performance. Model evaluation metrics included MAE and Pearson's *r*. Age bias correction was applied using the approach outlined by Beheshti et al.[89]."

It would be helpful to see a justification for the choice of variables used in each model, and minimally perform sensitivity analyses with a more comprehensive set of covariates.

We thank you very much for this comment. As also suggested by other reviewers, in the previous revision and our original manuscript, we indeed provided extensive sensitivity analyses (with your additional suggestion, now they are presented in **Supplementary Figure 1-17**). For example, for the GAM modeling alone, **Supplementary Figure 1-14** provides a set of sensitivity analyses (e.g., sex hormone) suggested by the other reviewers, as well as your additional suggestions. We believe that we have performed extensive sensitivity analyses to ensure the robustness of our modeling, acknowledging that hidden confounding factors might still not be accounted for.

The authors report sex differences in several BAGs, with some effects in opposite directions (i.e., accelerated aging in females based on brain MRI but the opposite pattern for the brain proteomic clock).

We thank you for this observation.

In our main GAM modeling results, indeed, we observed a significant difference in sex for both the brain MRIBAG and brain ProtBAG. In addition, we had explicitly shown the results that directly tested the sex-sleep interaction for the brain ProtBAG (**Line: 131**): "At the population mean level, females exhibited significantly higher brain ProtBAG values than males ($P_2<1\times10^{-20}$). The interaction between sex and sleep duration approached nominal significance ($P_3=0.06$), indicating a potential trend toward sex-specific associations". As detailed in **Supplementary File 1**, for the brain MRIBAG, the sex-sleep interaction's P-value is 0.84. Therefore, using a nominal P-value threshold (P=0.05), the terms for sex-sleep interaction are not statistically significant, strictly speaking.

Of note, the curve we observed in **Fig. 1** is a model-based prediction generated from our GAM modeling. Specifically, we fit a penalized spline for the overall sleep–BAG association. Additionally, we included a sex-specific smooth term along with a parametric main effect of sex. In GAM, these smooth terms are represented using cubic regression spline bases (bs='cr') with complexity controlled by the basis dimension and penalization selected by REML; thus, the plotted lines reflect the estimated smooth functions after shrinkage/regularization rather than unconstrained interpolation. The displayed sex-specific curves, therefore, represent the expected (conditional mean) BAG across sleep duration under the fitted model. Importantly, even when the formal sex×sleep interaction smooth is not statistically significant, the two curves can still

appear visually separated because the sex main effect induces an overall vertical offset, and the penalized smooths can yield modest apparent differences due to smoothing and differences in data density across sleep ranges. Accordingly, we interpret visual divergence cautiously and rely on the interaction term for inferential claims about sex-specific sleep–BAG relationships.

The potential biological interpretation is detailed in the comment you listed below.

It would be helpful if the authors could comment on whether these effects align with previous literature.

For the brain MRIBAG, or brain age in the neuroimaging community, has been established over decades. It is very well established that males show larger brain MRIBAG. I listed some references here:

- https://pmc.ncbi.nlm.nih.gov/articles/PMC3865444/: Search: "Across all subjects, *BrainAGE* scores were higher in males as compared to females ($F = 7.7$; $p = 0.006$)"
- https://www.sciencedirect.com/science/article/pii/S1053811916300404?via%3Dihub: Search: "In addition, there was a main effect of Sex (Beta = 3.398, p = 0.009) indicating that female brains were estimated on average to be 3.4 years younger than male brains."
- This is also supported by the fact that men showed more severe brain atrophy than women using brain MRI: https://www.pnas.org/doi/10.1073/pnas.2510486122: Search: "Men showed decline in a greater number of brain regions, including multiple cortical regions such as the parahippocampal cortex and several subcortical structures in older adults, while women showed a decline in fewer regions with limited cortical changes and greater ventricular expansion."

For the brain ProtBAG, there have been relatively fewer papers because the first major multi-organ proteome-based aging clock was published in a Nature paper published by our team (From Hamilton Oh et al.: https://www.nature.com/articles/s41586-023-06802-1), and followed up with many other ProtBAG papers in the recent 2 years. To the best of our knowledge, no prior work has systematically evaluated sex-specific aging clocks in a truly multi-organ setting, although some of them performed sensitivity analysis for each sex. This paper (https://www.nature.com/articles/s41591-024-03164-7) uses the non-organ-specific ProtBAG to demonstrate that their model is robust across sexes (showing a high Pearson's *r*), but did not explicitly compare the BAG across males and females.

A further caveat is that organ-specific clocks can be sensitive to how "organ-enriched" proteins are defined for model training. For instance, Oh and colleagues primarily relied on GTEx transcriptomic expression to assign proteins to organs; however, mapping from gene expression to circulating proteomics is imperfect, and additional organ-specific regulatory and post-translational processes may not be captured by transcript data alone. Complementary resources such as the Human Protein Atlas (HPA: https://www.proteinatlas.org/; using both gene expression data and proteomics data from various resources) may therefore offer a more comprehensive basis for defining organ-enriched proteins. This has been highlighted in our recent Comment: https://www.nature.com/articles/s43587-025-00928-9#Sec2 (see: Key considerations for aging clock generation and interpretation).

For sex differences using other modalities, like EEG data (https://www.nature.com/articles/s41591-024-03209-x), the author found that "Search: Sex differences were observed in the brain-age gaps when training in the non-LAC and testing in

LAC (Fig. 4a,b). Specifically, female participants in LAC exhibited a greater bias towards older brain age than males (fMRI: P = 0.04; EEG: P = 0.03).” So, we believe that such sex difference divergence across modalities might reflect the underlying biology of the features captured.

In the revised manuscript, we further discussed the underlying biology for these sex-specific patterns (**Line: 364**): “Several organs displayed significant BAG differences, BAG-by-sex interactions, divergence in the population mean, and sample minimum values of sleep time. For example, males exhibited higher brain MRIBAG, whereas females exhibited higher brain ProtBAG between sexes. This divergence likely reflects that these clocks index distinct layers of biology. The brain MRIBAG is driven by macro-structural MRI features (i.e., regional volumes) that are shaped by lifelong neurodevelopmental trajectories, hormonal, and sex-differential factors that have been repeatedly linked to sex differences in brain aging and neurodegeneration. In contrast, brain ProtBAG is derived from circulating brain-enriched proteins, which may be more sensitive to systemic inflammatory/immune signaling, endocrine regulation, blood–brain barrier permeability, and glia- and vasculature-related processes that influence the release, transport, or clearance of brain-linked proteins in plasma.”

Do these findings suggest that BAG estimation might benefit from sex-specific models? Could this also help explain differences in optimal sleep duration between males and females?

We thank you for this suggestion! Sex differences have been prominent in biological aging clock research, both evidenced from our own studies (listed below, which were trained on sex-pooled data from both males and females) and others.

- 11 ProtBAG: https://www.nature.com/articles/s43587-025-00928-9#Sec2 (see **Supplementary Note 2: Sex-stratified brain ProtBAG**)
- 5 MetBAG: https://www.nature.com/articles/s41467-025-59964-z (see **Supplementary Note 2**: Sex-stratified analyses for the metabolic MetBAG)
- 7 MRIBAG: https://www.nature.com/articles/s41591-025-03999-8 (see **Supplementary Figure 6**: Sex difference for the 7 MRIBAGs)

During peer review of our prior 3 studies (listed above) on 23 biological aging clocks, we observed signals suggestive of sex differences, but we did not examine them systematically. We did not fully evaluate key methodological choices such as whether age-prediction models should be trained in a sex-stratified versus sex-pooled manner (and how the fitted model behaves when applied across sexes), nor did we comprehensively compare interpretation strategies, including explicit sex–exposure interaction terms versus fully sex-stratified analyses, neither for downstream associations (e.g., to disease endpoints or cognition).

Building on evidence from the current study, our prior work, and the broader literature, we conclude that developing sex-specific (or explicitly sex-stratified) aging clocks is necessary to more fully capture the biology underlying human aging and disease.

Along these lines, we are preparing a manuscript that systematically examines sex-specific biological aging clocks, including their links to the X chromosome.

In the age-stratified analysis (Supplementary Figure 2), it appears that the optimal sleep duration differs between midlife and older adults, with older individuals showing shorter optimal sleep.

We thank you for this excellent observation!

Indeed, we agree that certain BAGs showed shorter “optimal” sleep duration in older age ranges than younger age ranges (e.g., brain ProtBAG). This pattern is plausible and consistent

with well-established age-related changes in sleep physiology and sleep measurement. With aging, total sleep time often decreases and sleep becomes more fragmented, with reductions in slow-wave sleep and changes in circadian regulation; therefore, the sleep duration associated with the most favorable biological-aging profile may shift toward slightly shorter values in older adults. In addition, "optimal" sleep in our analysis is defined empirically as the sleep duration that minimizes the model-predicted BAG within each age stratum, conditional on covariates, rather than as a normative clinical recommendation. Consequently, it can also reflect age-dependent differences in comorbidity burden, medication use, and the distribution of sleep duration itself (e.g., fewer older participants reporting long sleep, and greater heterogeneity in sleep quality), all of which can influence the estimated smooth function and its minimum. We therefore interpret the shorter "optimal" sleep observed in older individuals as a data-driven shift that likely captures a combination of genuine physiological aging of sleep and age-dependent health context, and we report it as a sensitivity/stratified finding – this is also in line with what you have suggested regarding the terminology "optimal".

The age-related pattern could be presented in parallel with the sex-related differences in optimal sleep duration. Has this pattern been formally tested- were age*sex interactions included?
Thank you a lot for this suggestion!

To address your comment, in the revised manuscript, we now added a new sensitivity analysis by including age-sex interaction term. The updated results are in **Supplementary Figure 5**.

TEXT REDACTED

FIGURE REDACTED

Generalized additive models were fit in the UK Biobank after adding the age-sex interaction term in the original GAM model. Dashed vertical lines indicate the sex-specific sleep duration corresponding to the minimum predicted BAG ("optimal" sleep). Colored icons correspond to the 9 significant BAGs identified in **Fig. 1**. The age–sex interaction term was significant for several BAGs, consistent with the age-stratified results shown in **Supplementary Fig. 4**, where the estimated "optimal" sleep duration differs between the two age groups. Nevertheless, the overall U-shaped association remains, with only minor shifts in the curves. Sex differences in the sleep–BAG associations may also be influenced by sex-specific reporting/measurement differences in self-reported sleep duration, as well as by other sleep exposures with marked sex disparities, such as higher prevalence of insomnia/restless legs syndrome in females and sleep apnea in males, which could confound or modify the observed relationships.

As we can see, the U-shaped pattern persisted, although minor changes exist regarding the shape of the curve and the 'optimal' sleep duration time.

Please consider other reasons for sex differences, such as differences in reporting of sleep by males/females; differences in other sleep exposures that differ by sex (insomnia/RLS is more common in females and sleep apnea in males).

We agree that the observed sex differences may also reflect sex-specific measurement and exposure profiles. In particular, self-reported sleep duration can differ systematically by sex, and the prevalence of other sleep disorders/exposures that are sex-biased (e.g., insomnia and RLS are more common in females; sleep apnea is more common in males, as you noted here) could confound or modify sleep–BAG associations.

Of note, we adjusted for disease status as a covariate in our primary GAM models and additionally conducted sensitivity analyses restricting to participants without these recorded sleep disorders (**Supplementary Figure 6**). We also added the potential sex-difference reason you suggested in the figure caption in **Supplementary Figure** 5: "Sex differences in the sleep–BAG associations may also be influenced by sex-specific reporting/measurement differences in self-reported sleep duration, as well as by other sleep exposures with marked sex disparities, such as higher prevalence of insomnia/restless legs syndrome in females and sleep apnea in males, which could confound or modify the observed relationships."

The BAG estimates across clocks were normalized for the analysis but providing information or a reference on the biological age prediction performance for each modality would be helpful. Among all clocks used, which one showed the best age prediction accuracy?

We thank you for raising this important question. We have added details in **Method 3: 23 Multi-organ aging clocks** for data splits and cross-validation. We present the modeling details on which population the age-prediction model was derived (i.e., the CN training/validation dataset, which is the normative/reference population of these clocks):

- CN training/validation dataset: 80% of the remaining CN population was used for the inner loop 10-fold cross-validation for hyperparameter selection.
- CN cross-validated test dataset: 20% of the remaining CN population was used for the outer-loop 50 repetitions.

This ensures that the 23 BAGs were derived in a coherent ML framework within a healthy population. However, we have to mention that the imaging, proteomics, and metabolomics populations in the UK Biobank do not entirely overlap.

Junhao (Hao) Wen, PhD
Department of Radiology

Regarding your second question: "Among all clocks used, which one showed the best age prediction accuracy?" In our previous study, we benchmarked the performance of these clocks (as I mentioned above, the training population might differ from ProtBAG, MRIBAG, and MetBAG by design as they do not cover the full UKB population):

- 11 ProtBAG: https://www.nature.com/articles/s43587-025-00928-9#Sec2 (see **Fig. 1a**)
- 5 MetBAG: https://www.nature.com/articles/s41467-025-59964-z (see **Fig. 1b**)
- 7 MRIBAG: https://www.nature.com/articles/s41591-025-03999-8 (see **Fig. 1b**)

One critical question in biological aging clock research: "***What is the goal/aim of developing such a clock?***"

In our recent Comment (https://www.nature.com/articles/s43587-025-00928-9), we highlighted that developing an age prediction model is not to achieve a state-of-the-art model performance for predicting chronological age per se (i.e., 'perfect clock'). Instead, the goal of an age-prediction model is to use these clocks as a biomarker to predict other clinical outcomes, e.g., cognitive decline (see **Fig. 2e-d** in https://www.nature.com/articles/s43587-025-00928-9). or disease endpoints (see **Fig. 3** in https://academic.oup.com/brain/article/143/7/2312/5863667?login=false).

This has also been independently demonstrated in a newly published Comment (11 December 2025) in Nature Aging: https://www.nature.com/articles/s43587-025-01038-2/figures/1, where the authors state that a 'perfect clock' (MAE ~1-2 years, Pearson's r ~1 between predicted age vs. chronological age) showed no additional prediction power for clinical outcomes.

This reflects a common machine-learning phenomenon: the tradeoff between within-domain fit and cross-domain prediction power, and accurately predicting chronological age (including variance from model error/noise and variance in true underlying biology) is not the ultimate goal.

In our 3 previous papers, we compared the prediction power for cross-domain variables (e.g., cognition or disease endpoints) of the different aging clocks (one additional study for 9 PhenoBAGs: https://www.nature.com/articles/s43587-024-00662-8):

- 11 ProtBAG: https://www.nature.com/articles/s43587-025-00928-9#Sec2 (see **Fig. 5,** where we compared the 11 ProtBAG vs. 9 PhenoBAG for predicting disease categories and mortality, although the populations do not fully overlap):
- 5 MetBAG: https://www.nature.com/articles/s41467-025-59964-z (see **Supplementary Figure 11**: **23 multi-organ, multi-omics BAGs provide incremental R^2 on top of calendar/chronological age for prediction of all-cause mortality and respiratory disease categories**, where we evaluated the additional contributions (incremental R^2) of the 7 PhenoBAG (excluding brain and heart due to smaller sample sizes after merging the samples), 11 ProtBAGs (there was a typo in that figure), 5 MetBAGs, to predict disease category of respiratory diseases and all-cause mortality; copy here for your direct reference):

TEXT REDACTED

FIGURE REDACTED

For the analysis of disease risk based on ICD codes, it would be important to clarify in the results section that this analysis was performed in a dataset independent of the main analytic sample (UKBB).

We thank you for this comment and agree that validation in an exact independent dataset is important.

However, to our best knowledge, there is currently no other biobank-scale resource comparable to UK Biobank in both sample size and depth of phenotyping and genotyping. For this reason, we used TriNetX to provide a complementary, partial "validation" of the multi-organ involvement of sleep-related traits (here, insomnia and hypersomnia). In the revised manuscript, we have explicitly stated the dataset used for each analysis.

In addition, including case–control sample sizes (e.g., as additional labels in Figure 3a) would improve clarity.

We thank you for your suggestion. Due to readability and heavy information in the figure, the other reviewers suggested putting this information, including other detailed statistics, in Supplementary Files. **Supplementary File 6-7** includes the sample sizes information for the UKBB and TriNetX survival analyses.

To address your comment, in the revised manuscript, we now added the sample size (N_case and N_control) information to **Fig. 3** and **Extended Data Figure 5**.

TEXT REDACTED

Junhao (Hao) Wen, PhD
Department of Radiology

FIGURE REDACTED

a) Using normal sleep duration, [6-8] hours, as the reference, we treated sleep duration patterns as categorical variables ([4–6) hours for short sleep, and (8–10] hours for long sleep) to estimate

622 West 168th Street, New York, NY 10032

their association with 726 comorbidity-free disease endpoints defined by ICD-10 codes (limited to diseases with more than 50 cases). After applying Bonferroni correction for multiple comparisons ($P < 0.05/726$), we highlight the significant associations and annotate representative systemic diseases across various organ systems. **b**) Short and long sleep duration are associated with a higher risk of all-cause mortality. An online interactive webpage is available at https://labs-laboratory.com/sleepchart/sleep_cox.html to facilitate visualization.

TEXT REDACTED

FIGURE REDACTED

Using insomnia and hypersomnia as proxies for short and long sleep duration, respectively, we replicated the significant associations shown in **Fig. 5** within the UK Biobank cohort. Disease

endpoints that remained significant after Bonferroni correction (0.05/111, accounting for 110 comorbidity-free disease endpoints and all-cause mortality) are indicated by colored shapes, while non-significant associations are shown in gray. One additional disease category was also tested for all types of dementia. Ctrl: health control.

Please consider the concern that insomnia and hypersomnia are not proxies for short and long sleep as defined here.
We have addressed this from your previous comment. Thank you again for your insightful comment on this!

Please provide the rationale for re-running the GWAS analyses for short and long sleep. Previous studies (PMID: 37770476 and PMID: 30846698), which included UKBB data, identified more genome-wide significant loci than reported here. Please clarify how this analysis differs from or extends those prior efforts and explain the differences in genetic signals reported.
We thank you for raising this comment and suggestion on these pioneering sleep GWAS studies! We have cited one of the studies in the original paper, and now cite both studies in the revised manuscript.

There are minor differences in our sleep GWAS compared to previous literature. We ran our own GWASs for practical considerations (differences in bin definition for short/normal/long sleep duration and GAM modeling for outliers, etc.), but we did not claim that our GWAS is "better" or "optimal". We detail the rationale of our choice below, state the differences, and explain why the two studies you cited resulted in finding more loci (potentially due to larger sample sizes for the power issue).

We chose to perform our own GWAS for both practical and methodological reasons. Practically, it allowed us to implement consistent quality control and, importantly, to obtain precise information on linkage disequilibrium, as some prior GWAS do not publicly share allele frequency data and other critical information (e.g., effect size types and standard error, etc.). Methodologically, key design choices, such as how we define short/normal/long sleep, whether we treat sleep duration as a categorical versus continuous phenotype, and how we handle outliers (in our case, excluding sleep duration <4 hours or >10 hours), as well as the GWAS software/pipeline choices, can substantially affect all downstream analyses, and we wanted these decisions to be fully aligned with the aims and structure of our study and hypothesis/assumptions of our statistical modeling approaches (e.g., GAM).

We detail **the differences** between our study and the two you have mentioned here (PMID: 37770476 from Austin-Zimmerman, which was cited in our manuscript, and PMID: 30846698 from Dashti, we cited this in this revision; we knew this study, but we did not cite it because it also used sleep duration as a continuous phenotype in GWAS; although short and long sleep duration GWASs were also performed):

- **Dataset involved**:
  - Our study: We focused on only the UK Biobank for GWAS.
  - Austin-Zimmerman: They included multiple datasets from the UK Biobank and MVP cohorts for a meta-analysis, resulting in higher statistical power with larger sample sizes.
  - Dashti: They also used UK Biobank data only for discovery; their primary GWAS took sleep duration as a continuous variable, as well as taking it as a categorical

variable for short vs. normal and long vs. normal sleep duration. Additional studies, like CHARGE, were used as a replication set.

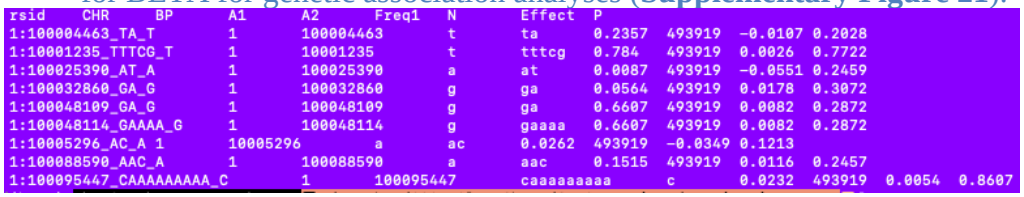
As we can see, the study from Austin-Zimmerman had the largest sample size because it included the additional MVP cohort for a meta-analysis.

- **Inclusion criteria**:
  - Our study: We limited the sleep duration to people between 4 and 10 hours [4-10, including 4 and 10 hours] to exclude the outlier influence of excessively long or short sleep duration on modeling methods such as GAM, which is very sensitive to outliers. We classified sleep duration into short (<6 hours), long (>8 hours), and normal ([6–8] hours) categories for all analyses. We also performed a sensitivity analysis for time to incident disease prediction using [7-9] hours as normal sleep duration (**Supplementary Figure 22**), as suggested by the National Sleep Foundation for young/older adults.
  - Austin-Zimmerman: They defined 'short' sleep duration as ≤5 h sleep, 'normal' as 7–8 h, and 'long' as ≥10 h; also, the criteria for questionnaire-based sleep data between UK Biobank and MVP slightly differed.
  - Dashti: They categorized sleep duration as either short (6 h or less: (-, 7) hours), normal (7 or 8 h: [7, 9) hours), or long (9 h or more: [9, +) hours) sleep duration. Extreme responses of less than 3 h or more than 18 h were excluded – this is less stringent than our excluding outlier criterion, resulting in a larger sample size.

As we can see, in our study, we used different binning schemes to define short, normal, and long sleep, and applied a more stringent inclusion range: [4–10] hours, which led to the smallest sample sizes among the 3 studies (short vs. normal: 16,872 short and 300,420 normal; long vs. normal: 25,049 long and 300,420 normal – we also removed outliers during preprocessing based on the BAG). This choice was motivated by the fact that GAM models are highly sensitive to individuals with extreme sleep durations at both tails, which can introduce spurious associations and bias. We believe that all these different choices might influence the GWAS signals identified in these studies.

In the revised manuscript, we downloaded the GWAS summary statistics from the 2 studies and performed comparison analyses (**Supplementary Note 3** & **Supplementary Figure 19-21**).

- **GWAS summary statistics for disseminations**:
  - Our study: we made our sleep GWAS summary data publicly available at our SleepChart portal: e.g., https://labs-laboratory.com/sleepchart/gwas_short.
  - Austin-Zimmerman: We downloaded the GWAS summary statistics from the portal: https://medicine.yale.edu/lab/gelernter/stats/. However, the summary statistics have a column called "Effect" (no s.e. information is provided), which does not specify whether this is Beta, OR, or a Z score. We assumed this column for BETA for genetic association analyses (**Supplementary Figure 21**).



rsid	CHR	BP	A1	A2	Freq1	N	Effect	P				
1:100004463_TA_T			1	100004463		t	ta	0.2357	493919	-0.0107	0.2028	
1:10001235_TTTCG_T			1	10001235		t	tttcg	0.784	493919	0.0026	0.7722	
1:100025390_AT_A			1	100025390		a	at	0.0087	493919	-0.0551	0.2459	
1:100032860_GA_G			1	100032860		g	ga	0.0564	493919	0.0178	0.3072	
1:100048109_GA_G			1	100048109		g	ga	0.6607	493919	0.0082	0.2872	
1:100048114_GAAAA_G			1	100048114		g	gaaaa	0.6607	493919	0.0082	0.2872	
1:10005296_AC_A 1			10005296		a	ac	0.0262	493919	-0.0349	0.1213		
1:100088590_AAC_A			1	100088590		a	aac	0.1515	493919	0.0116	0.2457	
1:100095447_CAAAAAAAA_C			1	100095447		caaaaaaaaa	c	0.0232	493919	0.0054	0.8607	

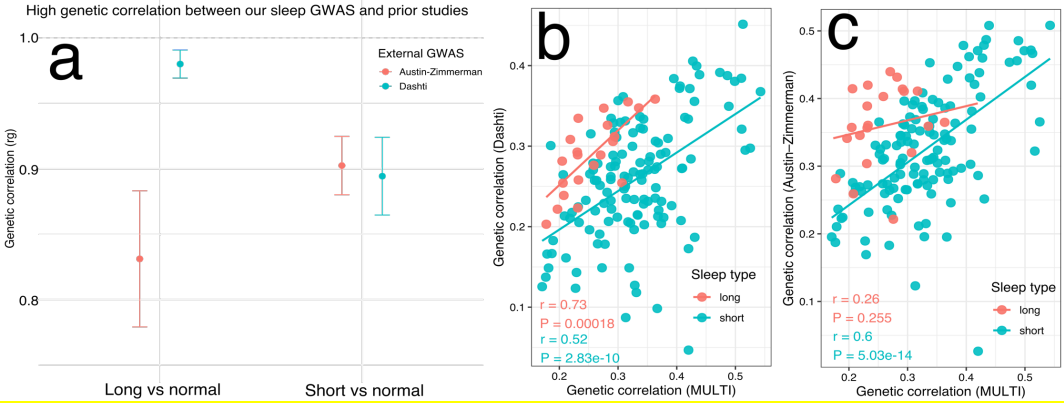
  - Dashti: We downloaded the GWAS summary statistics from https://sleep.hugeamp.org/dinspector.html?dataset=GWAS_UKBB_eu&phenotyp

e=SleepDuration. These summary statistics do not have the sample size (*N*) column, and we found the global sample size for each bin (short/normal/long) in their Supplementary Table 1: short=106,192; normal=305742; and long=34,184. Although this is not optimal, as effective sample sizes differ across individual SNPs.

We used the downloaded GWAS summary statistics from the two studies for genetic correlation analyses and compared the results with our sleep GWASs. Specifically, we estimated pairwise genetic correlations between these two GWAS and our own GWAS of short vs. normal and long vs. normal sleep duration.

- We observed very strong genetic correlations ($0.83<r_g<0.98$; **Supplementary Figure 19a**), indicating that the underlying genetic architecture between the 2 studies and our study is highly similar. Thus, the differences in the number of genome-wide significant loci are likely driven mainly by differences in statistical power and sample size, as well as by differences in how short, normal, and long sleep duration were defined.
- In addition, we performed genetic correlation between the GWAS from the two studies with the 525 FinnGen GWAS, and compared the significant results observed in our GWAS in Fig. 2c and found high concordance (**Supplementary Figure 19b-c**, **20**, and **21**).

Supplementary Figure 19: Genetic correlation between our binary sleep GWAS with two previous GWASs from Dashti et al and Austin-Zimmerman et al



a) We downloaded GWAS summary statistics from Dashti et al. and Austin-Zimmerman et al. and used LDSC to estimate pairwise genetic correlations between our GWAS and each of these studies for short vs. normal and long vs. normal sleep. We observed consistently high genetic correlations, indicating that differences in the number of genomic loci across studies are largely driven by differences in sample size and in the operational definitions of short, normal, and long sleep duration. **b**) Using FinnGen, we demonstrate that the significant DE-sleep signals we found in our GWAS correlate with signals in the sleep GWAS summary statistics of Dashti et al. **c**) Using FinnGen, we demonstrate that the significant DE-sleep signals we found in our GWAS correlate with signals in the sleep GWAS summary statistics of Austin-Zimmerman et al.

TEXT REDACTED

FIGURE REDACTED

Genetic correlations (LDSC; mean ± s.e.) were shown between short sleep vs. normal sleep (triangles) and long sleep vs. normal sleep (circles) and 309 disease endpoints (DEs) from FinnGen (P-value<0.05/527). Points are colored by organ system and ordered on the y-axis from the brain (top) to other (bottom) to reflect approximate human anatomy. Horizontal error bars

denote ±1 s.e., and the dashed red-vertical line indicates zero genetic correlation. For interpretability, the disease names shown as text annotations correspond to the top 10 representative DEs within each organ category (ranked by the smallest available p-value for the sleep–DE genetic correlation; otherwise by the largest absolute genetic correlation). **Supplementary File 5b** presents detailed statistics for our genetic correlation analysis.

TEXT REDACTED

FIGURE REDACTED

Genetic correlations (LDSC; mean ± s.e.) were shown between short sleep vs. normal sleep (triangles) and long sleep vs. normal sleep (circles) and 361 disease endpoints (DEs) from

FinnGen (P-value<0.05/527). Points are colored by organ system and ordered on the y-axis from the brain (top) to other (bottom) to reflect approximate human anatomy. Horizontal error bars denote ±1 s.e., and the dashed red-vertical line indicates zero genetic correlation. For interpretability, the disease names shown as text annotations correspond to the top 10 representative DEs within each organ category (ranked by the smallest available p-value for the sleep–DE genetic correlation; otherwise by the largest absolute genetic correlation). **Supplementary File 5c** presents detailed statistics for our genetic correlation analysis.

Genetic correlations for sleep duration also have been reported in multiple prior reports- please compare to literature and justify the novelty here (PMID: 30846698, PMID: 37770476, PMID: 38388528, PMID: 34482950, PMID: 39856408, etc).

We appreciate the additional references!

We were already aware of these studies and had cited some of them in earlier versions of the manuscript. In contrast to our genetic correlation analyses, we view these prior works as largely candidate and hypothesis-based approaches that i) focus on selected disease domains (e.g., psychiatric disorders) and ii) typically prioritize the single most highly powered GWAS for each disease. By comparison, our genetic analyses leverage large biobank resources (such as FinnGen) in a non-selective, data-driven manner and emphasize a multi-organ perspective (disease endpoints of their primary organ system). We provide detailed genetic correlation results in relation to each of these studies below.

- https://www.nature.com/articles/s41467-019-08917-4: They performed genetic correlation between sleep traits (including short and long sleep duration) with 45 selected traits spanning Cognition and Psychiatric, Anthropometrics, Glycemia and Cardiometabolic, Other Diseases, and Others (Supplementary Data 18 in their paper)
- https://www.nature.com/articles/s41467-023-41249-y: They used LDSC to assess the genetic correlation to traits based upon GWAS summary statistics downloaded from the PGC website and from the Sleep Disorder Knowledge Portal (http://www.kp4cd.org/dataset_downloads/sleep), which mainly targets brain mental conditions and sleep-related disorders (results in Supplementary Data 31 in their paper).
- https://pubmed.ncbi.nlm.nih.gov/34482950/: They pre-selected bipolar disorder (BIP) and schizophrenia (SCZ) traits using GWAS results from PGC, and obtained GWAS data for insomnia, chronotype, and sleep duration from publicly available UK Biobank–based summary statistics (as shown in their Figure 1).
- https://www.nature.com/articles/s42003-025-07514-0: They performed a genetic correlation association analysis by linking 6 Sleep Health Scores (SHS) with 375 phenotypes from the UK Biobank itself: "We estimated genetic correlations among SHS and with other self-reported and accelerometry sleep traits using LDSC44 and genome-wide SNPs mapped to the HapMap3 reference panel. To understand the genetic overlap with a range of common health problems, we selected 381 representative UKB traits by choosing the 232 most heritable traits using UKB hierarchical phenotype categories (selecting at most one 'level' per phenotype, at most 5 phenotypes per phenotype category, and at most 25 phenotypes per category group), as well as all 195 (partly overlapping) traits in the PanUKBB maximally independent set of phenotypes (https://pan.ukbb.broadinstitute.org/blog/2022/04/11/h2-qc-updated-sumstats/index.html)."

Therefore, we believe that our genetic correlation analyses are slightly different in that we leverage a different biobank-scale resource (FinnGen) and pursue a broader, data-driven hypothesis – that sleep disturbances are genetically linked to diseases across brain-body systems.

To fully address your two comments below, in the revised manuscript, we have these changes:

In the main manuscript (**Line: 231**): "**Supplementary Note 3**, **Supplementary Figure 19-21**, and **Supplementary File 5b-c** present a discussion comparing our GWAS and genetic relevance with previous studies[33,34,35,36,37]."

Decisions/limitations on data included: The use of specific sources of data were unclear. For example, regarding the discussion on objective vs. subjective sleep duration, both MESA and BLSA have objective sleep measures (and MESA has larger samples for sleep data than reported here.

We thank you for this comment. We are aware that the BLSA includes both questionnaire-based and actigraphy-based measures of sleep duration.

- Through our close collaboration with the BLSA investigators at NIA, sleep data were shared specifically for participants overlapping with the brain MRIBAG derived from neuroimaging on our side; consequently, the analyzable sample is limited (*N*=385).
- For MESA, similarly, we have access to questionnaire-based sleep duration, and after merging with the brain MRIBAG neuroimaging sample, the final analytic sample size was 583.

In the revised manuscript, we have added a detailed description of sleep duration ascertainment across the three cohorts with individual-level data (**Methods 1**). However, these two studies' sample sizes, after merging with the neuroimaging population, are very limited regarding statistical power.

Have the authors considered using an actigraphy-based sleep duration PRS (from PMID: 30846698) as a proxy for objective sleep duration? This could be estimated in the UKBB to test whether the observed U-shaped trajectory replicates.

Thank you for this suggestion. Although we included actigraphy-based sleep duration from the BLSA, the sample size is relatively small.

We are in the process of requesting actigraphy data from the UK Biobank; however, obtaining and processing these data will take additional time and specialized expertise. Moreover, as noted by other reviewers in the previous revision round, questionnaire-based and objective sleep duration capture different aspects of sleep, so we do not necessarily expect the U-shaped pattern to replicate exactly. They overlap to a certain extent, but they're not measuring the same aspect of sleep biology. Such discordance has been largely proven in previous literature, and we have discussed these in the revised manuscript (**Line: 103**): "Self-reported measures are less objective than actigraphy or polysomnography, capturing different yet complementary aspects of sleep biology, with only moderate correlations between modalities[25,26]."

In **Extended Data Figure 4**, we indeed linked the questionnaire-based sleep duration with the PRS of the 23 BAGs, demonstrating that such U-shaped patterns are not evident for PRS-BAG, although our PRS only accounts for common genetic variants (particularly supporting that such U-shaped patterns are not largely driven by underlying genetics).

Junhao (Hao) Wen, PhD
Department of Radiology

For these reasons and given the fact that we have included actigraphy data from BLSA and modeled the relationship using BAG-PRSs, we believe a comprehensive actigraphy-based analysis is better suited for future work.

On p. 11 lines 287-288, the hazard ratios for all-cause mortality associated with sleep duration do not match the results shown in Figure 3b (particularly for long sleep: 1.4 in the text vs. 1.1 in the figure). Please verify and correct this discrepancy.
Thank you for catching this typo, and we apologize for the oversight. The statistics reported in the text are correct. We have now corrected **Fig. 3b**, and all detailed results, including sample size, p-values, and hazard ratios, are provided in **Supplementary File 6**.

To address the possibility that differences in the estimates for optimal sleep vary across different clocks could be attributable to sample differences, it would be helpful to conduct sensitivity analyses that restrict analyses to participants with overlapping MRI, proteomic, and metabolomic data. If the differences in optimal sleep duration persist in this subsample, it would suggest that the effect is not primarily driven by sample composition.
Thank you for this insightful comment. We did initially consider this approach. However, a practical limitation in the UK Biobank is data availability: the overlap between participants with imaging, proteomics, and metabolomics data is quite small. As a result, the shared sample across these modalities is very limited (*N*=800). Even for the MRI populations, it is known that the brain, heart, and abdominal MRI populations do not largely overlap with each other. Please see the screenshot below showing the overlap after merging the 23 BAG datasets with the sleep-duration cohort in UKB:

```
> bag_of_interest = bag[, c('participant_id', 'Reproductive_female_ProtBAG', 'Pulmonary_ProtBAG', 'Heart_ProtBAG',
+                 'Brain_ProtBAG', 'Eye_ProtBAG', 'Hepatic_ProtBAG', 'Renal_ProtBAG',
+                 'Reproductive_male_ProtBAG', 'Endocrine_ProtBAG', 'Immune_ProtBAG', 'Skin_ProtBAG',
+                 "Endocrine_MetBAG", "Digestive_MetBAG", "Hepatic_MetBAG", "Immune_MetBAG",
+                 "Metabolic_MetBAG", "Brain_MRIBAG", "Adipose_MRIBAG", "Kidney_MRIBAG", "Heart_MRIBAG", "Liver_MRIBAG",
+                 "Pancreas_MRIBAG", "Spleen_MRIBAG"), drop = FALSE]
> merged_clean <- inner_join(bag_of_interest, sleep_of_interest, by = "participant_id") %>%
+     tidyr::drop_na()
> dim(merged_clean)
[1] 800  25
```

Apparently, due to the subjective nature of questionnaire-based sleep duration measures and the inherent noise, *N*=800 is not statistically powerful to detect this U-shaped pattern. With the rapid advance of the UK Biobank data collection, we anticipate that this will be tested in the near future when the entire 500K populations cover all the modalities.

Minor comments:
Missing reference for MetGAB estimation (assuming this should be reference #14).
We believe that we have cited the reference #14 at its first occurrence in the Introduction (**Line: 80**): "In parallel, the field of aging research has seen rapid progress through the development of biological aging clocks derived from imaging and multi-omics data, like magnetic resonance imaging (MRIBAG[1,13]), plasma proteomics (ProtBAG[2,14]), and metabolomics (MetBAG[3])." We believe that you might refer to the "**Method 3: 23 Multi-organ aging clocks**" section. We have now added all the 3 references for our previous aging clock papers.

In addition, as suggested by the editorial guidance, we now also cited the 3 previous studies for the 23 BAGs in the Abstract (**Line: 53**): "Optimal sleep plays a vital role in

promoting healthy aging and enhancing longevity. This study proposes a Sleep Chart to assess the relationship between self-reported sleep duration and 23 biological aging clocks across 17 organ systems or tissues and 3 omics data types (imaging[1], proteomics[2], and metabolomics[3]).”

Please list which six brain disease GWAS summary statistics were obtained from PGC.
We thank you for this suggestion. In the revised manuscript, we have added a **Supplementary Table 7** to detail the 6 PGC GWAS summary statistics, as well as the 521 FinnGen DEs.

The abbreviations “DE” and “IDP” are defined too late in the text (DE first mentioned on p. 22 but explained on p. 27; IDP first mentioned on p. 18 but defined in the legend of Extended Data Figure 1). Please, defining them upon first use.
Thank you! We have addressed it accordingly!

- **Line 112**: “Secondly, we assessed whether abnormal sleep duration patterns (i.e., short or long sleep duration) were adversely associated with all-cause mortality and systemic disease endpoints (DE) beyond the brain (**Fig. 2-3**).”
- **Line 155**: “We also observed this U-shaped pattern between sleep duration and 720 organ-specific imaging-derived phenotypes (IDP; **Extended Data Figure 1**)”

Typo in the “Ethnics” statement.
This has been corrected. Thank you for the careful proofreading. For this revision, we also had multiple co-authors proofread the main manuscript and supplementary materials to check for typos.

Thank you for your careful review and for the additional constructive comments that have helped strengthen our work!

Junhao (Hao) Wen, PhD

Department of Radiology, Computer Science, and Biomedical Engineering

We have addressed all the comments from the four reviewers and have no issues remaining.

622 West 168th Street, New York, NY 10032