Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Beyond the Hazard Ratio: Development of a Synthesis Readiness Framework for Effect-Scale Heterogeneity in Continuous Biomarker Evidence

Budi Utomo

budiutomo@fk.unair.ac.id

Airlangga University

Dian Ardiana

Universitas Hang Tuah

Shafira Meidyana

Airlangga University

Research Article

Keywords: Evidence synthesis, continuous biomarkers, effect-scale heterogeneity, meta-analysis methodology, metric harmonization, numerical demonstration, synthesis readiness

Posted Date: November 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11020044/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Quantitative synthesis of studies of continuous biomarkers assumes that reported effect estimates target sufficiently comparable estimands. However, differences in exposure scaling, cohort-specific variability, biomarker derivation, and reporting of distributional parameters may create effect-scale heterogeneity. Such heterogeneity can produce pooled estimates that are statistically precise but difficult to interpret on a common biological scale. We developed a proof-of-concept Synthesis Readiness Framework and applied it to associations between GrimAge-related measures and all-cause mortality.

Methods

We conducted a systematic review of prospective cohort studies, registered prospectively in PROSPERO (CRD420261449280) and reported in accordance with PRISMA 2020. We conceptualised heterogeneity in three layers: biological/demographic, analytical/methodological, and metric/effect-scale heterogeneity. We operationalised the Synthesis Readiness Index (SRI), a four-domain tool assessing estimand alignment, metric comparability, distributional availability, and calibration feasibility. Each domain was assigned a prespecified score from 0 to 2, yielding a total score ranging from 0 to 8. We also conducted a deterministic numerical illustration to demonstrate how an identical log-linear association can yield different hazard ratios when reported using cohort-specific standard deviation units.

Results

Five prospective cohorts comprising 7,438 participants were included. The SRI distinguished studies with high, conditional, and low readiness for effect-scale harmonisation. In several cohorts, incomplete reporting of standard deviations or equivalent distributional information precluded direct transformation of standardised estimates to a common absolute exposure scale without additional assumptions or external data. In the numerical illustration, a reference hazard ratio of 1.20 per 1-year increase corresponded to hazard ratios of 2.49 and 1.73 per 1-standard-deviation increase when the corresponding standard deviations were 5 and 3 years, respectively.

Conclusions

Consistent effect direction does not establish quantitative comparability. The Synthesis Readiness Framework provides a structured pre-pooling assessment of estimand alignment, exposure-scale comparability, distributional reporting, and calibration feasibility. The SRI is intended as a transparent methodological decision aid to guide the choice between quantitative harmonisation, structured narrative synthesis, and individual participant data meta-analysis; it is not a validated rule for determining meta-analysis eligibility.

INTRODUCTION

DNA methylation-based epigenetic clocks, particularly GrimAge, have emerged as robust predictors of all-cause mortality and biological aging [1–4]. First-generation clocks were developed to predict chronological age [5, 6], whereas second-generation clocks, such as PhenoAge [7] and GrimAge [8, 9], were trained against phenotypic and mortality outcomes, and third-generation measures such as DunedinPACE capture the pace of biological aging [10]. Prospective cohort studies consistently demonstrate that higher GrimAge acceleration is directionally associated with increased mortality risk, positioning it as a prime candidate for clinical risk stratification and precision geroscience [8, 11]. However, as the evidence base expands, a fundamental methodological challenge has emerged: the assumption that these directionally consistent associations are quantitatively comparable across different populations and study designs.

The interpretability of these associations is frequently compromised by heterogeneity. Traditional systematic reviews often categorize heterogeneity into clinical (biological/demographic) and methodological (analytical) layers. While these layers are well-recognized, a third, under-appreciated layer, metric or effect-scale heterogeneity, poses a critical barrier to evidence synthesis. In the GrimAge literature, effect estimates are reported using vastly different units of exposure contrast, such as absolute one-year increments versus population-specific one-standard deviation (1-SD) z-score increments. In addition, the reliability of clock estimates depends on computational and measurement properties that vary across cohorts and assays [12].

Conventional meta-analysis rests on the foundational assumption that pooled effect estimates represent comparable estimands [13, 14]. However, the current literature does not consistently satisfy this assumption due to incomplete reporting of distributional parameters and heterogeneous exposure scaling. Because a 1-SD increment in one cohort may represent a different underlying exposure contrast than a 1-SD increment in another, numerically divergent hazard ratios cannot be mathematically or biologically equated without rigorous calibration and, where necessary, effect-scale transformation [15]. Consequently, deriving a pooled hazard ratio from the current evidence base may yield a summary estimate that is statistically precise but biologically uninterpretable.

Therefore, evaluating the synthesis readiness of the evidence, rather than aggregating non-commensurable metrics, must precede any attempt at quantitative synthesis. Accordingly, this study aims to: (1) introduce a three-layer heterogeneity framework to categorize barriers to comparability in continuous biomarker evidence; (2) develop and operationalize a proof-of-concept Synthesis Readiness Index (SRI) to translate methodological concerns into reproducible synthesis decisions; and (3) evaluate the synthesis readiness of current GrimAge-mortality evidence as an empirical case study to demonstrate the framework's utility.

METHODS

2.1. Search Strategy and Study Selection

We conducted a systematic review to identify prospective cohort studies examining associations between GrimAge-related measures and all-cause mortality. The review protocol was registered prospectively in

PROSPERO (CRD420261449280). We searched PubMed, Embase, Web of Science, and Scopus from database inception to July 2026. The complete electronic search strategies for all databases, including search dates, controlled vocabulary terms, free-text terms, and applied filters, are provided in Supplementary Table S3.

Eligible studies were prospective cohort studies that reported hazard ratios (HRs), or other time-to-event effect estimates that could be interpreted as relative hazards, for the association between a GrimAge-related measure and all-cause mortality. We excluded cross-sectional studies, studies that did not report a usable mortality effect estimate, studies without a GrimAge-related exposure measure, duplicate reports from overlapping cohorts when a more complete or relevant report was available, and reports with insufficient information to characterise the reported exposure contrast.

Two reviewers independently screened titles and abstracts and subsequently assessed full-text reports for eligibility. Disagreements were resolved through discussion; when consensus could not be reached, a third reviewer adjudicated. Duplicate records were identified and removed using EndNote. Rayyan was used to support title and abstract screening. The study-selection process is reported using a PRISMA 2020 flow diagram.

2.2. Data Extraction and Audit

Two reviewers independently extracted data using a piloted, standardised data-extraction form. Extracted information included study setting, recruitment period, participant characteristics, sample size, duration of follow-up, number of deaths, definition and derivation of the GrimAge-related measure, laboratory platform and preprocessing procedures were reported, outcome ascertainment, covariate adjustment set, reported effect estimate and 95% confidence interval, and the exposure contrast used to estimate the association.

For the effect-scale audit, we additionally extracted information required to evaluate whether published estimates could be expressed on a common exposure scale. This information included whether the exposure was reported per absolute unit, per standard deviation, per interquartile range, or using another study-specific contrast; the reported mean, standard deviation, variance, interquartile range, or other distributional information; and whether the reported information was sufficient for a transparent mathematical transformation. When essential information was not reported, we recorded the absence of information rather than imputing distributional parameters.

2.3. Development of the Synthesis Readiness Index (SRI)

We developed the Synthesis Readiness Index (SRI) as a proof-of-concept framework for evaluating whether study-level effect estimates for continuous biomarkers are sufficiently specified for transparent effect-scale harmonization. The SRI was not designed to assess study validity, risk of bias, or certainty of evidence. These constructs were evaluated separately using established tools and approaches.

The initial domains were derived from methodological requirements for valid quantitative synthesis of continuous exposures: comparability of the target estimand, alignment of exposure metrics, availability of distributional information necessary for scale transformation, and feasibility of calibration without

unverifiable assumptions. The framework was informed by literature on estimands in epidemiology and survival analysis, re-expression of regression coefficients, effect-size transformation, and harmonization of biomarker data.

Four domains were prespecified: (1) estimand alignment; (2) metric comparability; (3) distributional availability; and (4) calibration feasibility. Each domain was scored as 0, 1, or 2, corresponding respectively to substantial limitation, limited readiness, and high readiness. Domain definitions, decision rules, examples, and scoring justifications are reported in Table 2 and Supplementary Table S1.

The total score was calculated by summing the four domain-specific scores, with possible values ranging from 0 to 8. Scores of 0–3 were interpreted as low readiness, scores of 4–6 as conditional readiness, and scores of 7–8 as high readiness for effect-scale harmonization. These categories were prespecified as conceptual decision aids rather than empirically validated thresholds. Accordingly, total scores were interpreted alongside domain-specific limitations; a favorable total score did not override a critical deficit in distributional availability or calibration feasibility.

2.4. Development Process of the SRI

The four domains of the SRI were not selected arbitrarily; they were derived inductively from fundamental assumptions required for valid quantitative synthesis of continuous exposures. Specifically, the domains map directly to the methodological literature on: (1) estimand theory and comparability in survival analysis [13, 14]; (2) exposure metric alignment and standardization in epidemiological synthesis [15]; (3) the statistical requirements for variance-based transformations and effect-size conversion [17]; and (4) the feasibility of calibration in the absence of individual-level data [13, 17]. By anchoring the SRI in these established methodological principles, the index provides a theoretically grounded, reproducible audit of synthesis readiness.

2.5. Synthesis Approach

Given the identified metric heterogeneity, a conventional random-effects meta-analysis was deemed methodologically inappropriate. This decision is consistent with Cochrane guidance that quantitative pooling should be avoided when effect estimates are not sufficiently comparable or when required distributional information is unavailable [17]. This decision was based on the absence of comparable exposure contrasts and insufficient distributional information required for defensible effect-scale transformation. Instead, we employed a critical narrative synthesis structured around the three-layer heterogeneity framework, focusing on the interpretability and comparability of the reported effect estimates.

2.6. Numerical Demonstration of Effect-Scale Distortion

To illustrate the mathematical consequences of reporting an identical underlying association on different exposure scales, we constructed a deterministic numerical demonstration rather than a stochastic simulation study. We specified a reference hazard ratio (HR) of 1.20 per 1-year increase in the biomarker and derived the corresponding HRs for standardized contrasts using the log-linear relation between exposure scale and the reported effect estimate.

Under this specification, a 1-SD contrast with an SD of 5 years corresponds to an HR of approximately 2.49, whereas a 1-SD contrast with an SD of 3 years corresponds to an HR of approximately 1.73. These values represent mathematically equivalent expressions of the same specified per-year association under different exposure distributions.

The demonstration was used to show how numerically different HRs can arise solely from differences in exposure scaling, without implying that the resulting estimates are directly poolable. No random-effects meta-analysis was performed in this demonstration because precision estimates, sample sizes, and between-study variance parameters were not specified. The purpose was therefore to illustrate effect-scale distortion rather than to estimate a meta-analytic summary effect.

2.7. Risk of Bias Assessment

Risk of bias was assessed separately from synthesis readiness. Because the included studies were prospective observational studies evaluating a prognostic biomarker in relation to mortality, risk of bias was assessed using the Quality In Prognosis Studies (QUIPS) framework [26]. QUIPS evaluates six domains relevant to prognostic factor research: study participation, study attrition, prognostic factor measurement, outcome measurement, study confounding, and statistical analysis and reporting [26].

Two reviewers independently assessed the included studies using the predefined QUIPS domains. Disagreements were resolved through discussion and, when necessary, adjudication by a third reviewer. Domain-level judgments were categorized as low, moderate, or high risk of bias.

Risk-of-bias judgments were not incorporated into the Synthesis Readiness Index because risk of bias and effect-scale comparability represent distinct methodological dimensions. The risk-of-bias assessment therefore informed interpretation of the evidence base but did not determine SRI scores or the study-level synthesis-readiness category.

RESULTS

3.1. Study Selection and Characteristics

The systematic search identified 4,726 records. After removal of duplicates and screening, five prospective cohort studies met the inclusion criteria, encompassing a total of 7,438 participants: the Women's Health Initiative (WHI) [19], the ESTHER cohort study [20], the InCHIANTI study [21], the Irish Longitudinal Study on Ageing (TILDA) [22], and Generation Scotland [23]. All included studies reported positive, directionally consistent associations between GrimAge-related measures and all-cause mortality, as summarized in the PRISMA flow diagram (Fig. 1) [16].

3.2. Risk of Bias Assessment

Risk-of-bias assessment using the QUIPS framework revealed that the included cohorts generally exhibited high methodological rigor, with no studies classified as having a high overall risk of bias. Specifically, two

studies (WHI and Generation Scotland) were judged as having a low risk of bias across most domains, while three studies (ESTHER, InCHIANTI, and TILDA) presented a moderate risk of bias. The most consistent source of moderate risk was identified in the study confounding domain, primarily due to incomplete reporting of adjustment for estimated blood cell proportions. Detailed domain-level judgments for each included cohort are presented in Supplementary Table S2.

3.3. Layer 1 and 2 Heterogeneity: Biological and Analytical

As expected, the included studies exhibited clinical and methodological diversity. Cohorts varied in baseline age, sex distribution, and follow-up duration (Layer 1). Analytically, studies utilized different covariate adjustment models and varying implementations of the GrimAge algorithm (Layer 2). While these layers introduce expected heterogeneity, they are traditionally manageable through standard meta-analytic subgrouping or meta-regression.

3.4. Layer 3 Heterogeneity: The Metric and Effect-Scale Barrier

The critical barrier to synthesis emerged at the third layer. The exposure scales used to report hazard ratios were not directly commensurable without additional calibration information. One study (WHI) reported the HR per one-year increase in GrimAge acceleration, while the remaining four studies reported HRs per one-standard deviation (1-SD) increase or study-specific standardized contrasts. Crucially, the studies utilizing the 1-SD metric did not consistently report the underlying standard deviation values, rendering it impossible to mathematically translate the 1-SD contrast back into an absolute year-based contrast without making unsupported distributional assumptions.

Framework for evaluating the comparability of continuous biomarker effect estimates. Biological/demographic and analytical/methodological heterogeneity represent recognized sources of variation that may be addressed through subgroup analyses or meta-regression. Metric/effect-scale heterogeneity represents a critical barrier when exposure contrasts are expressed using non-comparable units, such as absolute time increments versus population-specific standardized contrasts. A metric harmonization assessment is required before determining synthesis readiness and selecting quantitative pooling, narrative synthesis, or IPD-based harmonization.

3.5. Empirical Application of the SRI

To demonstrate the utility of the SRI, we applied the structured methodological index to the five included prospective cohorts (Table 1). The scoring revealed critical metric barriers that dictate the appropriate harmonization pathway. The three studies with SRI scores of 5–6 were classified as having conditional synthesis readiness. However, their domain-level profiles indicated limitations relevant to quantitative harmonization, particularly regarding distributional information and calibration feasibility. These limitations do not imply that quantitative synthesis is categorically precluded, but indicate that additional information, explicit harmonization assumptions, or sensitivity analyses would be required before combining estimates. Accordingly, the total SRI score was interpreted together with the domain-specific profile rather than as a standalone criterion for meta-analysis eligibility.

Table 1
Characteristics of Included Cohort Studies and Reported GrimAge–Mortality Effect Estimates

Cohort	Country	N	Baseline age (years)	GrimAge measure	Exposure contrast	Mortality association (adjusted)
WHI	USA	1900	63.2 ± 7.1	AgeAccelGrim	Per 1-year increase	HR 1.063 (1.042–1.084)
ESTHER	Germany	1771	61.7 ± 6.6 / 62.0 ± 6.7	AgeAccelGrim	Per 1-SD increase	HR 1.47 (1.32–1.64)
InCHIANTI	Italy	699	63.1 ± 15.7	Standardized baseline DNAmGrimAge	Per 1-SD increase	HR 1.38 (1.23–1.55)
TILDA	Ireland	490	63.4 ± 8.9	GrimAge-related measure	Standardized / study-specific	Study-specific estimate; not directly comparable
Generation Scotland	UK (Scotland)	2578	50.0 ± 12.5	Age-adjusted DNAm GrimAge	Per 1-SD increase	HR 2.10 (1.36–3.25)

Footnote: All-cause mortality was the common outcome across cohorts. Study-level SRI domain scores, total SRI scores, and the corresponding synthesis pathway are provided in Supplementary Table S1.

Table 2
Operational criteria for the Synthesis Readiness Index (SRI)

SRI domain	High readiness (2)	Limited readiness (1)	Substantial limitation (0)
Estimand Alignment	Exposure definition, contrast, outcome, and time-to-event estimand are sufficiently aligned across studies	Core estimand is broadly aligned, but differences require explicit interpretive assumptions	Exposure-response estimands are materially different or cannot be considered sufficiently comparable
Metric Comparability	Exposure contrasts are reported on identical or directly comparable scales	Exposure contrasts differ but can potentially be expressed on a common scale using reported parameters	Exposure contrasts are non-comparable and cannot be expressed on a common scale with available information
Distributional Availability	Required distributional parameters (e.g., SD, variance, or equivalent information) are completely reported	Some required parameters are available, but additional transparent assumptions are needed	Required distributional information is unavailable, preventing defensible scale conversion
Calibration Feasibility	Scale harmonization is directly possible using reported information without unsupported assumptions	Harmonization is possible with explicit, transparent, and sensitivity-tested assumptions	Harmonization would require unsupported or unverifiable assumptions
Footnote: Each domain is scored from 0 to 2 according to the predefined criteria shown above. The complete study-level scoring audit, including domain-specific scores, total SRI scores, and scoring justification, is provided in Supplementary Table S1.			
SRI Score (0–8)			

Visual summary of SRI scoring results for the five included prospective cohorts (WHI, Generation Scotland, ESTHER, InCHIANTI, TILDA). Each cohort is represented by coloured blocks corresponding to their total SRI score (range 0–8), where each block represents one point. Cohorts are classified into three harmonization pathways based on SRI scores: High Readiness for Harmonization (7–8, green blocks), Conditional Harmonization (4–6, orange blocks), and Structured Narrative/IPD Required (0–3, red blocks). This figure demonstrates the practical utility of the SRI framework in guiding defensible synthesis decisions for continuous biomarker evidence.

Notably, three cohorts (Generation Scotland, ESTHER, and InCHIANTI) scored in the 'Conditional Harmonization' range (5–6). However, a closer domain-level inspection (Supplementary Table S1) reveals that these cohorts received a score of 0 or 1 in the Distributional Availability domain. Consequently, despite a moderate aggregate score, direct quantitative pooling remained mathematically impossible without external distributional data. This highlights a critical feature of the SRI: the framework prevents the false security of a moderate total score by highlighting domain-specific limitations, particularly regarding distributional reporting.

3.5. Numerical Demonstration of Effect-Scale Distortion

The numerical demonstration showed that an identical underlying association can generate substantially different published hazard ratios when exposure contrasts are expressed on different scales. Under a specified reference effect of HR = 1.20 per 1-year increase, the corresponding HR was 2.49 for a 1-SD contrast when the cohort-specific SD was 5 years and 1.73 when the SD was 3 years. Thus, the same underlying per-year association can appear substantially different when expressed using population-specific standard deviation units.

These differences arise from the exposure scale rather than from a change in the specified underlying association. The example therefore illustrates why numerical similarity of reported hazard ratios does not establish quantitative comparability and why exposure-scale harmonization should precede quantitative pooling.

DISCUSSION

4.1. Summary of Findings

This critical methodological synthesis demonstrates that while GrimAge is a robust predictor of mortality [8, 9, 22], the current evidence base is fragmented by substantial effect-scale heterogeneity. The assumption that directionally consistent hazard ratios are quantitatively comparable cannot be reliably justified when exposure contrasts are expressed in non-commensurable units [14, 15]. The prognostic signal associated with GrimAge remains consistently observed across prospective cohorts; the challenge lies in the methodological translation of that signal across heterogeneous study designs and reporting practices.

4.2. Clarifying the Scope of Synthesis

It is important to clarify the scope of our conclusions. This critical synthesis does not suggest that GrimAge-mortality associations are invalid, nor does it imply that meta-analysis is inherently inappropriate for biomarker research. Rather, it highlights a fundamental prerequisite: quantitative synthesis requires the alignment of exposure metrics and interpretational units before aggregation [13, 17]. The biological signal of GrimAge remains robust; the challenge lies in ensuring that the magnitude of that signal is faithfully represented when evidence is synthesized across studies.

4.3. The Synthesis Readiness Index as a Methodological Tool

The SRI represents a proof-of-concept attempt to operationalize the pre-pooling audit of continuous biomarker evidence. By decomposing synthesis readiness into four discrete, scorable domains, the SRI provides systematic reviewers with a transparent, reproducible framework to determine the appropriate synthesis pathway before engaging in quantitative aggregation. Importantly, the SRI is not a substitute for risk-of-bias assessment using validated tools like QUIPS [26]; rather, it addresses a complementary methodological dimension—the commensurability of effect estimates.

The numerical demonstration further illustrates the practical importance of this audit. When the same underlying association is expressed using different exposure scales, the resulting hazard ratios can differ substantially even when the underlying per-unit relationship is held constant. The example is intended to demonstrate effect-scale distortion, not to estimate a universal magnitude of pooling bias. It therefore reinforces the need to assess exposure-scale comparability before quantitative aggregation [13, 15]. Figure 4 illustrates how a metric harmonization-informed pathway differs from a conventional aggregation pathway by placing an explicit assessment of effect-scale comparability before any decision to pool estimates.

4.4. Generalizability: The Continuous Biomarker Synthesis Problem

While demonstrated here using epigenetic aging biomarkers, the effect-scale heterogeneity identified in this study represents a potentially widespread methodological challenge in the synthesis of continuous biomarker evidence broadly. The Synthesis Readiness Framework and SRI may be transferable across diverse domains of biomedical research. In Aging and Geroscience, the synthesis of PhenoAge [7], DunedinPACE [10], and telomere length is similarly hampered by varying assay platforms and scaling metrics [1–3]. In Genetics and Multi-omics, the pooling of Polygenic Risk Scores (PRS) or metabolomic risk scores is frequently compromised because the variance of the score differs drastically across ancestral populations, making 1-SD contrasts non-commensurable without SRI-guided calibration and portability assessment [24]. In Inflammation and Clinical Biomarkers, continuous markers such as C-reactive protein (CRP), interleukin-6 (IL-6), and cardiac biomarkers (NT-proBNP, high-sensitivity troponin) often suffer from heterogeneous distributional reporting and assay variability [25], preventing defensible quantitative aggregation. Furthermore, in Cardiovascular and Metabolic Disease research, biomarkers such as blood pressure, lipid profiles, HbA1c, and fasting glucose exhibit similar challenges with unit standardization and population-specific reference ranges. In all these domains, the SRI provides a standardized, pre-pooling audit to ensure that researchers do not inadvertently generate statistically precise but biologically artifactual summary estimates."

4.5. Future Directions: From Aggregation to Harmonization

Ultimately, the inability to achieve quantitative comparability from the current aggregate evidence underscores a broader limitation in biomarker synthesis. When distributional parameters and scaling metrics are irreconcilable at the study level, the next frontier for evidence synthesis must shift toward individual-level data harmonization and IPD meta-analyses [30]. Future research should prioritize the development of reference-anchored calibration frameworks to mathematically translate heterogeneous exposure contrasts into a common biological scale and the establishment of standardized reporting guidelines—building upon existing frameworks like PRISMA [16]—that mandate the reporting of exposure distributional parameters. When aggregate harmonization fails, advanced dose-response modelling [31] and narrative harmonization techniques [32] offer alternative pathways for evidence synthesis.

4.6. Limitations

Several limitations should be acknowledged. First, the number of available prospective cohorts with published GrimAge-mortality associations remains limited, which constrains the empirical demonstration of the SRI. Second, the assessment of synthesis readiness depends on the completeness and transparency of published reporting. Third, the proposed SRI framework serves as a proof-of-concept decision aid; its scoring criteria and thresholds require external validation in future methodological studies across different biomarker domains.

4.7. Conclusion

Directional consistency is not synonymous with quantitative comparability. The proposed Synthesis Readiness Framework and Index provide a structured, pre-pooling audit tool to prevent the aggregation of non-commensurable continuous biomarker metrics. By systematically evaluating estimand alignment, metric comparability, distributional availability, and calibration feasibility, the SRI guides researchers toward defensible synthesis pathways and helps determine when quantitative pooling, conditional harmonization, or structured narrative/IPD synthesis is appropriate. The framework is intended as a methodological decision aid rather than a validated statistical decision rule, and its thresholds require external validation.

Declarations

Clinical trial number: Not applicable.

Ethics Approval and Consent to Participate: Not applicable. This study is a methodological synthesis of previously published aggregate data and does not involve human participants or individual-level data.

Consent for Publication: Not applicable.

Availability of Data and Materials: All data analyzed in this study are derived from previously published sources. The extracted dataset and study-level information used for synthesis are available from the corresponding author upon reasonable request. The PROSPERO protocol is publicly available at CRD420261449280.

Competing Interests: The authors declare no conflicts of interest.

Funding: The authors received no specific funding for this work.

Authors' Contributions: BU conceptualized the study, developed the Synthesis Readiness Framework, performed the numerical demonstration, and drafted the manuscript. DA contributed to the systematic search, data extraction, and critical revision of the manuscript. SM contributed to data extraction, SRI scoring, and critical revision of the manuscript. All authors read and approved the final manuscript.

Acknowledgements: The authors thank the investigators of the WHI, ESTHER, InCHIANTI, TILDA, and Generation Scotland cohorts for making their findings publicly available.

References

1. Field AE, Robertson NA, Wang T, Havas A, Ideker T, Adams PD. DNA methylation clocks in aging: categories, causes, and consequences. Mol Cell. 2018;71(6):882–95.
2. Ferrucci L, Levine ME, Kuo PL, Simonsick EM. Time and the metrics of aging. Circ Res. 2018;123(7):740–4.
3. Ferrucci L, Gonzalez-Freire M, Fabbri E, Simonsick E, Tanaka T, Moore Z, et al. Measuring biological aging in humans: a quest. Aging Cell. 2020;19(2):e13080.
4. Horvath S, Raj K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. Nat Rev Genet. 2018;19(6):371–84.
5. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell. 2013;49(2):359–67.
6. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013;14(10):R115.
7. Levine ME, Lu AT, Quach A, Chen BH, Assimes TL, Bandinelli S, et al. An epigenetic biomarker of aging for lifespan and healthspan. Aging. 2018;10(4):573–91.
8. Lu AT, Quach A, Wilson JG, Reiner AP, Aviv A, Raj K, et al. DNA methylation GrimAge strongly predicts lifespan and healthspan. Aging. 2019;11(2):303–27.
9. Lu AT, Binder AM, Zhang J, Yan Q, Reiner AP, Satoh A, et al. DNA methylation GrimAge version 2. Aging. 2022;14(24):9484–516.
10. Belsky DW, Caspi A, Corcoran DL, Sugden K, Poulton R, Arseneault L, et al. DunedinPACE, a DNA methylation biomarker of the pace of aging. eLife. 2022;11:e73420.
11. Moqri M, Herzog C, Poganik JR, Justice JN, Belsky DW, Higgins-Chen AT, et al. Validation of biomarkers of aging. Nat Med. 2024;30(2):360–72.
12. Higgins-Chen AT, Thrush KL, Wang Y, Minteer CJ, Kuo PL, Wang M, et al. A computational solution for bolstering reliability of epigenetic clocks: implications for clinical trials and longitudinal tracking. Nat Aging. 2022;2(7):644–61.
13. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to Meta-Analysis. 2nd ed. Hoboken (NJ): Wiley; 2021.
14. Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. Am J Epidemiol. 1987;125(5):761–8.
15. Hamling J, Lee P, Weitkunat R, Ambühl M. Facilitating meta-analyses by deriving relative effect and precision estimates for alternative comparisons from a set of estimates presented by exposure level or disease category. Stat Med. 2008;27(7):954–70.
16. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71.
17. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, editors. Cochrane Handbook for Systematic Reviews of Interventions. Version 6.5. Cochrane; 2024.
18. Wells GA, Shea B, O'Connell D, Peterson J, Welch V, Losos M, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Ottawa: Ottawa Hospital Research Institute; 2000.

19. Beydoun MA, Beydoun HA, Ashe J, Georgescu MF, Horvath S, Lu A, et al. Relationships of depression and antidepressant use with epigenetic age acceleration and all-cause mortality among postmenopausal women. Aging. 2024;16(11):8446–71.
20. Li X, Zhang Y, Gao X, Holleczek B, Schöttker B, Brenner H. Comparative validation of three DNA methylation algorithms of ageing and a frailty index in relation to mortality: results from the ESTHER cohort study. EBioMedicine. 2021;74:103686.
21. Kuo PL, Moore AZ, Tanaka T, Belsky DW, Lu AT, Horvath S, et al. Longitudinal changes in epigenetic clocks predict survival in the InCHIANTI cohort. Nat Aging. 2024;4(4):534–40.
22. McCrory C, Fiorito G, Hernandez B, Polidoro S, O'Halloran AM, Hever A, et al. GrimAge outperforms other epigenetic clocks in the prediction of age-related clinical phenotypes and all-cause mortality. J Gerontol Biol Sci Med Sci. 2021;76(4):741–9.
23. Hillary RF, Stevenson AJ, McCartney DL, Campbell A, Walker RM, Howard DM, et al. Epigenetic measures of ageing predict the prevalence and incidence of leading causes of death and disease burden. Clin Epigenetics. 2020;12(1):115.
24. Duncan L, Shen H, Gelaye B, Meyers J, Ressler K, Feldman M, et al. Analysis of polygenic risk score usage and performance in diverse human populations. Nat Commun. 2019;10(1):3328.
25. Lippi G, Sanchis-Gomar F. Preanalytical and analytical variability of high-sensitivity C-reactive protein. Clin Chem Lab Med. 2022;60(1):12–8.
26. Hayden JA, van der Windt DA, Cartwright JL, Côté P, Bombardier C. Assessing bias in studies of prognostic factors. Ann Intern Med. 2013;158(4):280–6. 10.7326/0003-4819-158-4-201302190-00009.
27. Linakis SA, Trikalinos TA. Re-expressing regression coefficients for meta-analysis of continuous exposures. BMC Med Res Methodol. 2024;24(1):112.
28. Gallardo-Gómez L, Delgado-Rodríguez M, Olmedo-Requena R. Impact of the standard deviation used for standardization on the meta-analysis of continuous outcomes. Res Synth Methods. 2023;14(4):567–78.
29. Jung S, Aloe AM. Harmonized standardized mean difference for meta-analysis of continuous outcomes. Res Synth Methods. 2025;16(1):45–59.
30. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ. 2010;340:c221.
31. Dekkers OM, Vandenbroucke JP, Cevallos M, Renehan AG, Altman DG, Egger M. COSMOS-E: guidance on conducting systematic reviews and meta-analyses of observational studies of etiology. PLoS Med. 2019;16(2):e1002742. 10.1371/journal.pmed.1002742.
32. Iqbal K, Intemann T, Börnhorst C, Zhang J, Aleksandrova K. Approaches for harmonization of biomarker data from multiple studies: a narrative methodological review. AJE Adv. 2025;1(1):uuaf005. 10.1093/ajeadv/uuaf005.

Figures

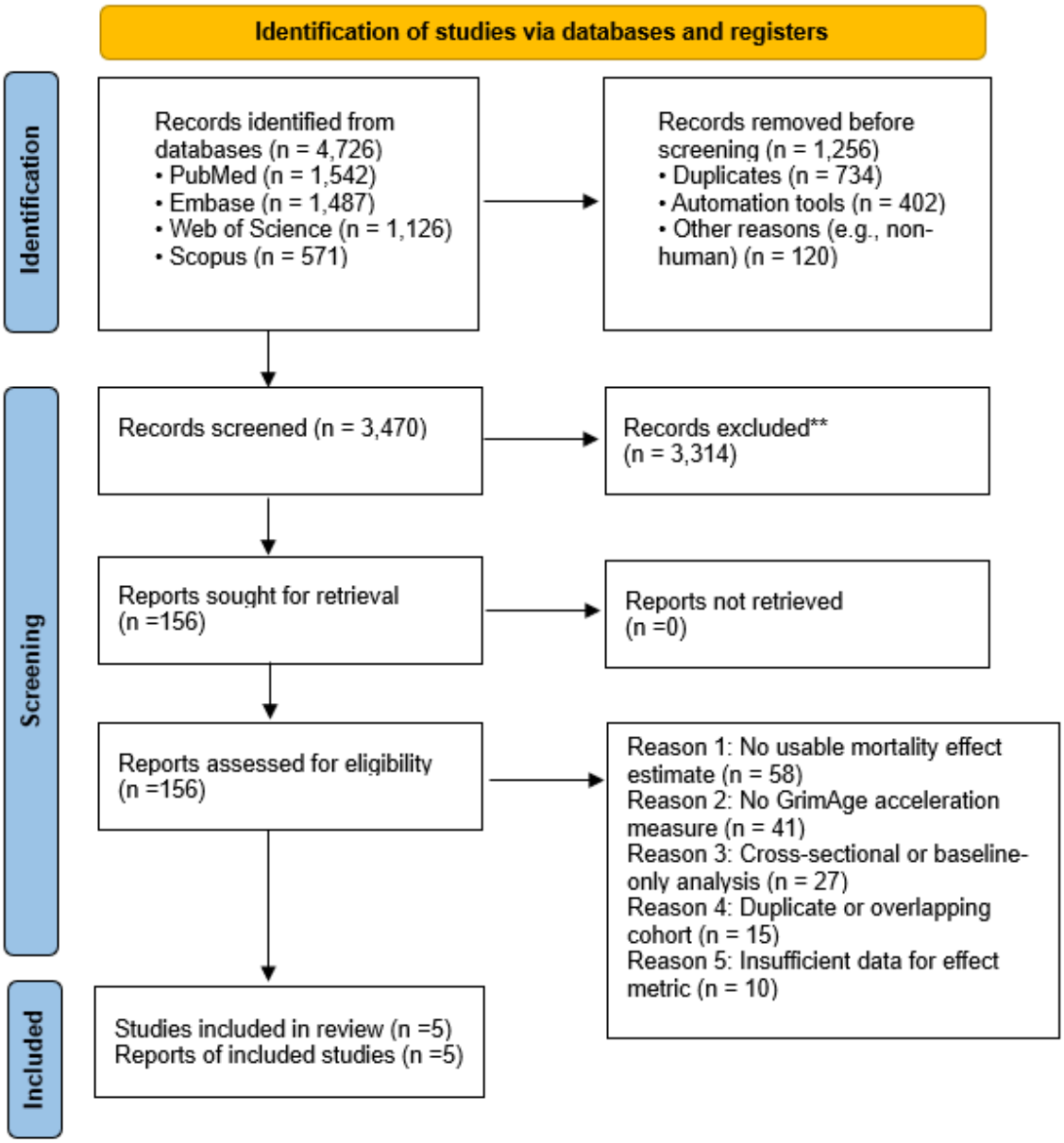


Figure 1

PRISMA 2020 flow diagram of study identification, screening, eligibility assessment, and inclusion.

Source: Page MJ, et al. BMJ 2021;372: n71. DOI: 10.1136/BMJ.n71.

This work is licensed under CC BY 4.0. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

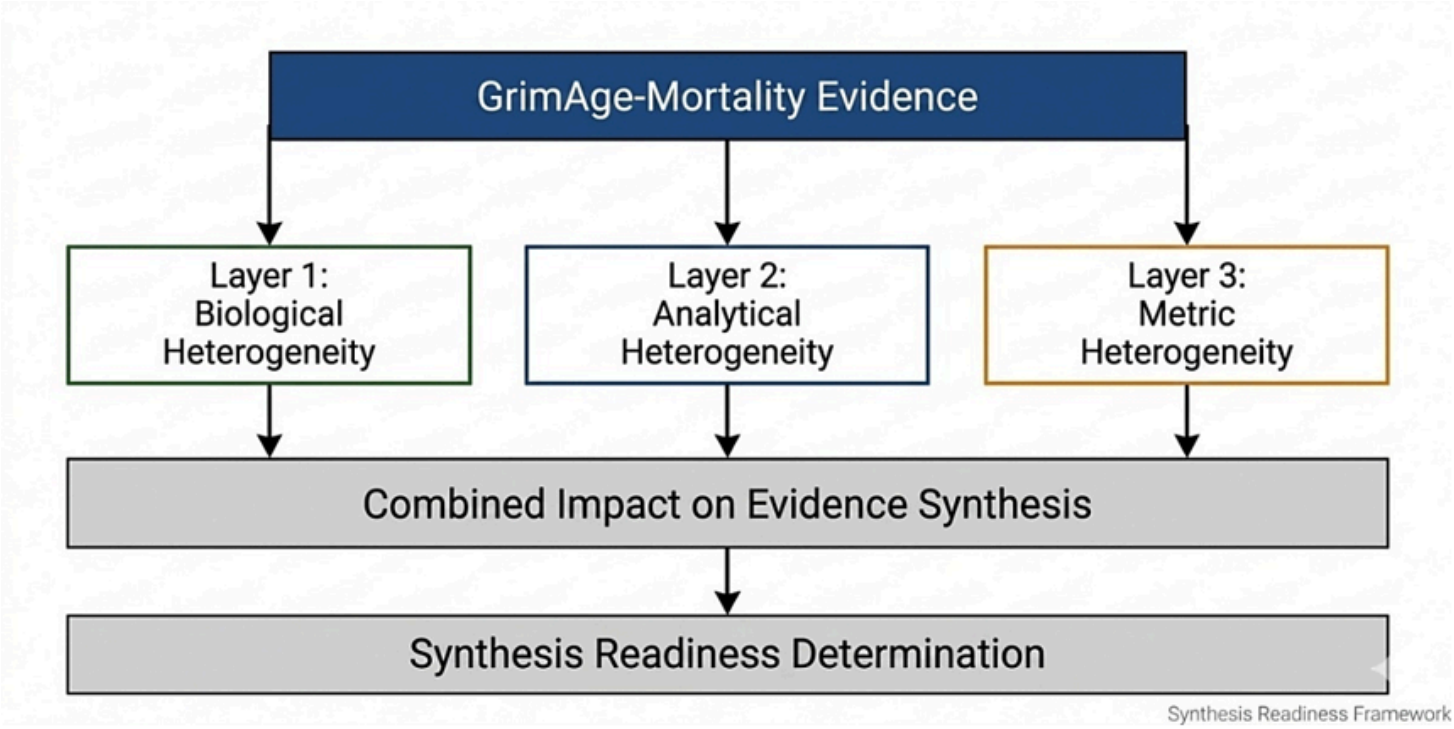


Figure 2

Three-layer heterogeneity framework

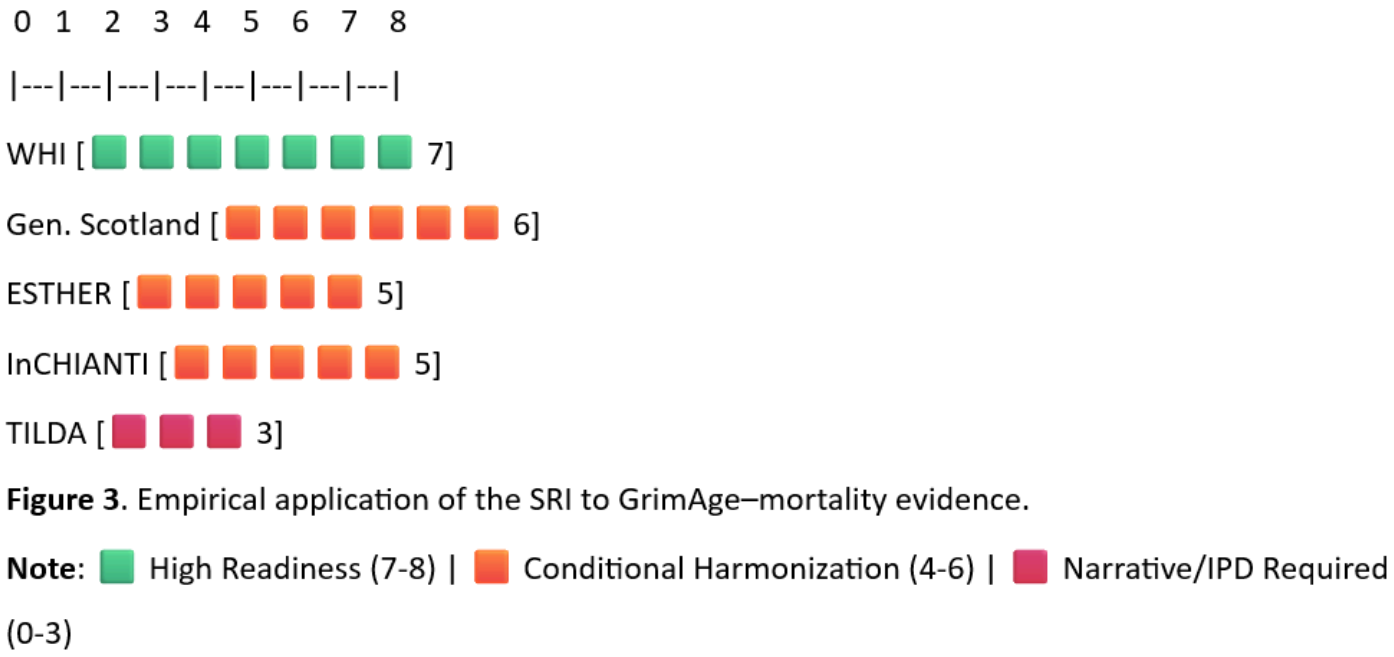


Figure 3. Empirical application of the SRI to GrimAge–mortality evidence.

Note: High Readiness (7-8) | Conditional Harmonization (4-6) | Narrative/IPD Required (0-3)

Figure 3

See image above for figure legend.

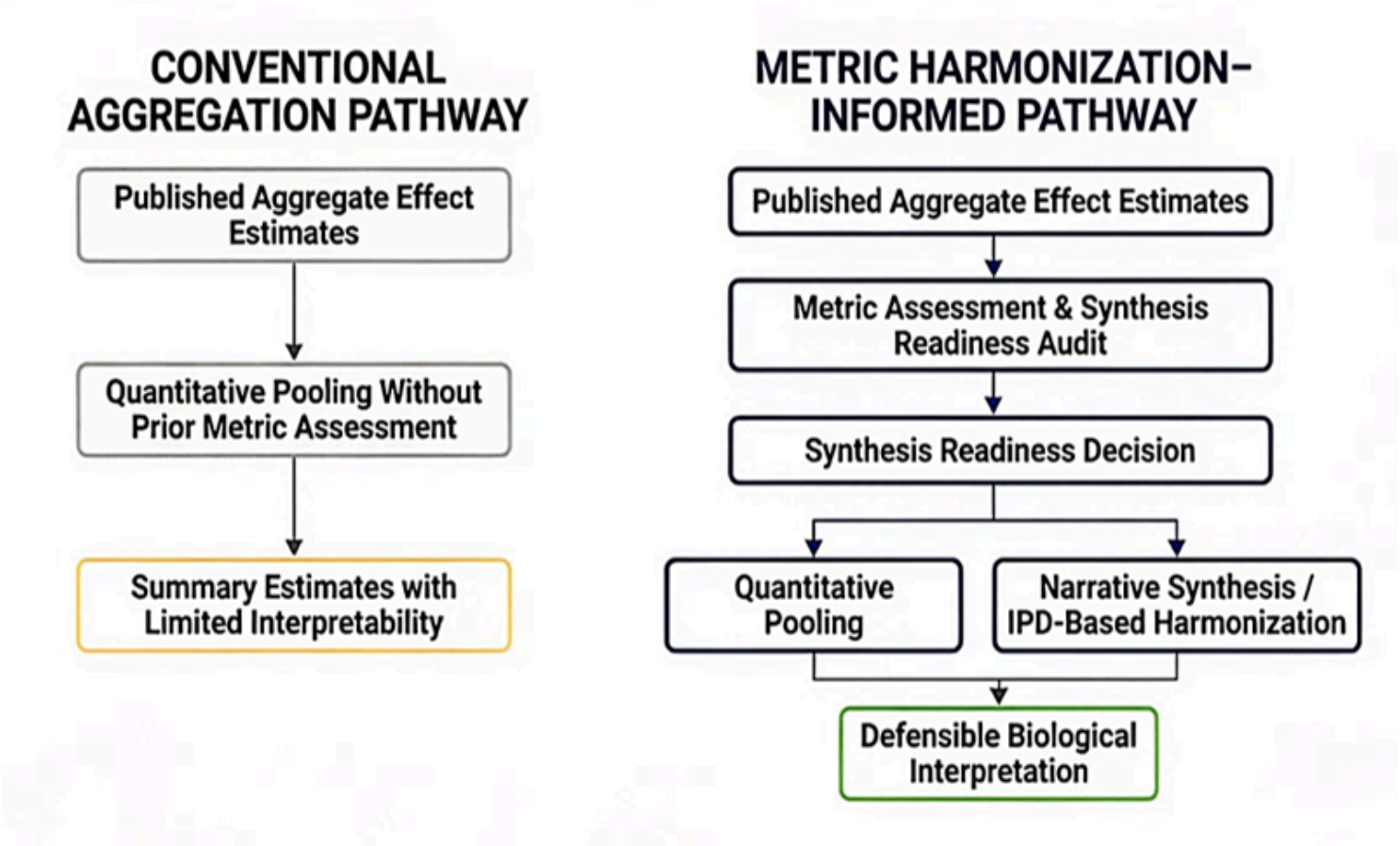


Figure 4

Metric harmonization-informed pathway versus conventional aggregation pathway.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTableS1.StudylevelauditoftheSynthesisReadinessIndexSRIforincludedcohorts.docx
- SupplementaryTableS2.StudylevelriskofbiasassessmentusingtheQUIPSframework.docx
- SupplementaryTableS3.CompleteelectronicsearchstrategiesforalldatabasesconductedinJuly2026.docx