Faculty of Medicine

**SYSTEMS BIOLOGY**

*Department of Human and Medical Genetics*

*Arnas Stučinskas, second year, group 1*

*Master's thesis*

*Dideliais kalbos modeliais paremta sistema teiginiams apie ilgaamžiškumo veiksnius ir jų įrodymų kokybei vertinti mokslinėje literatūroje*

*A Large-Scale LLM-Driven System for Extracting Claims About Longevity Factors and Assessing Evidence Quality from Scientific Literature*

Supervisor Prof. PhD Audronė Jakaitienė

Head of the department or clinic Prof. MD, PhD Algirdas Utkus

Vilnius, 2026

Student's email arnas.stucinskas@mf.stud.vu.lt

# CONTENTS

## LIST OF ABBREVIATIONS

| Abbreviation | Full term |
|---|---|
| ACU | Atomic Claim Unit |
| ADL | Activities of Daily Living |
| API | Application Programming Interface |
| BMI | Body Mass Index |
| EBM | Evidence-Based Medicine |
| EHR | Electronic Health Record |
| EPMC | Europe PubMed Central |
| GPT | Generative Pre-trained Transformer |
| HAGR | Human Ageing Genomic Resources |
| HALD | Human Aging and Longevity knowledge graph |
| HIIT | High-Intensity Interval Training |
| IADL | Instrumental Activities of Daily Living |
| ICO | Intervention, Comparator, Outcome |
| JSON | JavaScript Object Notation |
| JSONL | JavaScript Object Notation Lines |
| LLM | Large Language Model |
| MeSH | Medical Subject Headings |
| MR | Mendelian Randomisation |
| NLP | Natural Language Processing |
| PICO | Population, Intervention, Comparator, Outcome |
| PMID | PubMed Identifier |
| RCT | Randomised Controlled Trial |
| SD | Standard Deviation |
| SEF | Structured Extraction Frame |
| UMLS | Unified Medical Language System |

# INTRODUCTION

## 1.1 Research Problem and Relevance

Ageing is the single greatest risk factor for most non-communicable diseases, including cardiovascular disease, type 2 diabetes, neurodegeneration, and cancer (Kennedy et al., 2014) – a recognition that has given rise to geroscience, the field that seeks to understand the genetic, molecular, and cellular mechanisms by which ageing drives the onset of these conditions. As the global population aged over 60 is projected to double from approximately 1 billion in 2020 to more than 2 billion by 2050 (UN DESA, 2023), the societal burden of age-related morbidity has become one of the defining public health challenges of the twenty-first century. In response, the global market for longevity-focused products – including dietary supplements, nutraceuticals, lifestyle programmes, and pharmaceutical interventions targeting mechanisms of ageing – has expanded substantially. Annual global expenditure on anti-ageing products is estimated in the hundreds of billions of dollars, driven by consumer demand for evidence-based guidance on how to live longer and remain healthy.

Yet the evidence base underpinning longevity claims is profoundly uneven. Thousands of peer-reviewed publications are indexed in biomedical databases each year reporting on the effects of interventions on ageing-related outcomes, yet this literature is heterogeneous in quality, fragmented across disciplines, and not presented in a form that is easily comparable or actionable. A single intervention – resistance training, for example – may be studied in dozens of trials with different populations, outcome measures, follow-up durations, and methodological standards. A clinician, researcher, or health authority seeking to understand the aggregate state of evidence must synthesise across this literature manually, a task that is time-consuming, susceptible to selection bias, and practically impossible to keep up to date given the pace of publication.

Compounding this challenge is the nature of the claims themselves. Biomedical abstracts typically report multiple findings within a single paragraph, conflating different outcomes, populations, and effect sizes. These compound statements resist direct comparison and aggregation. Evidence synthesis methods such as systematic review and meta-analysis partially address this problem but are slow, expensive, and typically narrowly scoped. Living systematic reviews represent a step forward but remain labor-intensive and do not scale to the full breadth of the longevity literature.

The reproducibility of findings in biomedical research adds a further layer of complexity. The concern that a substantial proportion of published research findings may not replicate – articulated by Ioannidis, 2005 in the context of small-sample epidemiological research (Ioannidis, 2005) – is especially acute in geroscience, where studies are typically underpowered, outcome measures are heterogeneous, and publication bias towards positive findings is well-documented. Understanding

which longevity claims are supported by multiple independent studies, which rest on a single small trial, and which are contradicted by subsequent evidence is essential for any rational assessment of the field.

The biomedical informatics community has developed a range of computational tools to assist with literature synthesis. Natural language processing methods for extracting structured information from clinical trial abstracts – including population, intervention, comparator, and outcome (PICO) elements – have matured considerably over the past decade. Large language models have demonstrated the capacity to extract structured clinical information from unstructured text with a level of accuracy approaching human expert performance. These developments create an opportunity to address the evidence synthesis challenge in longevity research at a scale and speed that was previously impractical.

### 1.2 Research Gap and Novelty

Despite the proliferation of longevity-focused databases and computational tools, no systematic, quality-graded, claim-level evidence database for longevity interventions currently exists. Quality grading here refers to assessing evidence across multiple dimensions: study design (e.g. randomised trial vs. observational study), endpoint proximity (whether the measured outcome is a surrogate biomarker such as a blood marker, or a hard clinical outcome such as survival or functional independence), effect credibility, population relevance, and replication across studies. Existing resources tend to catalogue genes, variants, drugs, or pathways rather than decomposing individual claims and grading the evidence supporting each claim.

The concept of "atomic claim units" – minimal, self-contained representations of a single intervention–outcome relationship derived from a specific study – has emerged in the fact-checking and natural language processing literature as a powerful framework for claim decomposition and evidence structuring (Wanner et al., 2024; Hu et al., 2025; Metropolitansky & Larson, 2025; Zheng et al., 2026). Applied to biomedical literature, this approach enables systematic comparison across studies measuring the same or related intervention-outcome relationships, while preserving the study-level provenance of each extracted finding. For example, if an abstract reports that resistance training improved grip strength but did not change inflammatory biomarkers in older adults, the statement can be decomposed into separate ACUs for resistance training–grip strength and resistance training–inflammatory biomarkers, each retaining its own direction, endpoint type, and source publication.

The present thesis applies this claim decomposition and evidence-quality assessment framework to the longevity domain, combining automated PubMed retrieval, large language model-based relevance screening, structured extraction of atomic claim units using a frontier language model, multi-stage

filtering, entity normalisation, and a multi-dimensional evidence-tiering framework to produce a systematic, quality-graded, claim-level evidence database for longevity interventions in the published human biomedical literature.

### 1.3 Aim

The aim of this Master's thesis is to develop and evaluate an end-to-end natural language processing pipeline that extracts, filters, and quality-grades longevity-related evidence claims from the biomedical literature, producing a transparent and updatable structured evidence database linking interventions and exposures to ageing outcomes.

### 1.4 Objectives

1. To define a structured representation of longevity-related claims as Atomic Claim Units (ACUs) capturing factor (intervention or exposure), outcome, effect direction, study design, and population characteristics.
2. To implement and evaluate a pipeline for retrieving and screening longevity-related biomedical publications from PubMed using large language models.
3. To develop and apply an automated extraction approach for generating structured ACU-level information from unstructured abstract text.
4. To establish a multi-dimensional evidence-tiering framework that reflects study design quality, endpoint proximity to meaningful health outcomes, and effect credibility.
5. To characterise the structural properties of the resulting longevity evidence database, including factor (intervention or exposure) coverage, endpoint distribution, replication patterns, and hallmark of ageing coverage.
6. To make the resulting evidence database publicly accessible as a web resource (longevityevidence.org), enabling interactive querying of longevity evidence by intervention or exposure, outcome, evidence tier, and hallmark of ageing.

## LAY DESCRIPTION

Every week, headlines report new discoveries about what may help people live longer and healthier lives. One week the focus is exercise, the next week it is a supplement, a diet, or a medicine. For readers, clinicians, and researchers, it is difficult to know which findings are strong, which are early signals, and which are based only on indirect measurements.

The problem is not a lack of research. There are tens of thousands of scientific papers on ageing, healthspan, lifestyle, medicines, and biological markers. The problem is that the evidence is scattered across many abstracts, study types, outcomes, and technical terms. A single paper can contain several separate findings, and different papers often use different words for the same intervention or outcome.

This thesis describes a computational system that reads biomedical abstracts, filters the relevant studies, extracts individual claims, checks whether those claims are supported by the original abstract, and grades the quality of the evidence. The system also corrects cases where a wording change could reverse the meaning of an effect, such as when a decrease in muscle loss should be interpreted as an improvement in muscle mass.

The final evidence database contains 2,987 quality-graded evidence claims from 1,797 publications. These claims cover 793 factor nodes, such as exercise, vitamin D supplementation, or metformin, and 769 outcome nodes, such as survival, frailty, blood pressure, or biological age markers.

The resulting resource makes longevity evidence easier to inspect and compare. It does not replace expert systematic reviews, but it provides a transparent starting point for identifying stronger evidence, weaker evidence, and areas where more direct human research is still needed.

# LITERATURE REVIEW

## 2.1 Why Longevity Claims Are Hard to Evaluate?

The scientific study of human longevity is distinguished from most other fields of biomedical research by the exceptional complexity of its subject matter. Ageing is not a single process but an accumulation of interacting molecular, cellular, and systemic changes that unfold across decades, manifest differently across individuals and populations, and resist reduction to any single measurable variable (López-Otín et al., 2023). This biological complexity translates directly into methodological challenges that make the evaluation of longevity claims substantially harder than the evaluation of claims in acute medicine.

The first challenge concerns the diversity of study designs employed across the field. At one extreme, animal lifespan studies in model organisms such as Caenorhabditis elegans, Drosophila melanogaster, and mice offer the most direct measure of longevity effects and enable rapid, controlled testing of interventions. However, translating findings from short-lived invertebrates to humans is notoriously unreliable, and even murine models differ sufficiently from humans in metabolic rate, immune function, and genetic background to make direct extrapolation problematic (Scheibye-Knudsen, 2020). At the other extreme, randomised controlled trials in human populations offer the highest level of causal evidence but are prohibitively expensive and logistically demanding when survival is the primary endpoint. Trials powered to detect a meaningful reduction in all-cause mortality in healthy older populations require thousands of participants followed for many years, at a cost that few funding bodies are willing to bear (Newman et al., 2016; Rolland et al., 2023). The result is that the great majority of human longevity trials use surrogate or intermediate endpoints – measures that are expected to correlate with longevity but do not directly capture it.

The second challenge is the heterogeneity of endpoints. “Successful ageing” lacks a consensus definition (Newman et al., 2016), and the field employs a vast and poorly standardised array of outcome measures ranging from grip strength and gait speed to epigenetic clocks, inflammatory markers, telomere length, and composite frailty indices. Different studies of the same intervention may measure entirely different outcomes, making direct comparison impossible and aggregate conclusions unreliable. The surrogate endpoint problem is particularly acute in geroscience: a supplement may robustly lower an inflammatory biomarker without having any detectable effect on survival or functional independence (Cummings & Kritchevsky, 2022).

Third, the literature quality in ageing research is highly variable. Parish et al., 2025 conducted a methodological appraisal of the DrugAge database – a curated collection of longevity-relevant drug studies – and documented widespread deficiencies in reporting quality, including failure to report

effect sizes, inadequate description of statistical methods, and inconsistent use of controls. Small sample sizes are endemic, particularly in early-phase trials of novel interventions. Publication bias towards positive findings is well-established across biomedical research (Ioannidis, 2005) and is unlikely to spare this field.

Fourth, the language of longevity claims is imprecise. A published abstract may describe an intervention as having "improved ageing-related outcomes" without specifying the nature, magnitude, or clinical relevance of the improvement. Claims in the popular press and on supplement packaging typically involve further distortion. The absence of standardised language for reporting longevity claims makes systematic comparison across studies highly challenging without computational assistance.

Taken together, these challenges explain why, despite decades of research and thousands of publications, a coherent, quality-graded picture of the longevity evidence base has not previously been assembled. Addressing this gap requires tools that can operate at scale, impose consistent quality criteria, and decompose compound claims into comparable atomic units. The following sections review the literature relevant to each of these challenges: Section 2.2 examines evidence hierarchies and the endpoint problem; Section 2.3 surveys existing longevity knowledge resources and their limitations; Sections 2.4 and 2.5 address computational approaches to structured information extraction and claim decomposition; and Sections 2.6-2.8 cover entity normalisation and the state of the art in claim verification and evidence quality assessment.

### 2.2 Evidence Types and Endpoints in Ageing Research

The hierarchy of evidence in clinical medicine – with randomised controlled trials at the apex and expert opinion at the base – applies to ageing research but requires important modifications given the biology and economics of longevity science.

The gold standard for causal inference in clinical research is the randomised controlled trial, in which participants are randomly allocated to treatment or control conditions, eliminating the confounding that renders observational evidence susceptible to bias. In ageing research, randomised trials are feasible for many outcomes, and their use is growing. In the final analysis of this thesis, 2,288 of 2,987 ACUs (76.6%) derived from interventional studies, and randomised controlled trials alone accounted for 1,592 ACUs (53.3%). However, the specific challenge of survival as an endpoint means that randomised trials dominate the literature for clinical and functional outcomes rather than for longevity itself, which is predominantly studied in prospective cohort designs (Rolland et al., 2023).

The selection of appropriate endpoints is a central methodological challenge in geroscience. Justice et al. (2018) proposed a framework for selecting blood-based biomarkers for geroscience-guided

clinical trials, distinguishing between biomarkers that reflect underlying biological ageing processes and those that capture clinical manifestations of those processes. Cummings & Kritchevsky (2022) argued for a tiered approach to endpoint selection that prioritises health outcomes (survival, functional independence, incident disease) above physiological biomarkers, whilst acknowledging that biomarkers serve an important role in mechanistic studies and early-phase trials. Sanders (2023) described strategic considerations for biomarker development in translational geroscience, emphasising the need for validated, clinically meaningful proxies.

The frailty phenotype has been proposed as an integrative endpoint that captures the cumulative effects of ageing across multiple organ systems and correlates strongly with adverse outcomes including hospitalisation, institutionalisation, and death (Dent et al., 2019). Frailty assessment tools such as the Fried phenotype and the Rockwood frailty index have been increasingly adopted as primary or secondary endpoints in clinical trials of interventions targeting the ageing process.

Epigenetic clocks and other biological age estimators constitute a distinct endpoint category that has attracted considerable attention in recent years (Kriukov et al., 2025; Min et al., 2024; Li et al., 2025). These tools attempt to quantify the rate at which an individual is ageing relative to chronological age, using methylation patterns or other molecular signatures. In principle, a validated biological age clock that accurately predicts remaining lifespan would offer a practical surrogate endpoint for longevity trials, reducing the required follow-up time by years. In practice, however, the predictive validity and clinical utility of existing clocks remain subjects of active debate (Kriukov et al., 2025). Different clocks show limited agreement with one another, their responsiveness to interventions is uncertain, and the mechanistic interpretation of clock acceleration or deceleration is not fully established (Min et al., 2024). Li et al. (2025) reviewed the evidence for using biological age to predict healthspan and disease risk, concluding that while promising, these tools require further validation before they can serve as primary endpoints in regulatory settings.

The endpoint classification scheme employed in the present study – E1 (longevity/survival), E2 (clinical/functional), E3 (physiological/imaging), E4 (molecular/ageing clocks) – reflects this hierarchy. The proximity of an endpoint to what patients and society ultimately care about – longer, healthier lives – is a dimension of evidence quality that is not captured by study design alone. A perfectly designed randomised trial measuring a molecular biomarker of uncertain clinical significance provides less actionable evidence than a well-conducted cohort study following a survival outcome. In the base evidence-tiering model, the full weight vector sums to 1.00: endpoint proximity 0.30, study design quality 0.25, effect credibility 0.18, comparison structure 0.10, confounding adjustment 0.08, sample size 0.05, and population relevance 0.04. These author-

assigned weights were chosen a priori from clinical evidence-appraisal principles and are specified in Section 3.14, with sensitivity scenarios reported in Appendix C.

### 2.3 Longevity Resources and Knowledge Graphs

The past two decades have seen the development of several curated databases and knowledge resources to organise evidence on the biology and pharmacology of ageing. These resources represent significant achievements in evidence synthesis but also illustrate the limitations that motivate the automated approach developed here.

The Human Ageing Genomic Resources (HAGR), as described in Tacutu et al. (2018), is the most comprehensive collection of ageing-related databases, including GenAge (genes associated with ageing in model organisms and humans), AnAge (comparative biology of ageing), and DrugAge (longevity-related drugs). These databases are manually curated by domain experts, which ensures a high level of accuracy for the entries they contain. DrugAge, described by Barardo et al., 2017, catalogues compounds with evidence of lifespan extension in model organisms, providing a starting point for translational investigation. However, manual curation introduces inevitable constraints on coverage and update frequency: the databases reflect the state of the literature at the time of the last curation cycle, and rapidly accumulating evidence in active areas may not be reflected in a timely manner.

LongevityMap, developed by Budovsky et al. (2013), focuses specifically on human genetic variants associated with longevity, providing a curated catalogue of association studies. The Aging Atlas, developed by Aging Atlas Consortium (2021), integrates multi-omics data across cell types and tissues with the goal of characterising the molecular landscape of ageing. AgeFactDB, maintained by Hühne et al. (2014), adopts a broader scope, cataloguing ageing factors across species and integrating data from multiple upstream databases.

A more recent development is the Human Aging and Longevity knowledge graph (HALD), described by Wu et al. (2023). HALD represents a step towards automated evidence integration, constructing a structured knowledge graph from text-mined data sources to support geroscience analyses. However, the text mining approaches employed in knowledge graph construction are typically limited to entity and relation extraction and do not produce the claim-level decomposition with quality grading that is required for a systematic evidence database.

The common limitations of these resources – beyond their dependence on manual curation or limited text mining – include the absence of systematic quality grading at the claim level, no decomposition of compound findings into atomic units, and a focus on the genomics and pharmacology of ageing rather than the broad range of lifestyle, nutritional, and clinical interventions that constitute the

longevity evidence base as a whole. The evidence database developed in the present study is designed to complement, rather than replace, these resources by providing a claim-level layer focused on extracted findings and evidence quality.

## 2.4 From Text to Structure: Extracting Biomedical Study Information

### 2.4.1 PICO Extraction and Its Limits

The extraction of structured information about clinical trials from unstructured text has been an active area of research in biomedical natural language processing for over a decade. The PICO framework – Population, Intervention, Comparator, Outcome – provides a natural organising schema for clinical evidence and has been widely adopted as the basis for annotation and extraction systems.

The EBM-NLP corpus, introduced by Nye and et al. (2018), provided the first large-scale annotated dataset for PICO extraction from biomedical abstracts, enabling the training and evaluation of supervised machine learning models. The corpus contains annotations of population, intervention, and outcome spans in 5,000 clinical trial abstracts, and has been widely used as a benchmark for structured information extraction systems. Trialstreamer, described by Marshall et. al (2020), extended this work to create a living, automatically updated database of clinical trial reports extracted from PubMed, demonstrating the feasibility of continuous automated literature surveillance.

However, PICO extraction as traditionally formulated is a span-labelling task – it identifies the textual spans corresponding to each PICO element but does not extract the effect direction (whether the intervention improved or worsened the outcome) or a structured representation of the finding. Furthermore, the diversity of language used to describe PICO elements in biomedical abstracts is vast, making generalisation across domains challenging. In the longevity domain specifically, the heterogeneity of interventions and outcomes is particularly marked, with hundreds of distinct exposures and outcome measures appearing across the literature.

Reason et al. (2024) provided a proof-of-concept demonstration of using generative artificial intelligence to extract over 680,000 PICO elements from clinical study abstracts at scale, identifying both the practical potential and the quality challenges of LLM-based extraction. Their work showed that while precision and recall were not perfect, the scale achievable through automated extraction was orders of magnitude greater than what was possible through manual curation, and that careful prompt design could substantially improve output quality.

### 2.4.2 Extracting Findings and Effect Direction: Evidence Inference

The Evidence Inference dataset, introduced by Nye and colleagues and extended in Evidence Inference 2.0 by DeYoung et al. (2020), moved beyond span labelling to address the task of extracting structured evidence statements – specifically, given an intervention, comparator, and outcome, what did the study find? The dataset provides annotations of evidence statements in the form of (intervention, comparator, outcome, finding direction) tuples, enabling models to be trained on the task of reading clinical trial full texts and extracting structured, directional findings.

This framing is directly relevant to the construction of claim-level evidence databases: by extracting not just what was studied but what was found, Evidence Inference-style extraction produces the kind of structured, directional claim that can be compared across studies and quality-graded according to the evidence that supports it. The limitation of Evidence Inference, as originally formulated, is that it requires full-text access and is designed for post-hoc annotation of a defined corpus, rather than large-scale, continuously updateable abstract-level surveillance.

The present study extends the Evidence Inference framing by retaining the intervention–outcome–direction structure and adding the information needed for evidence grading, including study design, effect size, statistical significance, sample size, population characteristics, and species. The Methods section implements this extension as two type-specific SEF schemas and a downstream ACU schema, summarised visually in Figs. 2 and 3 and listed in JSON format in Appendix B.

### 2.4.3 LLM-Based Structured Extraction at Scale

The advent of frontier large language models has transformed the feasibility of large-scale structured information extraction from biomedical text. Unlike task-specific models that require substantial labelled training data for each new domain, instruction-tuned large language models can be directed to extract structured output following a natural language schema specification, with performance that is often competitive with fine-tuned models on tasks for which limited training data are available (Ntinopoulos et al., 2025).

TrialSieve, described by Kartchner et al. (2025), demonstrated a comprehensive biomedical information extraction framework combining PICO extraction, meta-analysis support, and drug repurposing analysis using large language models. Ntinopoulos et al. (2025) evaluated large language models for data extraction from electronic health records, finding that frontier models performed well on structured extraction tasks with appropriate prompt engineering. Balasubramanian et al. (2025) demonstrated effective extraction of structured information from pathology reports using large language models, documenting both the potential and the quality control challenges of this approach.

The present study employs gpt-5.2 (OpenAI) for structured extraction, using a batch processing API to process the retained abstract corpus at scale. The choice of a frontier commercial model over a domain-specific fine-tuned model was motivated by two considerations: the breadth of the extraction schema (which encompasses study design, effect direction, effect size, population, and multiple outcome characteristics) and the diversity of the longevity intervention space (which includes hundreds of distinct interventions that would not be well-represented in biomedical NLP training corpora).

### 2.5 Claim Decomposition: Why Atomic Claims Matter

A fundamental challenge in evidence synthesis from biomedical abstracts is that individual papers typically contain multiple distinct claims, often conflated within a single paragraph or sentence. An abstract might report, for example, that "resistance training improved grip strength, gait speed, and quality of life in older adults with sarcopenia." This compound statement contains at least three distinct claims (resistance training versus grip strength; resistance training versus gait speed; resistance training versus quality of life), each of which may be supported by different levels of evidence, may behave differently under meta-analytic aggregation, and may have different clinical implications.

Claim decomposition – the task of breaking compound statements into minimal, self-contained atomic claims – has emerged in the natural language processing and fact-checking literature as a prerequisite for systematic evidence evaluation. Wanner et al. (2024) examined claim decomposition in the context of factuality evaluation, finding that atomisation of claims substantially improved the precision of fact-checking pipelines by enabling each claim to be evaluated against the most directly relevant evidence. Hu et al. (2025) examined the trade-offs introduced by claim decomposition, showing that whilst decomposition improved claim-level verification accuracy, it could also introduce errors when compound claims had non-compositional truth conditions.

Metropolitansky & Larson (2025) addressed the extraction and evaluation of factual claims, proposing a framework for quality-controlled decomposition that preserves causal and logical relationships between sub-claims. Zheng et al. (2026) demonstrated LLM-based atomic fact extraction and verification in a general domain, showing that chain-of-thought prompting could improve decomposition quality. In the biomedical domain, the additional requirement is that decomposed claims should align with the intervention–outcome structure of clinical evidence, so that each atomic claim unit corresponds to exactly one intervention affecting exactly one outcome in one study population.

In In this thesis, an ACU is defined as one analysis-ready factor–endpoint claim derived from one study: it links a single intervention or exposure to a single measured endpoint, assigns an effect stance such as benefit, harm, or no difference, and preserves the source publication and population context. The claim-decomposition logic is operationalised in Section 3.6, where Fig. 2 shows the interventional and observational SEF field structures and Fig. 3 shows the common ACU object; the detailed JSON schema summaries are provided in Appendix B.

### 2.6 Normalisation and Entity Linking in Biomedicine

Structured information extraction from biomedical text produces surface-level strings – the exact text of intervention and outcome mentions as they appear in abstracts – which must be normalised to canonical identifiers before claims from different papers can be compared. The same intervention may be referred to as “resistance exercise training”, “progressive resistance training”, “strength training”, or “weight-bearing exercise” across different publications. Without normalisation, these would appear as distinct entities in the evidence database, fragmenting what is actually a coherent body of evidence.

Medical Subject Headings (MeSH) (Lipscomb, 2000), maintained by the National Library of Medicine, is the most widely used controlled vocabulary for biomedical literature indexing. MeSH provides a hierarchical taxonomy of biomedical concepts that serves as a natural target for entity normalisation. The Unified Medical Language System (UMLS) (Bodenreider, 2004), also maintained by the National Library of Medicine, integrates over 150 biomedical vocabularies and ontologies into a unified framework, providing broader coverage and richer semantic relationships.

The challenge of synonym normalisation – mapping surface strings to canonical identifiers – has been addressed by a range of deep learning approaches. BioSyn, introduced by Sung et al. (2020), used synonym marginalisation to train entity representations that are robust to the synonymy problem. A subsequent refinement by Peng et al. (2022) extended this approach using interaction-based synonym marginalisation. Liu et al. (2021) introduced self-alignment pretraining for biomedical entity representations. BERN2 (Sung et al., 2022), an advanced neural biomedical named entity recognition and normalisation tool, integrated these advances into a practical system for end-to-end biomedical entity processing.

Benchmark evaluations of biomedical entity linking models have documented significant variation in performance across entity types and domains (Garda et al., 2023; Kartchner et al., 2023). ScispaCy, described by Neumann et al. (2019), provided a practical implementation of biomedical NLP tools built on the spaCy framework.

In the longevity domain, the entity normalisation challenge is particularly acute because the intervention space includes hundreds of compounds, dietary patterns, exercise modalities, and lifestyle factors – many of which have inconsistent MeSH coverage or lack standardised nomenclature entirely. The approach adopted in the present study combines a pre-built assignment lookup with a two-stage LLM (gpt-5-mini) taxonomy classifier to map surface strings to a fixed group-node hierarchy, using a curated YAML taxonomy as the canonical reference framework. This approach is described in detail in Section 3.7.

### 2.7 Claim Verification and Evidence Quality Assessment

The task of linking claims to supporting literature – determining whether a stated claim is supported, refuted, or neither supported nor refuted by a body of evidence – has been formalised as scientific claim verification or biomedical fact-checking and has attracted considerable research attention.

SciFact, introduced by Wadden et al. (2020), provides a dataset of scientific claims derived from citation sentences in biomedical papers, each paired with evidence from cited abstracts and a support/refute annotation. FEVER (Thorne et al., 2018), the broader fact extraction and verification benchmark, established the general framework for multi-document evidence-based claim verification.

In the health domain, HealthFC, described by Vladika et al. (2024), specifically targets health-related claims with evidence-based factchecking, and Sarrouti et al. (2021) addressed evidence-based factchecking of health-related claims with a focus on consumer health questions. Pradeep and colleagues et al. (2020) applied the VERT5ERINI system to scientific claim verification, demonstrating that generative approaches combining retrieval and generation could outperform earlier pipeline methods.

The binary support/refute framing of these systems – adequate for general factchecking – is insufficient for the needs of a longevity evidence database, where the relevant question is not simply whether a claim is true or false, but how strongly and directly it is supported by evidence of different qualities. The present study adopts a different framing: rather than verifying claims against a fixed corpus of evidence, it extracts claims from a systematic literature search and grades their quality according to the design, endpoint, effect, population, and replication information available in the source abstracts.

### 2.8 Towards an Updatable Claim-Level Evidence Database for Longevity Research

The preceding sections have established the state of the field across several relevant dimensions: the methodological challenges of longevity research, the endpoint hierarchy and its implications for

evidence quality, the landscape of existing longevity databases, and the technical toolkit for automated extraction, normalisation, and claim evaluation. A synthesis of these strands reveals the gap that the present study addresses.

No systematic, quality-graded, claim-level evidence database for human longevity interventions exists. Existing curated databases are invaluable but static, manually maintained, and not designed for claim-level decomposition or evidence quality grading. Evidence Inference-style systems can extract directional findings from full texts but are not designed for prospective large-scale surveillance. PICO extraction tools produce structural descriptions of trials but do not capture effect direction or quality. These limitations motivate the pipeline developed in this thesis.

The approach developed in the present study integrates automated retrieval, LLM-based relevance screening, structured ACU extraction using a frontier language model, multi-stage quality filtering, taxonomy-based entity normalisation, polarity correction, LLM-based claim validation, hallmark mapping, and a multi-dimensional evidence tiering framework into an end-to-end pipeline capable of generating, at scale, a quality-graded database of longevity claims from the published biomedical literature. The following Methods section describes this pipeline in detail.

# METHODS

## 3.1 Study Design and Overall Pipeline Architecture

The codebase for this thesis is available at https://github.com/arnasstucinskas/longevityevidence. An end-to-end natural language processing pipeline was designed and implemented to extract, filter, validate, and quality-grade longevity-related evidence claims from the biomedical literature. The final analysis combined two local runtime directories, `data/20260219_144933` and `data/20260403_054512`; these directories contain generated run outputs. The consolidated outputs were deduplicated by ACU identifier and publication record identifier, then written to `analysis/tiered_acus.jsonl` and `analysis/tiered_acus_merged.jsonl`.

Figure 1 summarises the final thesis pipeline from PubMed retrieval to quality-graded ACUs and merged claim pairs. Counts refer to the consolidated final analysis files unless otherwise specified. Numbers correspond to the step labels in Methods Sections 3.2-3.15. Blue boxes denote main pipeline processing stages, green boxes denote final outputs, and the yellow box denotes supporting publication selection and bibliometric enrichment rather than an ACU filtering step. The 29,332 type-routed extraction records are not a mutually exclusive publication count: they reflect routing into interventional and observational extraction streams, whereas 28,721 is the deduplicated retained-publication count after screening. Some figure steps combine adjacent methods sections where they implement one conceptual operation: Step 2 combines abstract normalisation and relevance screening, Step 7 combines entity quality filtering and taxonomy normalisation, and Step 11 combines evidence-score calculation with assignment of A/B/C/D quality tiers.

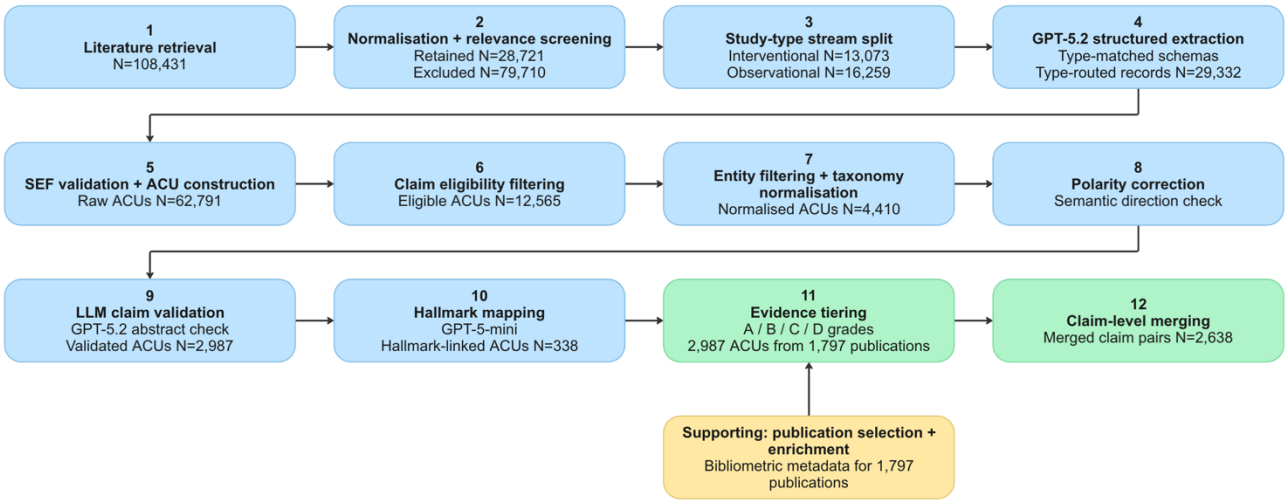


*Fig. 1. Final thesis pipeline from PubMed retrieval to quality-graded ACUs and merged claim pairs.*

### 3.2 Literature Retrieval (Step 1)

Publications were retrieved from PubMed using the National Center for Biotechnology Information Entrez application programming interface (API). A broad, high-recall query was constructed to capture human longevity and healthy-ageing research, combining ageing/longevity terms, intervention and exposure domains, and ageing-relevant outcome terms. The query was restricted to English-language publications and to publication types consistent with primary empirical research. Included publication types were Clinical Trial (all phases I-IV), Controlled Clinical Trial, Randomised Controlled Trial, Pragmatic Clinical Trial, Equivalence Trial, Adaptive Clinical Trial, Observational Study, Comparative Study, Multicenter Study, and Twin Study. Excluded publication types included reviews, meta-analyses, protocols, editorials, guidelines, case reports, retractions, and expressions of concern. The main retrieval was performed on 19 February 2026 and an incremental update was performed on 3 April 2026. Together, these runs retrieved 108,431 records, each stored with PMID, title, abstract, publication types, MeSH terms, DOI, PMCID, journal, publication year, and author metadata. The exact PubMed query is provided in Appendix A.

### 3.3 Abstract Normalisation (Step 2a)

Retrieved abstracts were subjected to text normalisation prior to screening (`s01_02_clean`). HTML entities and special characters were removed using rule-based processing. Whitespace sequences were collapsed to single spaces and Unicode characters were standardised to ensure consistent text encoding before downstream processing. This step produced clean, uniformly encoded abstract text whilst preserving all substantive content.

### 3.4 LLM-Based Relevance Screening (Step 2b)

The full corpus of 108,431 records was subjected to automated relevance screening. All runs used the Qwen 2.5-7B model (model identifier: qwen2.5-7b-gate), a 7-billion-parameter open-weight language model deployed locally via the Ollama inference server (Yang et al., 2024). Qwen 2.5-7B was selected for this first-pass screening task because it could be run locally at low cost, supported sufficient instruction-following for binary KEEP/EXCLUDE decisions, and avoided sending the full retrieval corpus to a commercial API before the more expensive structured-extraction stage. Five concurrent inference workers processed abstracts in parallel.

Each record was classified as KEEP or EXCLUDE, accompanied by a primary reason label drawn from a defined exclusion reason taxonomy. The screening criteria were designed to be inclusive (high recall, low precision), retaining records if they appeared to report original findings about any factor affecting any ageing-relevant outcome in humans. Key exclusion categories included: purely mechanistic or in vitro studies without direct human relevance; studies that exclusively examined age

as a biological variable without any modifiable factor; studies restricted exclusively to paediatric populations; and editorials, reviews, protocols, or commentaries without original data. The screening prompt did not require studies to demonstrate clinical significance or any threshold of evidence quality – these criteria were applied at later filtering stages.

Following screening, 28,721 records were retained (26.5% of the retrieved corpus), with 79,710 records excluded. This is a deduplicated publication-level count; the larger 29,332 count reported for Step 3 refers to type-routed extraction records after some retained publications were eligible for more than one extraction stream.

### 3.5 Interventional/Observational Stream Split (Step 3)

Screened records were partitioned into two parallel processing streams based on the reason label assigned during screening (`s01_03_split`). Records identified as reporting interventional or experimental designs were routed to `abstracts_interventional.jsonl`, whereas records reporting observational or epidemiological associations were routed to `abstracts_observational.jsonl`. Across the consolidated runs, 13,073 records were routed to the interventional stream and 16,259 to the observational stream. The stream total is larger than the retained-publication count because records could carry labels that made them eligible for more than one extraction pathway. This split enabled distinct extraction schemas optimised for each study type.

### 3.6 Structured Extraction: SEFs and ACU Construction (Steps 4-5)

#### 3.6.1 Structured Extraction Frame Definition

The retained and type-routed records were submitted to gpt-5.2 (OpenAI) via the OpenAI batch processing API for structured information extraction (`s01_04_extract_batch` and subsequent s02 parsing stages). The batch API was used to parallelise extraction across the retained corpus at reduced per-token cost relative to synchronous inference. In total, 29,332 type-routed extraction records were processed across the interventional and observational schemas.

A Structured Extraction Frame (SEF) was defined as the per-abstract extraction output – a structured JSON object capturing all comparisons or associations described within a single abstract. For interventional studies, each SEF contained a list of comparisons, each describing one intervention arm versus one comparator arm, with associated effect fields for each measured outcome. In observational studies, each SEF captured primary exposure variables and their associated endpoints.

SEFs represent the raw, abstract-level extraction unit: a single SEF corresponds to one publication and may contain multiple comparisons or associations.

The SEF schema, implemented in `schemas/interventional.py` and `schemas/observational.py` and summarised in Appendix B, required that every named entity (intervention, exposure, endpoint) be accompanied by a canonical label (the model's normalised term) and a `raw_string` anchor (the verbatim text from the abstract supporting that entity). This raw-string anchoring was essential for subsequent hallucination detection.

Two SEF schema variants were defined: an interventional SEF schema capturing comparator structure, randomisation, allocation concealment, arms, interventions, and endpoints; and an observational SEF schema capturing exposure contrast, model adjustment, comparison scope, exposures, covariates, and endpoint associations. Fig. 2 presents both SEF field structures; yellow shading marks fields shared across the interventional and observational schemas. These extraction schemas are distinct from the downstream ACU schema, which is constructed deterministically from SEFs and contains one factor, one endpoint, one stance, effect fields, study-design metadata, population descriptors, raw-string anchors, and source identifiers. Appendix B provides the corresponding JSON-format schema summaries.

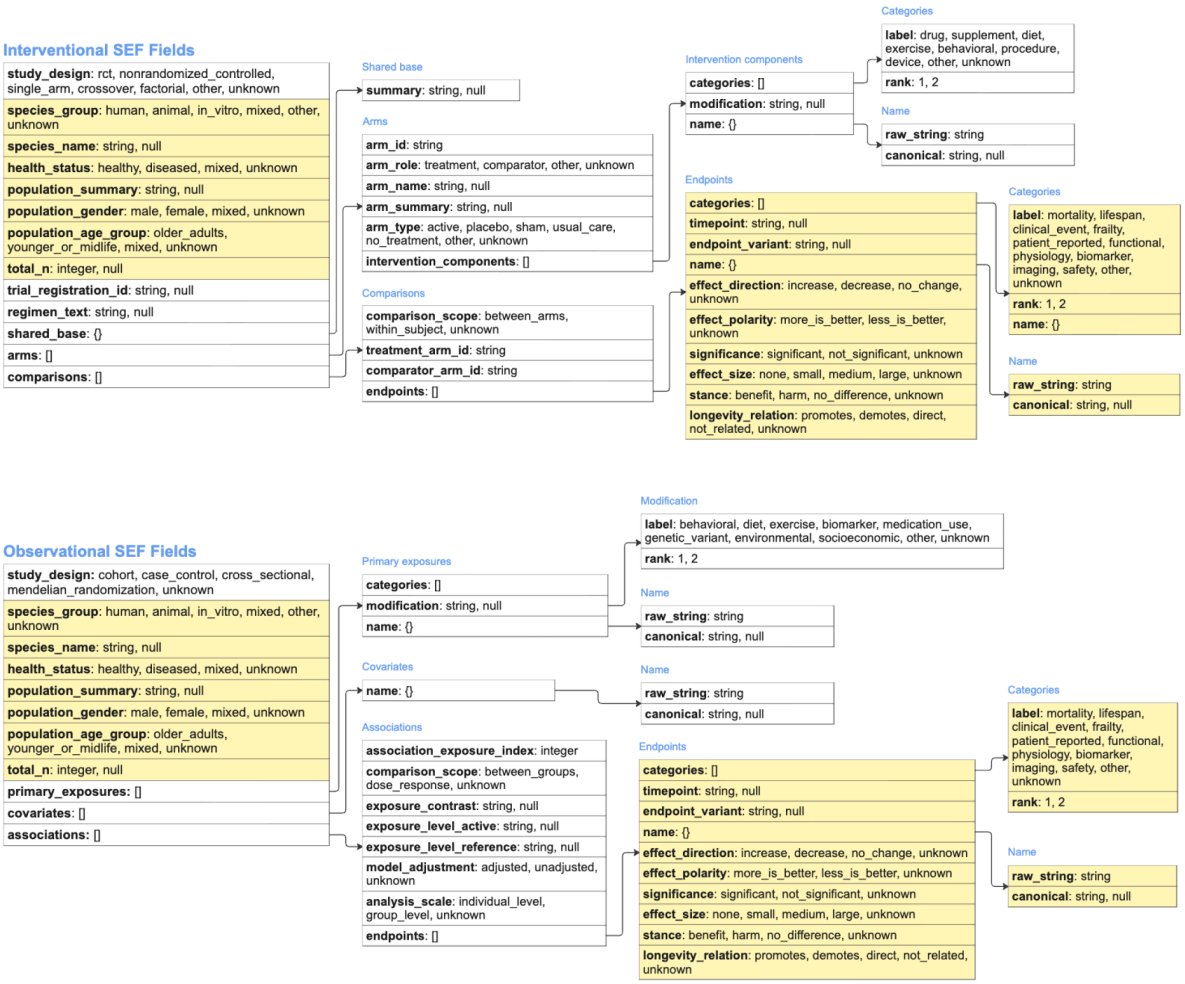


*Fig. 2. Interventional and observational SEF field schemas. Yellow shading denotes fields shared by both SEF variants.*

SEF is too broad to use directly as the unit of evidence, because a single abstract can describe multiple arms, comparisons, endpoints, and directional findings. The pipeline therefore converts each SEF into one or more ACUs. Each ACU represents one analysis-ready claim: a normalised factor, a normalised endpoint, an effect direction or stance, and the metadata needed for filtering, merging, and evidence scoring. Table 1 shows this conversion for one extracted endpoint finding.

*Table 1. Worked example showing how one endpoint finding from a study-level SEF is converted into a single ACU.*

| Field | Example value |
|---|---|
| Source record | pubmed:31557380 |
| Study type / design | Interventional / randomised controlled trial |
| Raw factor string | 1,700 mg/day metformin |
| Raw endpoint string | thigh muscle mass |
| Position within SEF | One endpoint finding within one study comparison |
| Normalised ACU factor node | Metformin |
| Normalised ACU endpoint node | Upper leg muscle mass |
| Corrected effect direction / stance | decrease / harm |
| Endpoint tier and category | E3; Imaging and neuro/structural measures |

### 3.6.2 Batch Polling and Output Parsing

Upon submission, the pipeline polled the OpenAI batch API at regular intervals until all batch jobs reported a terminal status (completed, failed, or expired) (`s02_01_extract_batch_check`). Completed batch output files were retrieved via the Files API and written to disk as raw `output.jsonl` files within each batch directory.

Parsed SEF records were extracted from the raw batch outputs (`s02_02_extract_batch_parse`). Each line of the batch output corresponded to one submitted abstract; the parser extracted the structured SEF JSON from the model's response and wrote it to `sef_interventional.jsonl` or `sef_observational.jsonl` as appropriate.

The structured-extraction workflow therefore followed a fixed sequence: retained abstract → study-type split → type-specific schema → OpenAI batch extraction → output parsing → raw-string validation → ACU construction. The interventional schema emphasised intervention arms, comparator arms, randomisation, and comparison scope. The observational schema emphasised exposure variables, model adjustment, analysis scale, and exposure-outcome association. This separation reduced ambiguity in the extraction prompt and allowed downstream scoring to treat causal and associational evidence differently.

### 3.6.3 SEF Validation

Raw-string validation was used as diagnostic quality control rather than as a filtering step. The script `s02_03_validate_raw_strings` checked whether each SEF `{canonical, raw_string}` span appeared as a verbatim substring of the cleaned source abstract after whitespace normalisation. Mismatching SEFs were reported and valid-subset files were produced, but mismatches were not excluded from final ACU construction because many failures reflected punctuation, abbreviation, Unicode, or whitespace differences; unsupported claims were removed later during LLM-based claim validation. Across records with an available source abstract, 13,088 of 13,331 SEF records passed all raw-string checks (98.2%) and 243 failed at least one check (1.8%); at the span level, 261 of 57,605 raw strings failed the substring check (0.5%).

### 3.6.4 ACU Construction from SEFs

Parsed SEFs were processed by two parallel ACU construction modules (`s02_04_build_acus_interventional, s02_05_build_acus_observational`). The key distinction between SEFs and ACUs is one of granularity: a SEF is an abstract-level record

that may describe multiple comparisons, each involving multiple endpoints; an ACU is a single (factor × endpoint) pair extracted from one comparison within one SEF.

Throughout this thesis, the term factor is used to encompass both interventions (experimental manipulations applied in controlled trials, such as a drug or exercise programme) and exposures (variables measured observationally, such as dietary patterns or lifestyle behaviours). This unified term reflects the pipeline's processing of both interventional and observational study designs within a single evidence database.

For interventional studies, each comparison in a SEF was expanded by taking the Cartesian product of interventions and endpoints: a comparison reporting one intervention against three outcomes produced three ACUs. For observational studies, each association object linked one indexed exposure to each of its endpoints, again producing one ACU per (exposure × endpoint) pair. Tautological ACUs in which the factor canonical name was identical to the endpoint canonical name were discarded.

Each ACU inherited structured fields from its parent SEF: canonical factor and endpoint names (with raw string anchors), effect direction, effect polarity, statistical significance, effect size category, study design, health status, population age group, population gender, total sample size, and species. The unique ACU identifier was constructed deterministically from the record ID, factor name, and endpoint name. Fig. 3 summarises ACU fields and the fields specific to interventional and observational ACUs.

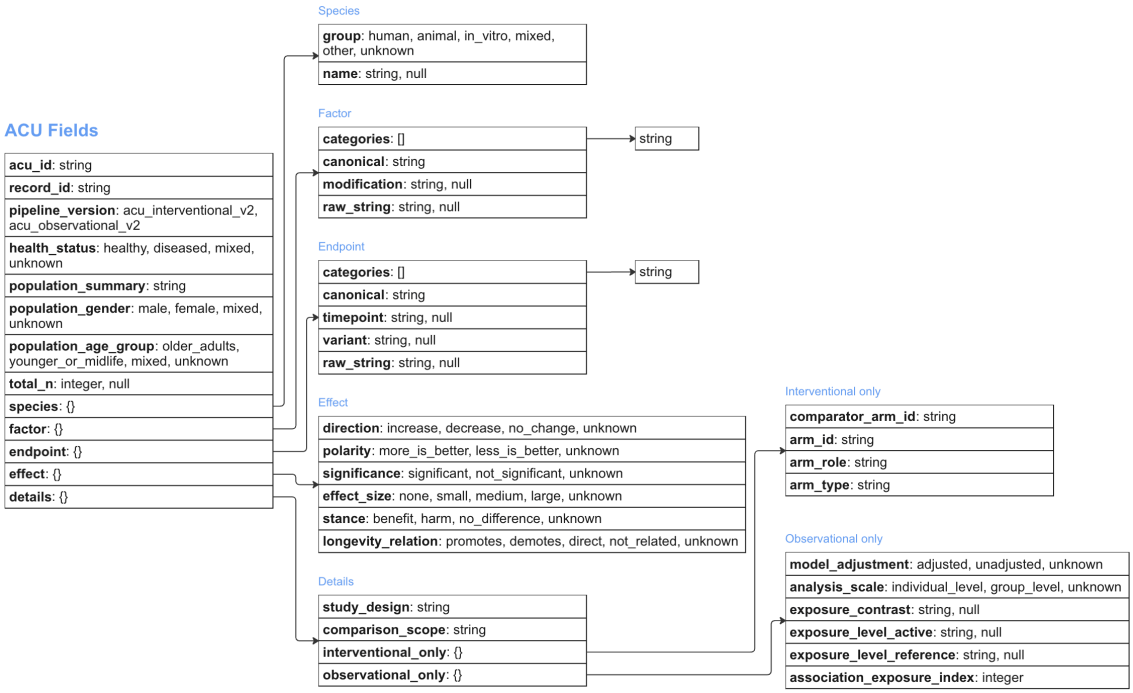


*Fig. 3. ACU field schema. The common ACU object stores publication identity, population context, species, factor, endpoint, effect, and design metadata, with additional details retained for interventional and observational streams.*

Each ACU inherited structured fields from its parent SEF: canonical factor and endpoint names (with raw string anchors), effect direction, effect polarity, statistical significance, effect size category, study design, health status, population age group, population gender, total sample size, and species. The unique ACU identifier was constructed deterministically from the record ID, factor name, and endpoint name.

The two ACU streams (`interventional` and `observational`) were merged into a single `acus_merged.jsonl` file (`s02_06_merge`). Raw extraction and ACU construction across the 28,721 retained records produced 62,791 ACUs. The mean number of ACUs per publication was 2.19.

### 3.7 ACU Claim Eligibility Filtering (Step 6)

The 62,791 raw ACUs were subjected to multi-criteria claim eligibility filtering to retain only human-relevant, directional, statistically supported claims (`s02_07_filter`). This stage retained ACUs for which the extracted effect was interpretable for downstream evidence assessment: human species, relevant health status, known effect direction and polarity, statistically significant effect, benefit/harm stance rather than no-difference or unknown stance, ageing relevance, and identifiable study design. ACUs meeting any of the following criteria were excluded:

1. **Non-human species:** ACUs where the species field indicated a non-human organism were excluded; only human studies were retained.
2. **Disease-specific populations**: ACUs where `health_status` was neither `healthy`, `mixed`, nor `unknown` were excluded, to focus the evidence database on prevention-relevant populations rather than disease treatment contexts.
3. **Unknown or absent effect direction:** ACUs where `effect.direction` was `unknown` or `none` were excluded.
4. **Non-significant or unknown significance:** ACUs where `effect.significance` was `not_significant` or `unknown` were excluded; only statistically significant claims were retained.
5. **No-difference or unknown stance:** ACUs where `effect.stance` was `no_difference` or `unknown` were excluded.
6. **Unknown effect polarity:** ACUs where `effect.polarity` was `unknown` were excluded.
7. **Not longevity-related:** ACUs where `effect.longevity_relation` was `not_related` were excluded.
8. **Absent effect size:** ACUs where `effect.effect_size` was `none` were excluded.

9. **Unknown study design:** ACUs where `details.study_design` was `unknown` were excluded.
10. **Out-of-scope endpoint:** ACUs with the endpoint value "falls" were excluded due to the specialist clinical context of falls prevention.

This stage reduced the ACU corpus from 62,791 to 12,565 ACUs (retention rate: 20.0%).

### 3.8 Entity Export and LLM-Based Entity Quality Filtering (Step 7a)

#### 3.8.1 Entity Export

Unique canonical factor and endpoint strings were extracted from the 12,565 filtered ACUs and written to `unique_effectors.txt` and `unique_endpoints.txt` (`s02_10_export_entities`). These lists contained every distinct canonical surface string appearing in the filtered corpus, in order of first occurrence.

#### 3.8.2 LLM-Based Entity Quality Filtering

Both factor and endpoint canonical strings were submitted to gpt-5.2 (reasoning effort: medium) in chunks of 200 terms per request for eligibility and specificity review (`s02_11_filter_entities`).

For factor terms, the filtering prompt instructed the model to retain a factor only if it was self-sufficient, identifiable without additional context, and plausibly health-relevant (i.e., something a longevity-focused reader could reasonably evaluate as healthy or harmful). Factors that were control/comparator labels (e.g., "placebo", "usual care"), underspecified (e.g., "treatment", "unspecified"), or purely administrative process labels were excluded.

For endpoint terms, the filtering prompt instructed the model to retain an endpoint only if it was relevant to longevity, healthspan, or ageing biology and had interpretable directionality (i.e., it was possible to judge whether more or less of the endpoint was desirable). Endpoints describing administrative or process outcomes (e.g., programme attendance, length of hospital stay), clearly exogenous injuries unrelated to ageing, or purely idiosyncratic goal-tracking measures were excluded.

The output of this stage was a `unique_effectors_keep.txt` list of approved factor strings and a `unique_endpoints_keep.txt` list of approved endpoint strings. Audit logs recording keep/drop decisions per chunk were written to `llm_filter_effectors_audit.jsonl` and `llm_filter_endpoints_audit.jsonl`.

### 3.8.3 ACU Filtering by Approved Entities

ACUs were retained only if both the factor canonical name appeared in `unique_effectors_keep.txt` and the endpoint canonical name appeared in `unique_endpoints_keep.txt` (`s02_12_filter_acus_by_entities`). ACUs failing either criterion were discarded. This step removed ACUs associated with underspecified or control-label factors and ACUs with endpoints outside the longevity-relevant scope.

## 3.9 Entity Taxonomy Normalisation (Step 7b)

Factor and endpoint canonical strings in the filtered ACU set were mapped to a structured group-node taxonomy to enable systematic comparison across ACUs that used different surface forms for the same concept (`s02_13_group`). A surface form is the raw phrase extracted from the abstract; a canonical string is the cleaned model-normalised phrase before final grouping; and a group-node taxonomy is the curated vocabulary in which related terms are organised into high-level groups and normalised nodes. For example, raw phrases such as "1,700 mg/day metformin" and "metformin therapy" can resolve to the factor node "Metformin", while endpoint phrases such as "thigh muscle mass" can resolve to "Upper leg muscle mass".

Normalisation proceeded in two stages. Before this step, the top-level effector and endpoint category lists were defined from earlier pipeline outputs and then treated as fixed: the normalisation step was not allowed to create new categories. A small initial set of canonical nodes was also built based on earlier outputs. First, each canonical string was looked up against assignment tables (`resources/effector_assignments.tsv` and `resources/endpoint_assignments.tsv`), which were empty at the beginning of the normalisation process and were then populated as terms were assigned. These TSV files contain exact-match mappings from canonical surface strings to group-node assignments within the taxonomy. Previously assigned terms can therefore be resolved deterministically in later runs without repeated LLM calls.

For terms not found in the assignment tables, a two-stage LLM classification procedure using gpt-5-mini was applied:

1. Gating call: the model received the list of top-level group names from the taxonomy YAML and selected the candidate groups most likely to contain the new term, narrowing the search space for the final assignment.

2. Decision call: the model received the candidate group definitions and existing nodes and assigned the term to an existing node or created a new node under an existing group. New groups could not be created, preserving the fixed taxonomy structure.

The final consolidated evidence database contains 793 normalised factor nodes and 769 normalised endpoint nodes across 2,987 ACUs. Example factor nodes include resistance training, vitamin D supplementation, metformin, Mediterranean diet, and cognitive training. Example endpoint nodes include gait speed, grip strength, all-cause mortality, inflammatory biomarkers, biological age clocks, and global cognitive function. Taxonomy normalisation groups terminology; it does not reinterpret effect direction, which is handled by the separate polarity correction step described below.

### 3.10 Polarity Correction (Step 8)

Endpoint taxonomy normalisation can invert the semantic meaning of an extracted effect direction. For example, a raw endpoint phrase such as "muscle loss" may be normalised to the positive endpoint node "Muscle mass". In that case, a decrease in the raw endpoint phrase is not equivalent to a decrease in the normalised node; the direction must be flipped before evidence tiering and benefit/harm interpretation.

Polarity correction was implemented in two stages. First, `s02_14_annotate_polarity_flip.py` read `resources/endpoint_assignments.tsv` and used gpt-5-mini to label each endpoint-term-to-node assignment as same, flipped, or unclear. Same means that an increase in the raw endpoint term has the same direction as an increase in the normalised endpoint node; flipped means the two directions are opposite; unclear means the relationship is too ambiguous for automated correction. Across 8,301 endpoint-assignment rows, 7,264 were labelled same (87.51%), 520 flipped (6.26%), and 517 unclear (6.23%).

Second, `s02_15_polarity_appending.py` applied these labels to grouped ACUs. ACUs mapped to same endpoints were retained unchanged and marked with the polarity-adjustment label. ACUs mapped to flipped endpoints had `effect.direction` reversed (`increase` ⇄ `decrease`) and `effect.polarity` reversed (`more_is_better` ⇄ `less_is_better`). ACUs mapped to unclear endpoints were excluded from downstream analysis. The corrected output was written to `grouped_acus_fixed.jsonl`. In the main run, 3,547 of 3,977 grouped ACUs were retained after polarity handling; in the incremental run, 29 of 33 grouped ACUs were retained. This step is a semantic normalisation safeguard, not a biological effect-size model.

### 3.11 LLM-Based Claim Validation (Step 9)

After polarity correction, each normalised ACU was re-checked against its source abstract. The purpose of this step was to remove clear false-positive normalised claims introduced by extraction, taxonomy grouping, or polarity handling before hallmark mapping and evidence tiering.

`s02_16_llm_validation_batch.py` prepared validation requests from `grouped_acus_fixed.jsonl` and the retained abstracts in `screen_llm_keep.jsonl`. Each request contained the original title and abstract, the normalised factor node, the normalised endpoint node, and the corrected effect direction. GPT-5.2 classified each claim as `supported`, `not_supported`, or `unclear`. The prompt was high-recall: it accepted paraphrases and conceptually equivalent wording but rejected claims contradicted by the abstract, claims with the opposite direction, and claims about a different factor or endpoint. Batch outputs were checked by `s02_17_llm_validation_batch_check.py` and parsed by `s02_18_llm_validation_batch_parse.py`.

Supported and unclear claims were retained, while `not_supported` claims were dropped. Unclear claims were retained because this stage was designed as a consistency filter for clear false positives rather than a maximum-precision human adjudication layer. In the main run, 3,547 ACUs were prepared and parsed: 2,716 were `supported`, 244 `unclear`, and 587 `not_supported`, leaving 2,960 retained ACUs. In the incremental run, 29 ACUs were prepared and parsed: 28 were supported and 1 was `not_supported`. The output `grouped_acus_claim_validated.jsonl` became the input to hallmark mapping. This validation is an automated LLM consistency check against abstracts, not independent human validation and not full-text adjudication.

### 3.12 Hallmark of Ageing Mapping (Step 10)

To characterise the mechanistic landscape of the longevity evidence corpus, endpoints from claim-validated ACUs were mapped to the 12 hallmarks of ageing defined by López-Otín et al. (2023): genomic instability, telomere attrition, epigenetic alterations, loss of proteostasis, disabled macroautophagy, deregulated nutrient sensing, mitochondrial dysfunction, cellular senescence, stem cell exhaustion, altered intercellular communication, chronic inflammation, and dysbiosis. Mapping was performed by `s02_19_map_to_hallmarks.py` using `grouped_acus_claim_validated.jsonl` as input. Only endpoint groups with plausible mechanistic content were eligible for hallmark mapping: molecular biomarkers and immune function, biological age and ageing clocks, physiology and clinical risk factors, and imaging/neurostructural measures. Deterministic keyword rules were applied first for high-specificity signals such as telomere

terms, DNA methylation and epigenetic clocks, inflammatory markers, mitochondrial or oxidative-stress terms, autophagy terms, nutrient-sensing markers, DNA-damage terms, stem-cell terms, and immune-response terms. Remaining eligible endpoint nodes were submitted to gpt-5-mini with fixed hallmark definitions. Endpoints without a defensible mechanistic hallmark assignment were labelled Non-hallmark outcome rather than forced into the hallmark taxonomy.

*Table 2. Deterministic hallmark-mapping rule families.*

| Rule family | Example signal | Primary rationale |
|---|---|---|
| Telomere | telomere, telomerase | Directly maps to telomere attrition. |
| Epigenetic clocks | DNA methylation age, epigenetic clock, GrimAge | Directly maps to epigenetic alterations / biological age measurement. |
| Inflammation | CRP, interleukin, cytokine, inflammatory marker | Directly maps to chronic inflammation. |
| Mitochondria / oxidative stress | mitochondrial terms, reactive oxygen species, oxidative stress | Maps to mitochondrial dysfunction. |
| Nutrient sensing | insulin, HOMA-IR, glucose, IGF-1, mTOR, AMPK, sirtuins | Maps to deregulated nutrient sensing. |
| Autophagy | autophagy, LC3, ATG, mitophagy | Maps to disabled macroautophagy. |

### 3.13 Publication Selection and Bibliometric Enrichment (Supporting Step)

Following hallmark mapping, publications contributing at least one final ACU were selected (`s02_20_select_publications.py`). The record IDs appearing in `acus_hallmarked.jsonl` were collected and matched against the screened publication records (`screen_llm_keep.jsonl`); only publications with matching record IDs were written to `selected_publications.jsonl`. The consolidated final evidence database contains 1,797 unique contributing publications.

Selected publications were then enriched with bibliometric metadata from two external APIs (`s02_21_enrich.py`): OpenAlex, used as the primary source for citation counts, open-access status, venue information, author affiliations, and DOI; and Europe PMC, used as a secondary source for PMCID, open-access flag, and cross-validation. API requests were rate-limited and cached to avoid repeated queries for the same PMID. The enriched publication records were written to `selected_enriched_publications.jsonl` and provided the citation counts, journal information, and geographic data reported in the Results section.

### 3.14 Evidence Tiering Framework (Step 11)

A multi-dimensional weighted linear scoring model was developed to assign an evidence quality score to each ACU (`s02_22_tier.py`). The model comprised seven active scoring dimensions,

each normalised to a 0-1 subscore and combined using a weighted linear formula scaled to a 0-100 evidence score.

Weights were assigned by the author based on established principles of clinical evidence appraisal. Endpoint proximity received the highest weight (0.30) following the core principle of evidence-based medicine that outcomes directly meaningful to patients - survival and functional independence – carry greater inferential value than physiological surrogates or molecular biomarkers (Cummings & Kritchevsky, 2022). Study design quality (0.25) reflects the standard evidence hierarchy, in which randomisation is the primary defence against confounding. Effect credibility (0.18) captures statistical robustness, which is a necessary but secondary condition given that even well-powered studies may measure clinically distant endpoints. Remaining dimensions (comparison structure, confounding adjustment, sample size, population context) received smaller weights reflecting their incremental rather than primary contribution to evidence quality. The full weight vector sums to 1.00.

*Table 3. Evidence-tiering dimensions, weights, implementation, and rationale.*

| Dimension | Weight | Implementation | Rationale |
|---|---|---|---|
| Endpoint proximity | 0.30 | Group-level proximity score: E1=1.00; E2 groups=0.55-0.85; E3=0.35-0.40; E4=0.20. | Indirectness is the central geroscience problem; survival and clinical outcomes carry more inferential value than surrogates. |
| Study design | 0.25 | RCT=1.00; crossover/factorial=0.90; non-randomised controlled=0.75; cohort=0.65; cross-sectional=0.30; other design-specific values. | Randomisation and stronger designs reduce causal uncertainty. |
| Effect credibility | 0.18 | Product of significance, effect-size, and stance weights; after filtering, variation is mainly driven by effect-size category. | A statistically supported and larger effect is more credible than an unknown or small effect. |
| Comparison structure | 0.10 | Between-arm interventional comparisons score higher than within-subject or unknown comparisons; observational contrast scope is also scored. | Explicit comparators strengthen interpretation. |
| Confounding adjustment | 0.08 | For observational ACUs, adjustment and analysis scale are combined and capped at 0.90; interventional ACUs receive 1.00. | Adjusted individual-level observational analyses are less confounded than unadjusted/group-level analyses. |
| Sample size | 0.05 | Log-scaled total sample size: log10(n+1)/2 with type-specific defaults for missing n. | Larger studies receive incremental credit with diminishing returns. |
| Population context | 0.04 | Age group, gender balance, health status, and species are multiplied into one population relevance score. | Older or mixed human populations closer to the target use case receive more credit. |

**Endpoint proximity** (weight: 0.30) was scored according to a four-level tier (E1–E4): E1 (Longevity/Survival) scored 1.00; E2 outcomes scored between 0.55 and 0.85 depending on their clinical directness; E3 outcomes (physiological/imaging) scored 0.35–0.40; and E4 outcomes (molecular/ageing clocks) scored 0.20.

Study design quality (weight: 0.25) was scored according to the standard hierarchy of evidence. For interventional studies: randomised controlled trials scored 1.00; crossover and factorial trials scored 0.90; non-randomised controlled trials scored 0.75; single-arm studies scored 0.40. For observational studies: Mendelian randomisation scored 0.85; cohort studies 0.65; case-control studies 0.50; cross-sectional studies 0.30.

**Effect credibility** (weight: 0.18) combined statistical significance, effect size category, and stance direction into a composite credibility score. Significant results with large or medium effect sizes scored highest; results with unknown effect sizes scored lowest.

**Comparison structure** (weight: 0.10) distinguished between between-arm comparisons (higher score) and within-subject pre–post comparisons (lower score), reflecting the greater causal inferential strength of parallel-group designs.

**Confounding adjustment** (weight: 0.08, applied to observational studies only) distinguished between analyses that adjusted for relevant confounders and unadjusted analyses.

**Sample size, log-scaled** (weight: 0.05) awarded incremental credit for larger sample sizes using a logarithmic transformation.

**Population context** (weight: 0.04) awarded credit for ACUs from studies of older adults or mixed populations.

The composite score for each ACU was computed as the weighted sum of component subscores and then scaled to a 0-100 range. The final quality tiers were defined as A >=85, B 71-84, C 55-70, and D <55. These thresholds should be interpreted as pragmatic cut-offs for separating stronger, moderate, weaker, and low-quality evidence within this corpus, not as externally validated clinical thresholds. Because earlier claim eligibility filtering removed non-significant findings and unknown-stance claims, the retained ACUs were concentrated in the middle and upper parts of the score range. The tier boundaries were therefore chosen after inspecting the empirical score distribution, with the aim of reserving Tier A for the strongest evidence while still distinguishing lower-scoring retained claims. The evidence scoring thresholds and sensitivity scenarios are summarised in Appendix C.

### 3.15 ACU-Level Claim Scoring and Merging (Step 12)

Following individual ACU tiering, ACUs sharing the same normalised (factor, endpoint) taxonomy node pair were merged into unified claim records for replication and convergence analysis (`s02_22_tier.py`). A claim-level quality score was computed for each merged claim, incorporating five components: (1) best evidence score (weight 0.30), calculated as the top-two average of supporting evidence scores; (2) convergence score (weight 0.25), reflecting consistency of effect direction across supporting ACUs; (3) non-contradiction score (weight 0.20), penalising minority-stance evidence; (4) replication score (weight 0.15), a log-scaled measure of concordant supporting studies; and (5) design breadth score (weight 0.10), rewarding evidence that spans multiple study-design classes. This produced 2,638 merged claim pairs in the final consolidated corpus.

The final merged output contains 2,638 unique normalised (factor, endpoint) claim records from the 2,987 individual ACUs.

# RESULTS

## 4.1 Corpus Composition and Growth

The pipeline retrieved 108,431 records from PubMed across all years up to the 3 April 2026 incremental update. Following LLM-based relevance screening, 28,721 records were retained (26.5% retention rate), with 79,710 records excluded as irrelevant. Multi-stage ACU extraction, polarity correction, claim validation, hallmark mapping, and evidence tiering yielded 2,987 high-signal ACUs drawn from 1,797 unique publications. Publication screening and ACU extraction yields are summarised in Tables 4 and 5, respectively.

*Table 4. Publication screening yield.*

| Pipeline stage | Records | Stage yield |
|---|---|---|
| PubMed retrieval | 108,431 | - |
| After LLM screening | 28,721 | 26.5% of retrieved |
| Unique publications contributing >=1 ACU | 1,797 | 6.3% of screened; 1.7% of retrieved |

*Table 5. ACU extraction and filtering yield.*

| Pipeline stage | ACUs / claim pairs | % retained |
|---|---|---|
| Raw ACUs constructed (GPT-5.2 extraction) | 62,791 | - |
| After claim eligibility filtering | 12,565 | 20.0% |
| After entity filtering, polarity correction, claim validation, hallmark mapping, and tiering | 2,987 | 23.8% |
| Merged factor x endpoint claim pairs | 2,638 | - |

The temporal distribution of publications in the final corpus followed a pronounced upward trend, reflecting both the growth of ageing research as a field and the expanding scope of longevity-relevant clinical trials over the past two decades (Fig. 4). Although the final corpus also includes a smaller number of publications from the 1980s, 1990s and 2026, Fig. 4 shows the plotted 2000-2025 window. Within this window, publication counts were substantially higher after approximately 2010 and peaked in 2025.

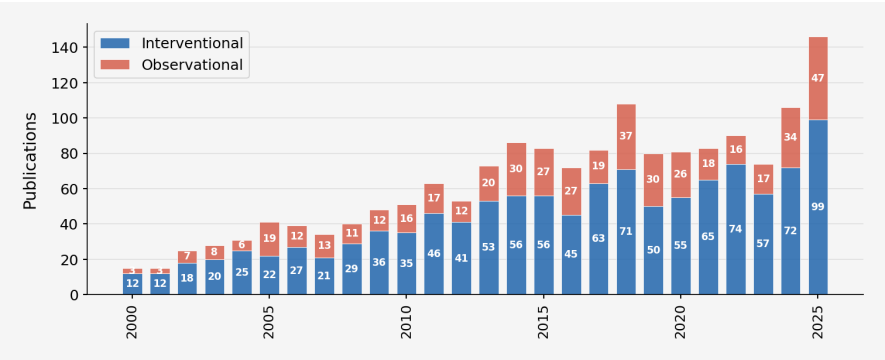


*Fig. 4. Annual number of publications contributing to the final ACU corpus for the 2000-2025 plotting window (n=1,647 plotted publications; 1,797 unique publications in the final corpus overall).*

The geographic distribution of the corpus was dominated by North American and European institutions, with the United States contributing the largest single share of first-author publications. The next largest first-author countries were the United Kingdom, Japan, Australia and China, with Germany, Canada and France contributing similar counts. These patterns show a broad but uneven international evidence base (Fig. 5).

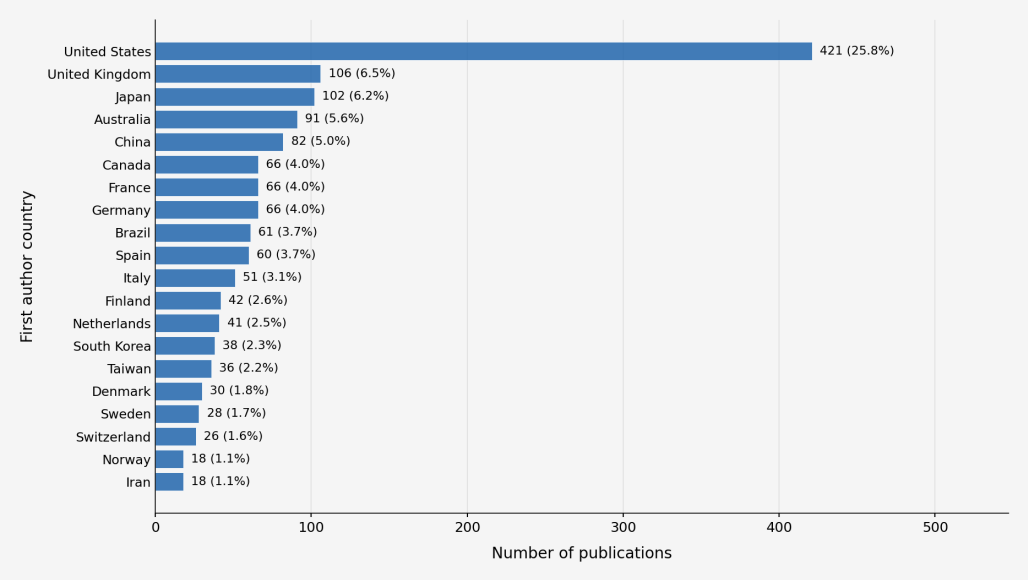


*Fig. 5. Geographic distribution of publications contributing to the final ACU corpus. The country of the first author's institution was used for attribution.*

The proportion of open-access publications in the corpus increased markedly over the study period, reflecting the broader open-access transition in biomedical publishing (Fig. 6).

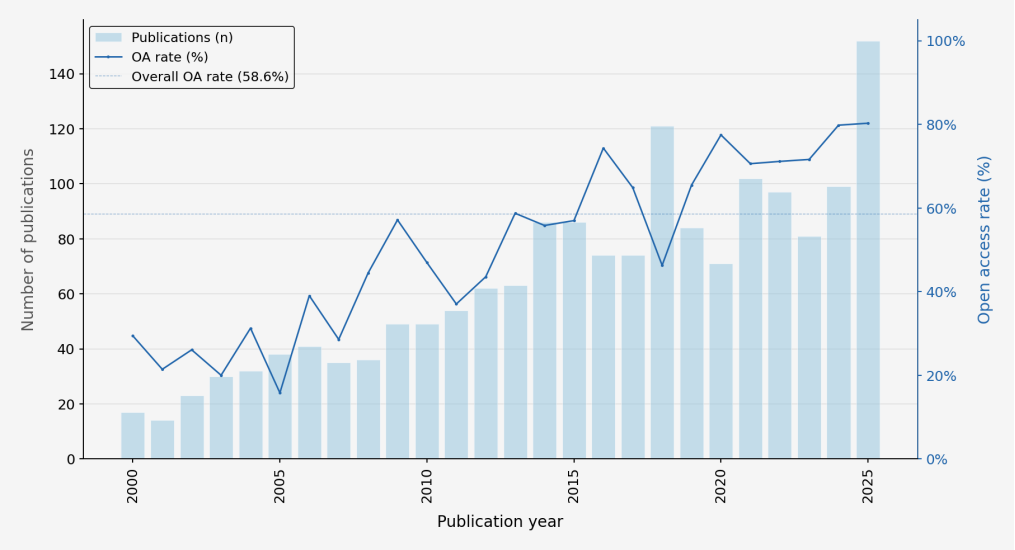


*Fig. 6. Open-access (OA) proportion of publications in the final ACU corpus by year. Bars show the number of contributing publications per year (left axis). The solid line shows the proportion of those publications available as open access (right axis). The dashed reference line indicates the overall open-access proportion across plotted years (58.6%).*

The distribution of ACUs per contributing publication was right-skewed: most publications contributed one or two ACUs (median = 1), whereas a small minority contributed five or more ACUs and four publications contributed ten or more (Fig. 7).

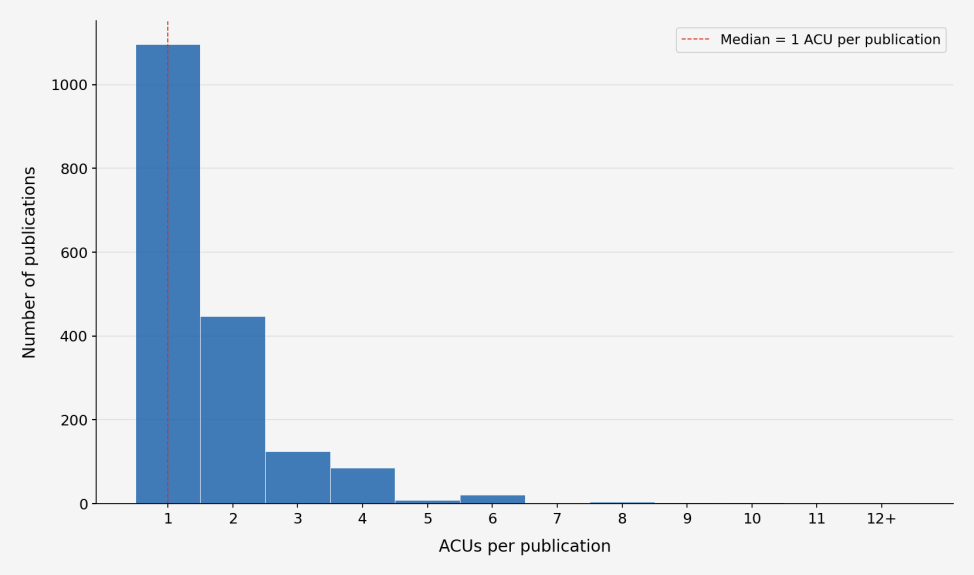


*Fig. 7. Distribution of the number of ACUs contributed per unique publication (N=1,797).*

The citation count distribution of publications in the final corpus was right-skewed: a small number of highly cited landmark studies accounted for a disproportionate share of total citations, while most contributing publications had low to moderate citation counts (Fig. 8). This is consistent with the broader power-law pattern of citation distributions in biomedical literature.

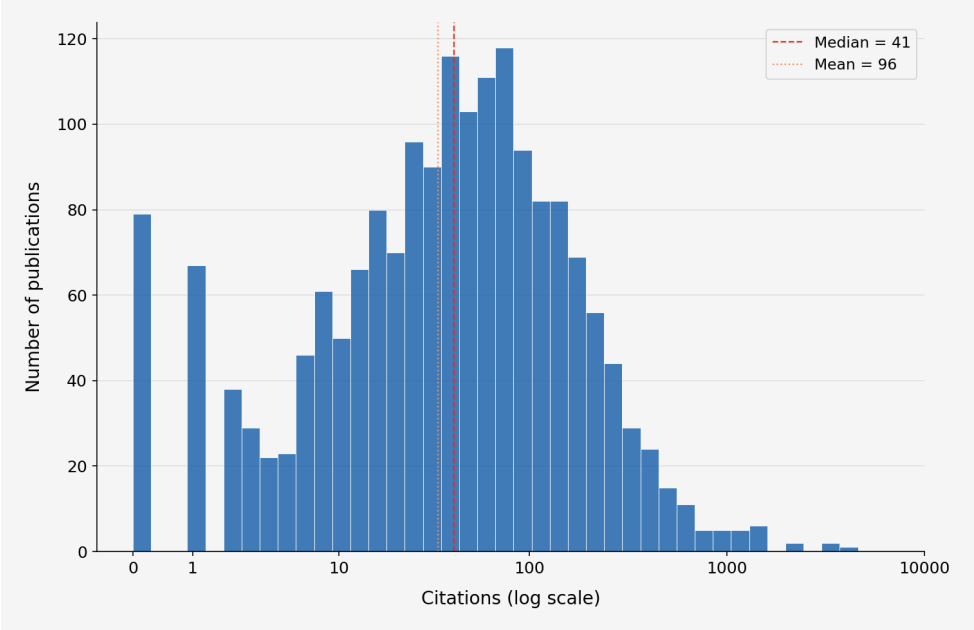


*Fig. 8. Distribution of citation counts for unique publications in the final ACU corpus (N=1,797).*

The top contributing journals are shown in Table 6. The Journal of the American Geriatrics Society and The Journals of Gerontology Series A contributed the most unique publications (73 and 70, respectively), corresponding to 108 and 109 ACUs. Nutrients ranked third, with 62 publications and 98 ACUs. ACU yield per publication was highest for The Journal of Nutrition, Health & Aging (1.94) and Experimental Gerontology (1.92), exceeding that of the leading geriatrics journals. Median citation counts were highest for The American Journal of Clinical Nutrition (99), the Journal of the American Geriatrics Society (81), and The Journals of Gerontology Series A (74), while several substantial contributors, including Nutrients, had markedly lower median citation counts.

*Table 6. Top 10 contributing journals by number of unique publications in the final ACU corpus.*

| Rank | Journal | Pubs | % Pubs | ACUs | % ACUs | ACU/Pub | Med. Citations |
|---|---|---|---|---|---|---|---|
| 1 | Journal of the American Geriatrics Society | 73 | 4.1 | 108 | 3.6 | 1.48 | 81 |
| 2 | The journals of gerontology. Series A, Biological sciences and medical sciences | 70 | 3.9 | 109 | 3.6 | 1.56 | 74 |
| 3 | Nutrients | 62 | 3.5 | 98 | 3.3 | 1.58 | 17 |
| 4 | The American journal of clinical nutrition | 55 | 3.1 | 84 | 2.8 | 1.53 | 99 |
| 5 | Experimental gerontology | 38 | 2.1 | 73 | 2.4 | 1.92 | 40 |
| 6 | BMC geriatrics | 33 | 1.8 | 44 | 1.5 | 1.33 | 27 |
| 7 | Age and ageing | 32 | 1.8 | 48 | 1.6 | 1.5 | 33 |
| 8 | The journal of nutrition, health & aging | 31 | 1.7 | 60 | 2.0 | 1.94 | 34 |
| 9 | PloS one | 31 | 1.7 | 53 | 1.8 | 1.71 | 39 |
| 10 | International journal of environmental research and public health | 31 | 1.7 | 56 | 1.9 | 1.81 | 13 |

### 4.2 Study Design Distribution

The study design composition of the final corpus reflected the dominance of interventional designs in ageing intervention research. Of the 2,987 ACUs, 2,288 (76.6%) were derived from interventional studies and 699 (23.4%) from observational studies. Within the interventional category, randomised controlled trials were the most common design, accounting for 1,592 ACUs (53.3% of all ACUs), followed by non-randomised controlled trials (332, 11.1%), crossover designs (166, 5.6%), single-arm studies (124, 4.2%), and factorial designs (52, 1.7%). Within the observational category, cohort studies contributed the largest share (393 ACUs, 13.2%), followed by cross-sectional studies (271, 9.1%), case-control studies (32, 1.1%), and Mendelian randomisation studies (3, 0.1%). The full design distribution is shown in Table 7.

*Table 7. Study design distribution of the final ACU corpus (N=2,987).*

| Design | n | % Total | Interventional n | Interventional % | Observational n | Observational % |
|---|---|---|---|---|---|---|
| RCT | 1,592 | 53.3 | 1,592 | 100.0 | 0 | 0.0 |
| Crossover | 166 | 5.6 | 166 | 100.0 | 0 | 0.0 |
| Factorial | 52 | 1.7 | 52 | 100.0 | 0 | 0.0 |
| Non-randomised controlled | 332 | 11.1 | 332 | 100.0 | 0 | 0.0 |
| Single arm | 124 | 4.2 | 124 | 100.0 | 0 | 0.0 |
| Cohort | 393 | 13.2 | 0 | 0.0 | 393 | 100.0 |
| Case-control | 32 | 1.1 | 0 | 0.0 | 32 | 100.0 |
| Cross-sectional | 271 | 9.1 | 0 | 0.0 | 271 | 100.0 |
| Mendelian randomisation | 3 | 0.1 | 0 | 0.0 | 3 | 100.0 |
| Other | 22 | 0.7 | 22 | 100.0 | 0 | 0.0 |
| Total | 2,987 | 100% | 2,288 | 76.6 | 699 | 23.4 |

The relationship between study design and evidence quality tier was notably structured. The heatmap (Fig. 9) shows the distribution of ACU counts across study designs and evidence quality tiers (A-D). The tiers used in this heatmap are defined as follows: Tier A, score >=85; Tier B, 71-84; Tier C, 55-70; and Tier D, <55. The detailed tiering formula is described in Methods Section 3.14, and the full tier distribution is shown in Table 8. Overall, 97.8% of Tier A ACUs derived from interventional studies, whereas 42.8% of Tier D ACUs did so. This gradient reflects the joint effect of the study-design quality dimension, which rewards randomised designs, and the endpoint-proximity dimension, which rewards clinically direct outcomes.

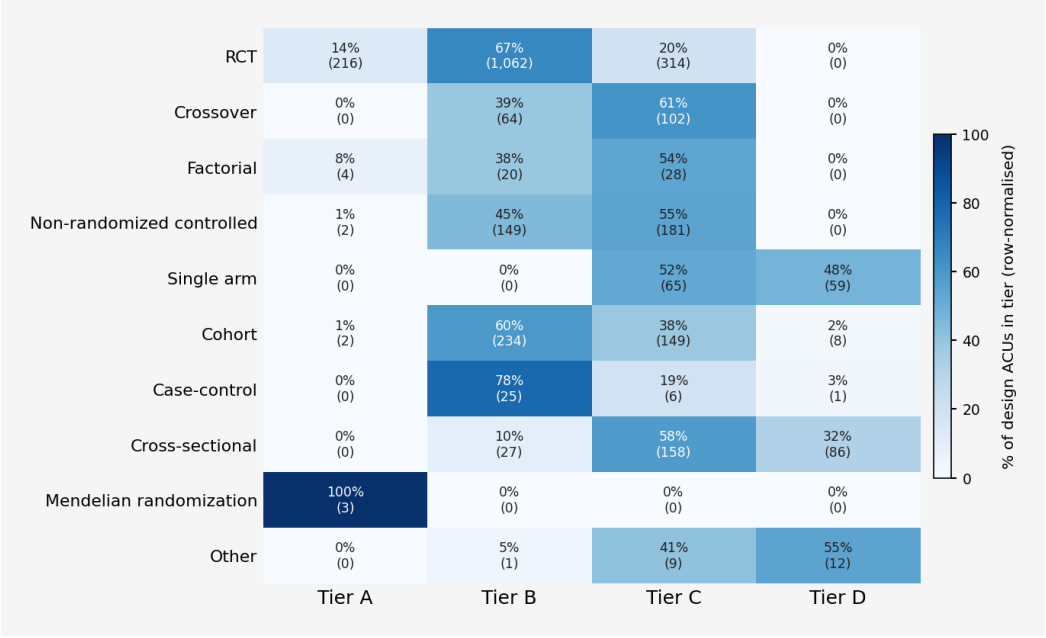


*Fig. 9. Heatmap of ACU count by study design (rows) and evidence quality tier (columns, A–D). Colour intensity indicates ACU count.*

The relationship between study design and endpoint proximity tier was also structured. Randomised controlled trials overwhelmingly measured E2 (clinical/functional) and E3 (physiological) outcomes, whilst survival (E1) endpoints were predominantly studied in cohort designs: of the 32 ACUs with E1 endpoints, only 9 (28.1%) derived from interventional studies. This reflects the practical impossibility of conducting randomised trials with mortality as the primary endpoint in most intervention contexts.

### 4.3 Evidence Score and Tier Distribution

The distribution of evidence scores across the 2,987 ACUs spanned a range of 42.40 to 95.50 points, with a mean of 72.31 (standard deviation 9.40) and a median of 72.63. The distribution was roughly normal with a slight left skew. The full tier distribution is summarised in Table 8 and Fig. 10.

*Table 8. Evidence quality tier distribution for the final ACU corpus (N=2,987).*

| Tier | Score range | n | % | Mean | SD | Median | Min | Max |
|---|---|---|---|---|---|---|---|---|
| A | ≥85 | 227 | 7.6 | 88.34 | 3.45 | 86.74 | 85.01 | 95.5 |
| B | 71–84 | 1,582 | 53.0 | 76.79 | 3.56 | 77.04 | 71.0 | 84.96 |
| C | 55–70 | 1,012 | 33.9 | 65.34 | 4.34 | 65.94 | 55.0 | 70.97 |
| D | <55 | 166 | 5.6 | 50.26 | 3.49 | 51.58 | 42.4 | 54.86 |
| All | 0–100 | 2,987 | 100.0 | 72.31 | 9.4 | 72.63 | 42.4 | 95.5 |

Tier B was the modal category, containing 1,582 ACUs (53.0%). Tier A, representing the highest quality evidence, contained 227 ACUs (7.6%). Tier A ACUs were predominantly from interventional studies (97.8%). Tier D, representing the lowest quality evidence in the corpus, contained 166 ACUs (5.6%), predominantly from observational studies (57.2%). The overall score distribution was approximately normal with a slight left skew, centred on the Tier B range.

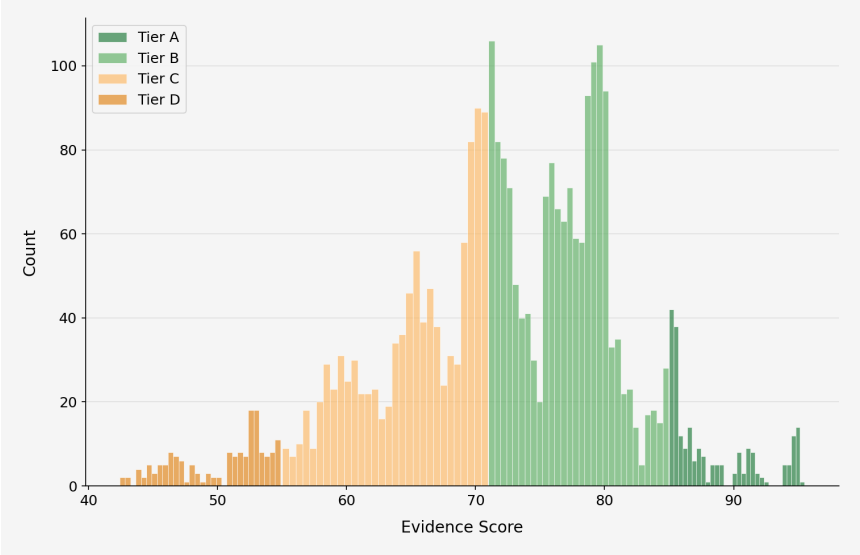


*Fig. 10. Distribution of evidence scores across the final ACU corpus (N=2,987), with evidence quality tier boundaries indicated.*

Correlations between evidence scoring dimensions were generally weak, except for the expected positive association between study design and adjustment (r = 0.62; Fig. 11). This stronger correlation reflects the scoring structure: interventional designs received full adjustment credit, whereas observational studies varied according to confounding adjustment. Endpoint proximity was weakly positively correlated with effect credibility (r = 0.22) and weakly negatively correlated with adjustment (r = -0.22). Study design was weakly negatively correlated with endpoint proximity and effect credibility (r = -0.11 and r = -0.08, respectively). Population relevance was largely independent of the other displayed dimensions (r = 0.02 to 0.07). These patterns indicate limited redundancy across most sub-scores, with the main dependence arising from the design-adjustment relationship.

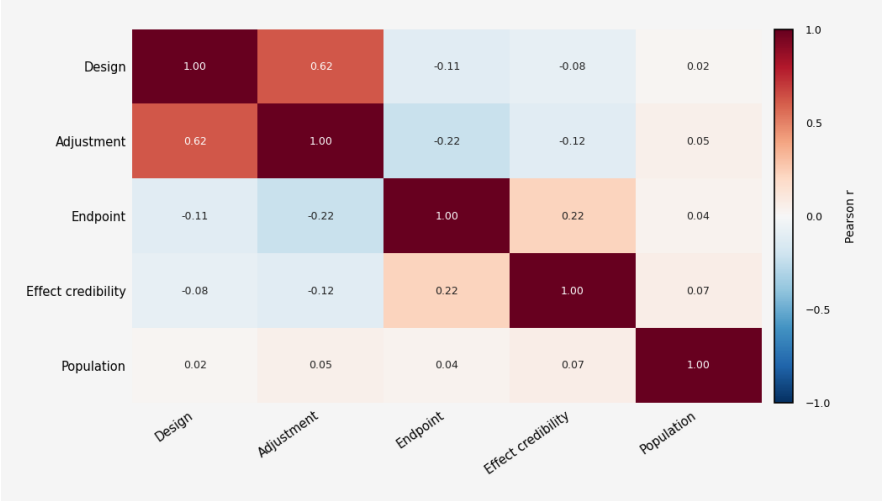


*Fig. 11. Pearson correlation heatmap between evidence scoring dimensions.*

### 4.4 Endpoint Proximity Distribution – The Proximity Gap

The distribution of ACUs across endpoint proximity tiers revealed a pronounced concentration of research at intermediate and surrogate endpoint levels. Of the 2,987 ACUs in the final corpus, only 32 (1.1%) targeted direct survival or longevity outcomes (E1). The majority of ACUs, 1,955 (65.5%), measured clinical or functional outcomes (E2). A further 728 ACUs (24.4%) measured physiological or imaging endpoints (E3), and 272 (9.1%) measured molecular or biological age endpoints (E4). This distribution, characterised in the present analysis as the proximity gap, reflects a fundamental structural feature of the longevity intervention literature (Table 9, Fig. 12).

*Table 9. Distribution of ACUs across endpoint proximity tiers.*

| Tier | Category | n | % |
|---|---|---|---|
| E1 | Longevity / Survival | 32 | 1.1 |
| E2 | Clinical / Functional | 1,955 | 65.5 |
| E3 | Physiological / Imaging | 728 | 24.4 |
| E4 | Molecular / Ageing clocks | 272 | 9.1 |
| Total | | 2,987 | 100 |

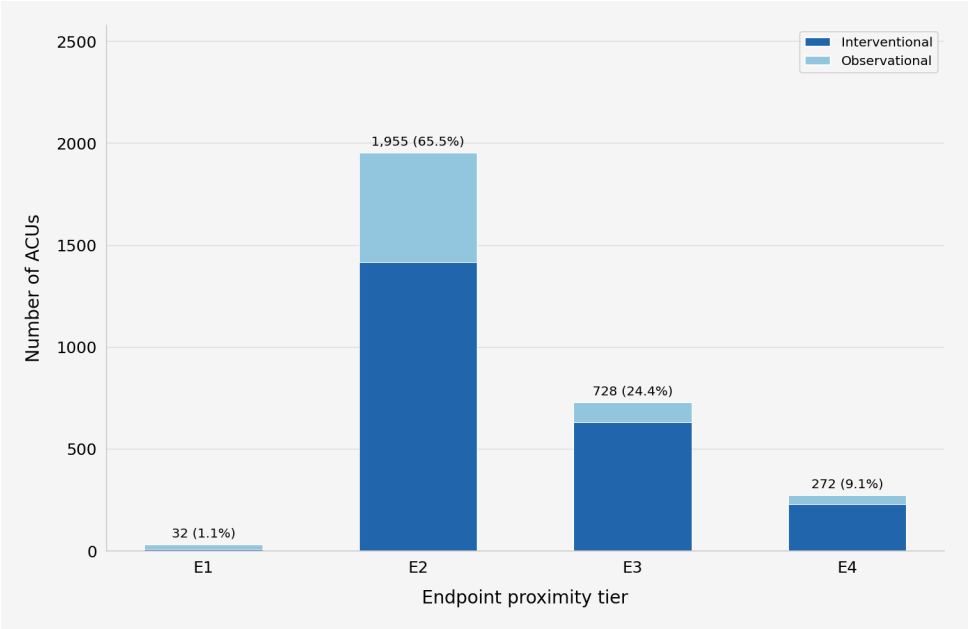


*Fig. 12. Distribution of ACUs across endpoint proximity tiers (E1–E4).*

Within the E2 category, physical function and performance constituted the largest endpoint group, accounting for 597 ACUs (20.0% of the total corpus), with a mean evidence score of 76.01 and a benefit rate of 96.5%. Physiology and clinical risk factors were the second largest endpoint group (450 ACUs, 15.1%), followed by clinical events and diagnoses (426 ACUs, 14.3%). The full endpoint group distribution is shown in Table 10.

*Table 10. Distribution of ACUs across endpoint groups.*

| Endpoint group | Proximity tier | n | % | Mean score | Benefit % |
|---|---|---|---|---|---|
| Physical function and performance | E2 | 597 | 20.0 | 76.01 | 96.5 |
| Physiology and clinical risk factors | E3 | 450 | 15.1 | 66.61 | 90.7 |
| Clinical events and diagnoses | E2 | 426 | 14.3 | 80.11 | 69.7 |
| Cognition and neuropsychological performance | E2 | 420 | 14.1 | 74.42 | 85.7 |
| Patient-reported outcomes and wellbeing | E2 | 310 | 10.4 | 73.64 | 89.7 |
| Imaging and neuro/structural measures | E3 | 278 | 9.3 | 65.97 | 94.6 |
| Molecular biomarkers and immune function | E4 | 228 | 7.6 | 60.81 | 93.9 |
| Frailty and geriatric syndromes | E2 | 120 | 4.0 | 74.53 | 75.0 |
| Disability, ADL/IADL, institutionalisation | E2 | 82 | 2.7 | 75.93 | 69.5 |
| Biological age and ageing clocks | E4 | 44 | 1.5 | 54.62 | 68.2 |
| Longevity / Survival | E1 | 32 | 1.1 | 82.73 | 87.5 |

The benefit direction varied systematically across endpoint proximity tiers. Physical function and performance ACUs, all in E2, had the highest benefit rate among the large endpoint groups (96.5%). Imaging and neuro/structural measures ACUs, which are E3 endpoints, also showed a high benefit rate (94.6%). By contrast, E1 (survival) ACUs showed a lower benefit rate of 87.5%, reflecting the inclusion of harmful exposures such as sedentary behaviour and tobacco smoking alongside beneficial interventions (Fig. 13). Biological age and ageing clock ACUs showed a

benefit rate of 68.2% with the lowest mean score (54.62). Physical function and cognition endpoint groups were dominated by randomised controlled trials, while survival (E1) and clinical events groups relied substantially on cohort study designs, reflecting the practical constraints on trial design when hard clinical endpoints are the target (Fig. 14).

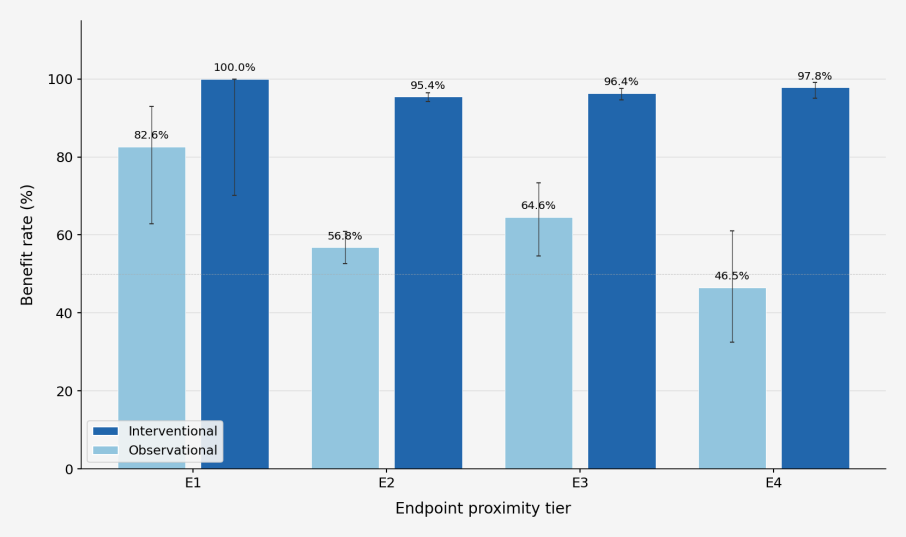


*Fig. 13. Benefit ratio (proportion of ACUs reporting beneficial effect direction) by endpoint proximity tier (E1–E4) and endpoint group.*

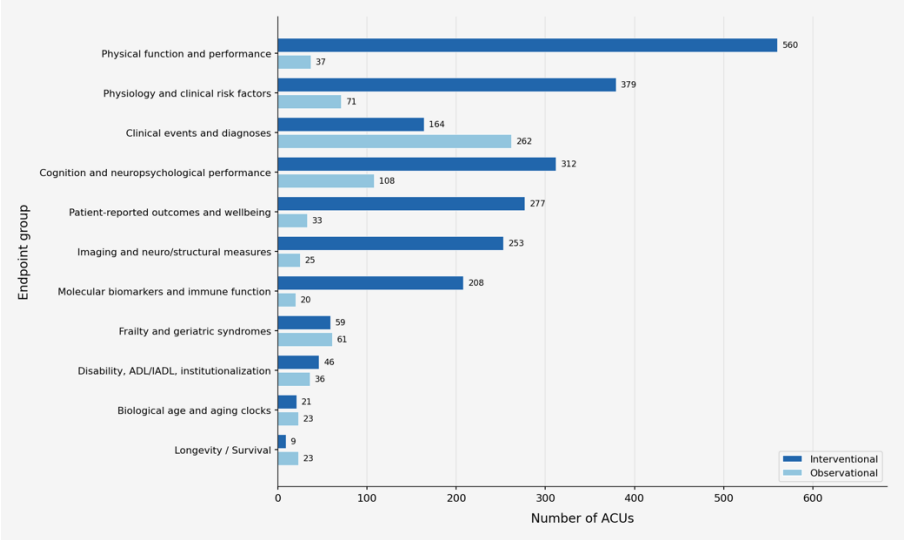


*Fig. 14. Breakdown of study design composition within each endpoint group.*

## 4.5 Population Characteristics

The population characteristics of the ACU corpus reflected the focus of longevity intervention research on older adult populations. The age group distribution showed the predominance of older adult populations, with a substantial proportion of studies conducted in mixed-age samples. Older adult populations were the most common study population in the corpus, studied across the full range of design types (Fig. 15).

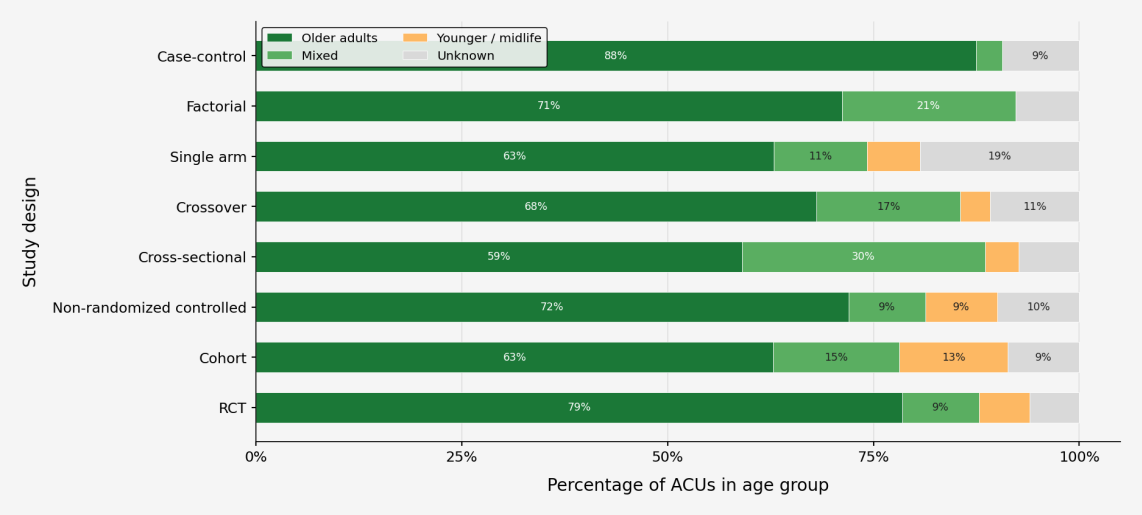


*Fig. 15. Distribution of study designs across population age groups.*

Sample sizes were highly variable and right skewed. Among ACUs with available sample-size data, the majority came from studies with fewer than 300 participants. Small randomised controlled trials (fewer than 100 participants) constituted a substantial proportion of the interventional ACU corpus, while large observational cohort studies contributed ACUs with sample sizes in the tens of thousands (Fig. 16).

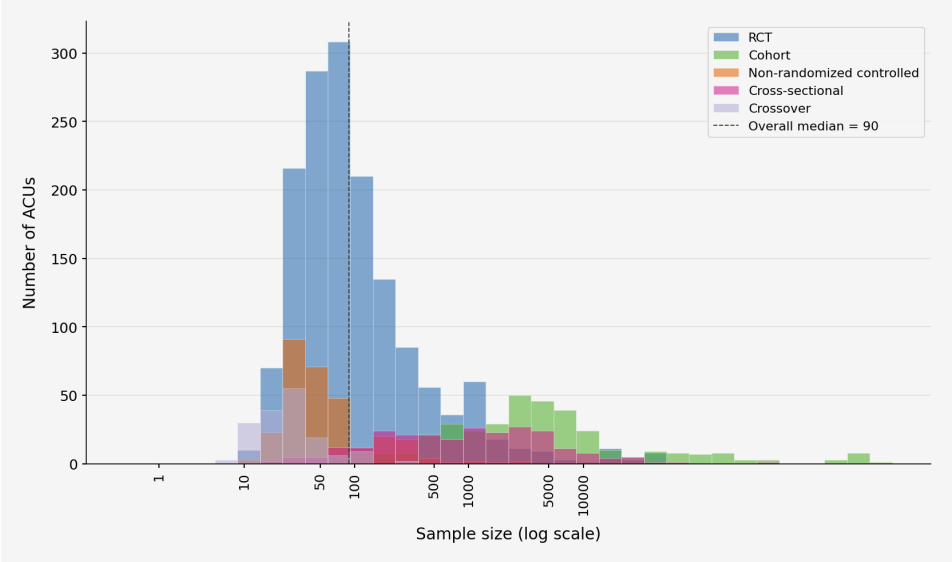


*Fig. 16. Distribution of study sample sizes (log scale) across ACUs with available sample-size data in the final corpus.*

## 4.6 Top Factors and Factor Categories

The final corpus covered 793 normalised factor nodes. The distribution of ACU counts across factors was highly skewed, with the most studied factors accounting for a disproportionate share of the evidence base. The top 20 factors by ACU count are shown in Table 11.

*Table 11. Top 20 factors by ACU count.*

| Rank | Factor | ACU count |
|---|---|---|
| 1 | Resistance training | 214 |
| 2 | Physical activity, exercise (general) | 141 |
| 3 | Aerobic endurance training | 65 |
| 4 | Cognitive training | 53 |
| 5 | Omega-3 fatty acids | 36 |
| 6 | Mediterranean diet | 34 |
| 7 | Multimodal exercise programme | 33 |
| 8 | Balance training | 30 |
| 9 | Alcohol consumption | 30 |
| 10 | High-intensity interval training | 29 |
| 11 | Tai chi | 25 |
| 12 | Social engagement | 23 |
| 13 | Exergaming | 23 |
| 14 | Memory strategy training | 20 |
| 15 | Dance | 19 |
| 16 | Calorie restriction | 18 |
| 17 | Balance and strength training | 18 |
| 18 | Walking programme | 17 |
| 19 | Resveratrol | 17 |
| 20 | Multivitamin / multimineral | 17 |

Resistance training was the single most frequently studied factor, contributing 214 ACUs – more than any other single factor and substantially more than the second-ranked general physical activity category (141 ACUs). The top 20 list was dominated by exercise modalities, reflecting the extensive research investment in exercise as a longevity intervention (Fig. 17).

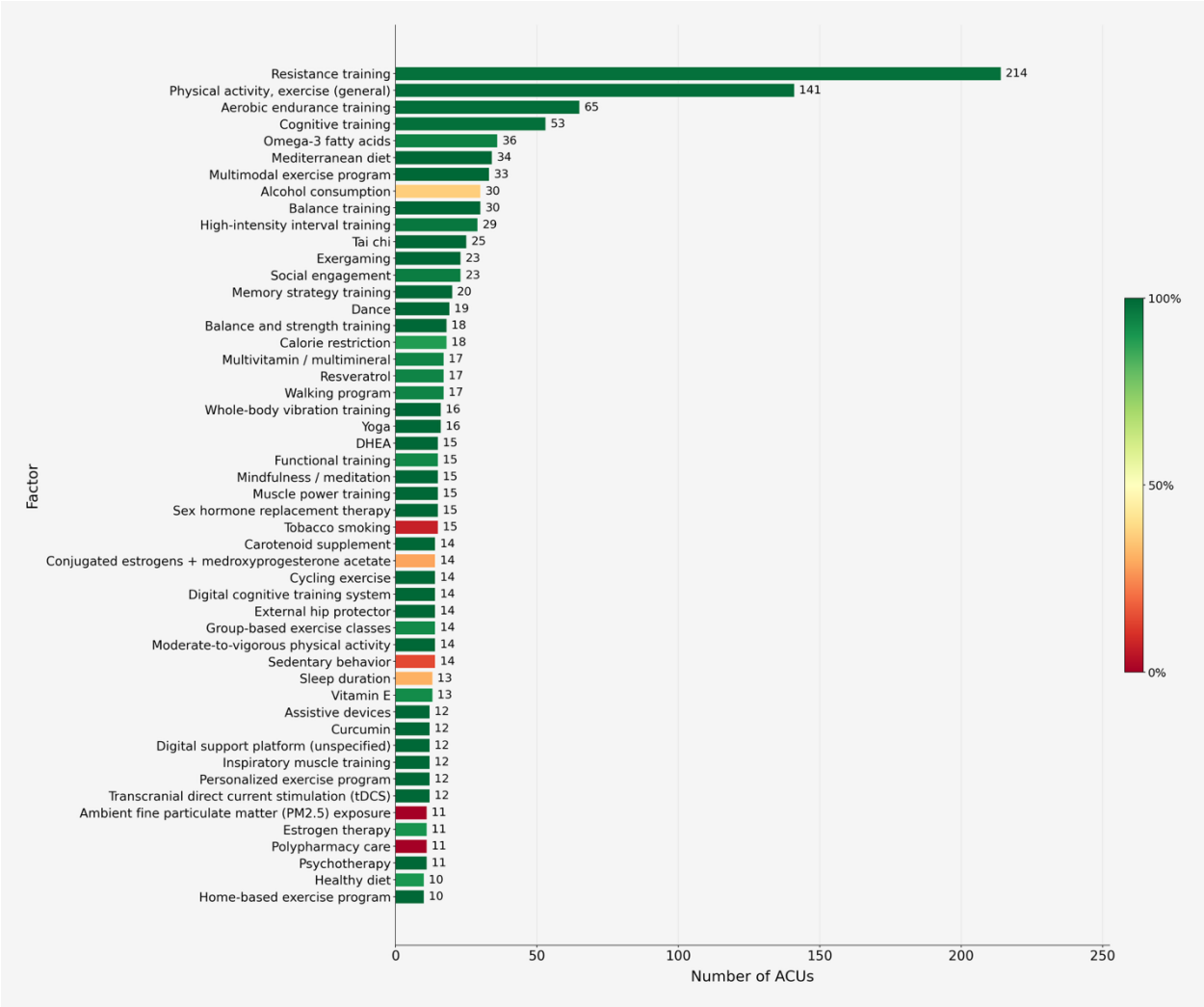


*Fig. 17. Top 50 factors by ACU count.*

At the category level, exercise and physical training constituted the largest factor category by far, accounting for 1,009 ACUs (33.8% of the total corpus) with a mean evidence score of 73.43 and a benefit rate of 96.2%. Supplements and nutraceuticals were the second largest category (463 ACUs, 15.5%), followed by medications and biologics (331 ACUs, 11.1%). The factor category distribution is shown in Table 12.

*Table 12. Top 7 factor categories by ACU count, mean evidence score, and benefit rate.*

| Category | n | % | Mean score | Benefit % |
|---|---|---|---|---|
| Exercise & physical training | 1,009 | 33.8 | 73.43 | 96.2 |
| Supplements & nutraceuticals | 463 | 15.5 | 70.42 | 96.8 |
| Medications & biologics | 331 | 11.1 | 72.3 | 65.0 |
| Diet patterns & eating schedules | 191 | 6.4 | 68.82 | 86.4 |
| Foods & beverages | 129 | 4.3 | 67.96 | 79.8 |
| Cognitive training & rehabilitation | 106 | 3.5 | 76.36 | 98.1 |
| Digital health & interactive tech | 95 | 3.2 | 74.82 | 97.9 |

The medications and biologics category had a markedly lower benefit rate (65.0%) than exercise (96.2%) or supplements (96.8%). This reflects the mixed directionality of pharmacological evidence in the corpus: some pharmacological agents appear as beneficial interventions in randomised trials, while others are documented as harmful exposures in observational studies (e.g., certain medication classes associated with adverse outcomes). It does not reflect a higher rate of null-effect findings, since non-significant ACUs were excluded during claim eligibility filtering. Among the seven largest categories in Table 12, cognitive training and rehabilitation showed the highest mean evidence score (76.36) and a near-universal benefit rate (98.1%).

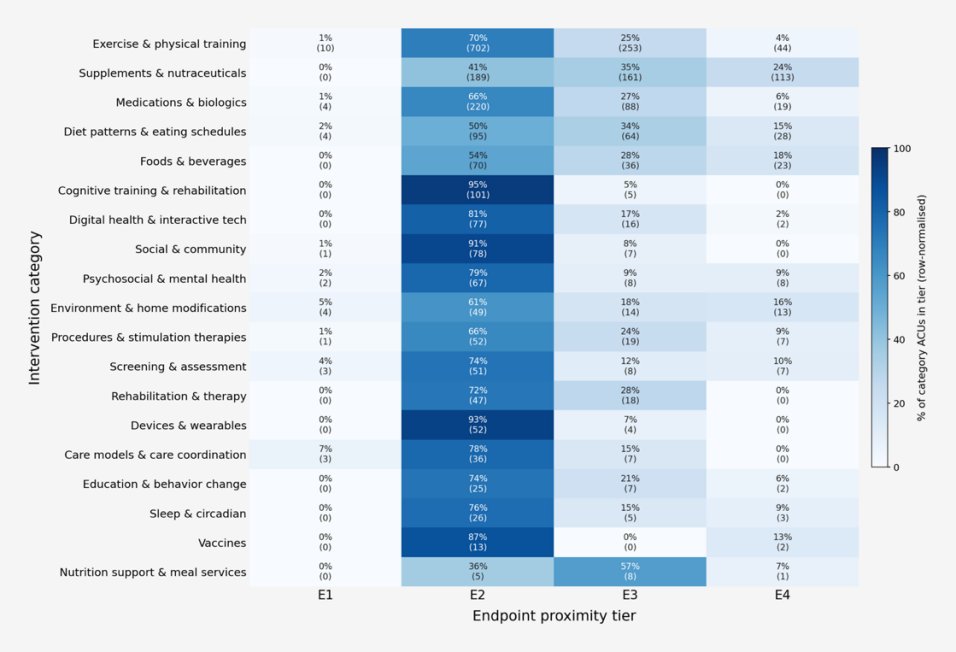


*Fig. 18. Factor categories stratified by endpoint proximity tier (E1–E4).*

The top factors by direct survival evidence (E1 endpoint proximity tier) are shown in Table 13. Physical activity and general exercise led with 6 E1 ACUs, all reporting benefit. Group-based exercise classes contributed 2 E1 ACUs with 100% benefit. Radiation exposure, sedentary behaviour, and tobacco smoking appeared as harmful exposures with exclusively harmful effect directions.

*Table 13. Factors with direct survival (E1) evidence in the final ACU corpus.*

| Factor | E1 ACUs | Benefit % | Best score |
|---|---|---|---|
| Physical activity, exercise (general) | 6 | 100.0 | 84.8 |
| Group-based exercise classes | 2 | 100.0 | 90.5 |
| Adherence to moderate-to-vigorous physical activity recommendations | 1 | 100.0 | 85.5 |
| Art activities | 1 | 100.0 | 90.5 |
| Beta-blockers | 1 | 100.0 | 86.8 |
| Blood transfusion dose | 1 | 0.0 | 84.4 |
| Calcium antagonists (calcium channel blockers) | 1 | 100.0 | 86.8 |
| Carvedilol | 1 | 100.0 | 86.3 |

| | | | |
|---|---|---|---|
| Childhood infectious disease epidemic exposure | 1 | 100.0 | 76.5 |
| Community paramedicine health promotion programme (CP@clinic) | 1 | 100.0 | 90.1 |
| Expressive writing (written emotional disclosure) | 1 | 100.0 | 90.5 |
| Flavonol intake | 1 | 100.0 | 66.3 |
| Group psychotherapy | 1 | 100.0 | 90.5 |
| Guide to Action for Care Homes fall prevention programme | 1 | 100.0 | 91.5 |
| Healthy lifestyle index | 1 | 100.0 | 78.0 |
| Lithium concentration in tap water | 1 | 100.0 | 67.2 |
| Mediterranean diet | 1 | 100.0 | 80.5 |
| Monounsaturated fat intake | 1 | 100.0 | 78.3 |
| Pole vaulting | 1 | 100.0 | 86.5 |
| Preventive care adherence programme | 1 | 100.0 | 90.0 |
| Radiation exposure | 1 | 0.0 | 80.1 |
| Sedentary behaviour | 1 | 0.0 | 78.8 |
| Time spent in physical activity | 1 | 100.0 | 78.6 |
| Tobacco smoking | 1 | 0.0 | 77.2 |
| Vasodilators | 1 | 100.0 | 86.8 |
| Western dietary pattern | 1 | 100.0 | 84.8 |

### 4.7 Evidence Replication and Convergence

The replication analysis assessed the degree to which individual (factor, outcome) claim pairs were supported by multiple independent studies. Of the 2,638 unique claim pairs in the merged corpus, 2,422 (91.8%) rested on a single study. A further 206 claim pairs (7.8%) were supported by 2-5 independent studies, 8 (0.3%) by 6-10 studies, and 2 (0.1%) by more than 10 studies. The full replication distribution is shown in Table 14 and Fig. 19.

*Table 14. Replication distribution for merged claim pairs (N=2,638).*

| Replication band | Merged claim pairs | % |
|---|---|---|
| Single-study (n=1) | 2,422 | 91.8 |
| 2-5 studies | 206 | 7.8 |
| 6-10 studies | 8 | 0.3 |
| >10 studies | 2 | 0.1 |

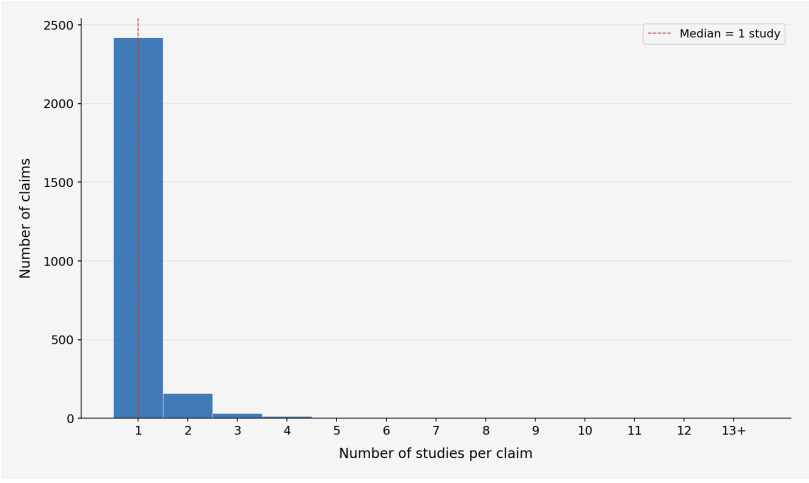


*Fig. 19. Distribution of unique claim pairs by number of supporting studies (replication count).*

The relationship between replication count and claim-level score was positive: claims supported by multiple studies scored higher on average, reflecting the positive contribution of the replication score component and the general pattern that more frequently studied factor–outcome pairs tend to involve more established designs and clinically meaningful outcomes (Fig. 20).

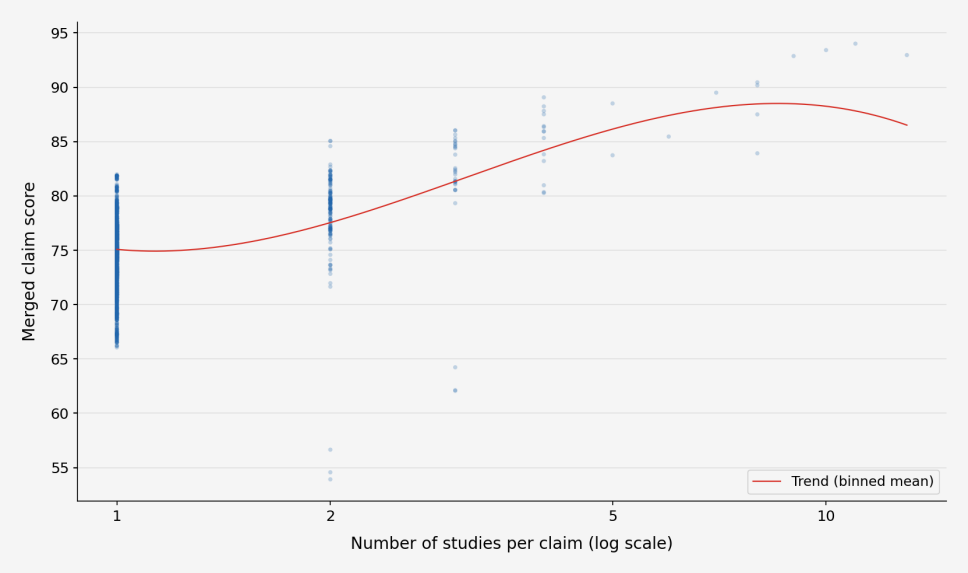


*Fig. 20. Mean evidence score by replication count (number of supporting studies per claim pair).*

The convergence score distribution for multi-study claim pairs showed that the great majority of replicated claims displayed consistent effect directions across supporting studies, with high convergence scores (>0.8) dominating the distribution (Fig. 21).

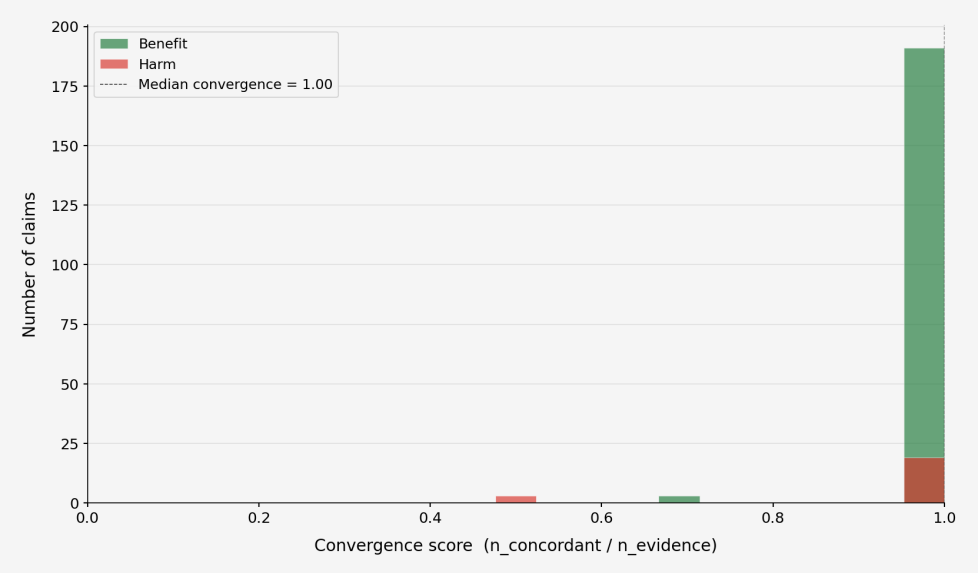


*Fig. 21. Distribution of within-claim convergence scores for multi-study claim pairs.*

Claims supported by evidence from multiple distinct study design types showed substantially higher mean scores than single-design claims, reflecting the design breadth component of the claim-level scoring model (Fig. 22).

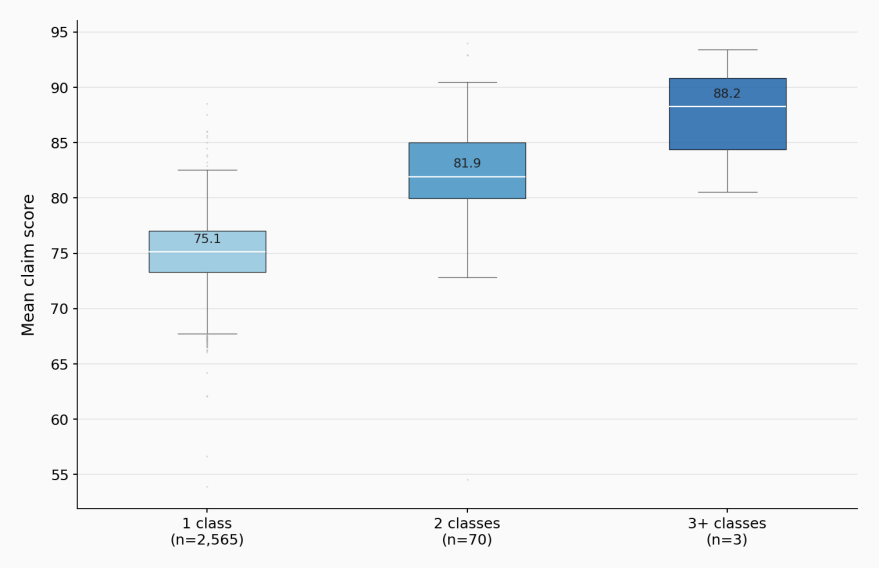


*Fig. 22. Mean evidence score by design breadth (number of distinct study designs among supporting studies). Claims with evidence from multiple design types show substantially higher scores.*

### 4.8 Hallmark of Ageing Coverage

The hallmark of ageing mapping revealed pronounced heterogeneity in the clinical evidence base across the twelve hallmarks. Hallmarks with well-established and easily measurable clinical proxies – particularly deregulated nutrient sensing and mitochondrial dysfunction – were represented by the largest numbers of hallmark-linked ACUs and high benefit ratios. By contrast, hallmarks that lack validated clinical endpoints – particularly cellular senescence, dysbiosis and disabled macroautophagy – showed sparse representation in the clinical ACU corpus (Fig. 23).

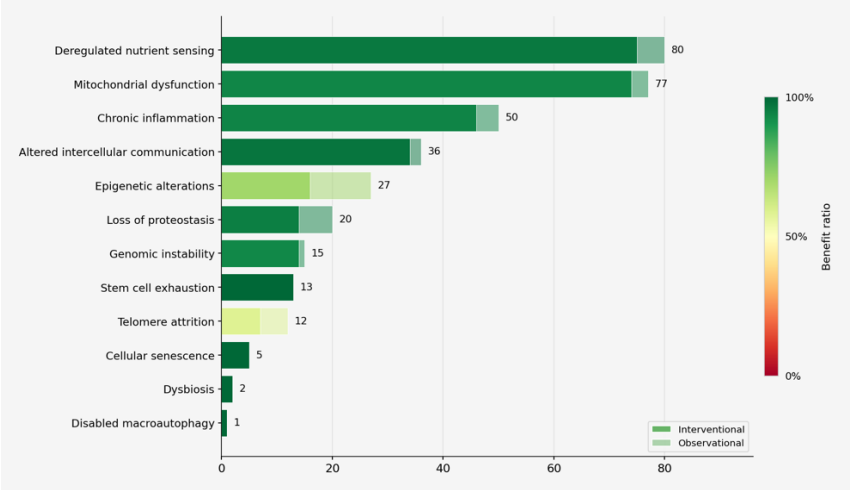


*Fig. 23. Coverage of the twelve hallmarks of ageing in the final ACU corpus.*

Chronic inflammation and altered intercellular communication – hallmarks that overlap with clinical immune function and inflammatory marker measurements – also showed substantial ACU representation. The epigenetic alterations hallmark was represented primarily by E4 (biological ageing clock) ACUs, reflecting recent clinical interest in epigenetic clocks as intervention response biomarkers.

# DISCUSSION

## 5.1 The Proximity Gap: A Structural Bias Towards Surrogate Endpoints

The most consequential finding of this analysis is the near-complete absence of direct survival evidence in the longevity intervention literature captured by this pipeline. Only 1.1% of the 2,987 high-signal ACUs in the final corpus targeted survival or longevity directly (E1 endpoint proximity tier). The remaining 98.9% measured intermediate endpoints – clinical function, physiological markers, or molecular biomarkers – none of which are guaranteed to translate into longer life. This "proximity gap" is not merely a statistical artefact but reflects a fundamental structural feature of the field: conducting trials powered to detect mortality effects of lifestyle or nutritional interventions is practically, ethically, and economically prohibitive for most academic research groups.

This finding closely mirrors concerns articulated by several groups working in geroscience methodology. Rolland et al., (2023) described the specific challenges of designing geroscience trials with meaningful endpoints, including the impracticality of survival as a primary outcome. Newman et al., (2016) argued that the field requires a new class of endpoints that bridge the gap between molecular biomarkers and survival. Cummings & Kritchevsky, (2022) proposed a tiered endpoint framework explicitly ranking survival above functional outcomes and functional outcomes above biomarkers – a hierarchy closely paralleling the E1–E4 endpoint proximity tier scheme employed in the present analysis.

The proximity gap has significant implications for the interpretation of the evidence database. The high benefit rates observed for exercise interventions (96.2% across all ACUs) and supplements (96.8%) are almost entirely driven by effects on physiological and functional endpoints. The 6 E1 ACUs for general physical activity (all reporting benefit) and 2 E1 ACUs for the group-based exercise classes (all reporting benefit) represent the strongest direct survival evidence in the corpus, but these figures are small relative to the overall volume of surrogate endpoint claims.

## 5.2 Lifestyle Factors Dominate; Supplements Remain Peripheral to the Survival Evidence Base

The dominance of exercise and physical training factors in the ACU corpus (33.8% of all ACUs) reflects both the practical feasibility of conducting exercise trials in older adult populations and the genuine strength of the mechanistic and clinical evidence for physical activity effects on ageing-relevant pathways. Resistance training alone contributed 214 ACUs.

In contrast, the supplement and nutraceutical category – which includes omega-3 fatty acids, vitamin D, resveratrol, and hundreds of other compounds that feature prominently in the popular longevity discourse – contributes 463 ACUs but with a different endpoint profile. Supplements are concentrated

in E3 and E4 endpoints (physiological biomarkers and molecular measures), with limited representation in the E1 and E2 clinical event tiers. The mean evidence score for the supplement category (70.42) was lower than for cognitive training (76.36), digital health (74.82), or exercise (73.43).

Parish et al., (2025) documented similar limitations in the DrugAge database, finding that longevity-relevant drug studies frequently lacked effect size reporting, used inadequate controls, and relied on model organism evidence with uncertain translational relevance. The present analysis reaches convergent conclusions through a complementary approach: the systematic extraction of human clinical trial evidence reveals that the supplement evidence base, whilst broad in terms of the number of factors tested, is shallow in terms of endpoint proximity and replication.

The medication and biologics category showed a lower benefit rate (65.0%) compared to lifestyle factors. This reflects the mixed directionality of pharmacological evidence: some agents appear as beneficial interventions in randomised trials, while others appear as harmful exposures in observational studies. The lower benefit rate does not indicate a higher frequency of null-effect findings, since ACUs with non-significant or no-difference effects were excluded during claim eligibility filtering.

### 5.3 The Corpus Is Broad but Shallow: Replication and the Evidence Quality Challenge

The most immediate practical limitation of the longevity evidence database is the extreme sparsity of replication. Nine in ten of the 2,638 unique (factor, outcome) claim pairs in the merged corpus rest on evidence from a single study. This finding is consistent with the broad-but-shallow characterisation of the longevity intervention literature: a very large number of distinct factor-outcome pairs have been investigated at least once, but very few have been subjected to independent replication.

This pattern directly reflects the concerns articulated by Ioannidis (2005) regarding the reliability of single-study findings in biomedical research. Single-study ACUs are assigned a replication score of zero in the claim-level scoring formula, structurally capping their claim-level scores relative to replicated claims. The two claim pairs with more than 10 supporting studies, and the three claim pairs with at least 10 supporting studies, all involved exercise or cognitive-training interventions on functional or cognitive outcomes.

The design breadth analysis reinforced this conclusion: claim pairs supported by convergent evidence from both randomised and observational designs scored substantially higher than those supported by a single design type. The convergence of randomised trial evidence (which establishes causal effect)

and observational cohort evidence (which establishes effect in naturalistic populations) represents a particularly strong evidential pattern that is rare in the current corpus.

### 5.4 Uneven Hallmark Coverage: Gaps Between Basic Science and Clinical Translation

The hallmark of ageing mapping revealed a pronounced mismatch between the mechanistic priorities of ageing biology and the clinical evidence base captured by this pipeline. Cellular senescence has emerged as one of the most exciting mechanistic targets in geroscience, with the first human trials of senolytic agents reporting promising results. Yet the clinical ACU corpus contains almost no evidence assigned to cellular senescence, reflecting the early stage of translational research in this area and the absence of validated clinical biomarkers for senescent cell burden.

Deregulated nutrient sensing is represented by a large body of clinical evidence largely because the metabolic outcomes associated with nutrient sensing – glycaemic control, insulin sensitivity, lipid profiles – are easily and reliably measured in clinical settings. The differential coverage of hallmarks thus reflects the availability of clinical measurement tools as much as the intrinsic importance of the hallmark.

This finding has practical implications for geroscience research prioritisation. The uneven coverage suggests that investment in the development of validated clinical biomarkers for underrepresented hallmarks – particularly cellular senescence and macroautophagy – could substantially expand the evidence base for interventions targeting these mechanisms.

### 5.5 Strengths and Limitations

The principal strength of the present work is its scale and systematicity. The retrieval of 108,431 publications, screening to 28,721 relevant records, and extraction of 2,987 high-signal ACUs covering 793 normalised factor nodes and 769 normalised endpoint nodes represents the most comprehensive automated, quality-graded claim database of the human longevity intervention literature yet assembled. The final pipeline also includes two safeguards added after the initial extraction workflow: polarity correction, which addresses semantic reversals introduced by endpoint normalisation, and LLM-based claim validation, which removes unsupported normalised claims before evidence tiering.

Several important limitations should be acknowledged. The analysis was restricted to publication abstracts, which are typically less detailed than full texts and may not capture nuances of study methodology, secondary outcomes, or subgroup analyses. The restriction to English-language publications introduces a language bias, potentially excluding relevant studies published in other languages, particularly from Asian research communities where ageing research is highly active.

LLM-based extraction introduces a source of noise that, while reduced by prompt design, raw-string validation, post-extraction filtering, polarity correction, and LLM-based claim validation, cannot be eliminated entirely. GPT-5.2 may mischaracterise study designs, misattribute effect directions, or generate plausible-sounding but incorrect endpoint descriptions. The claim-validation step should therefore be interpreted as an automated consistency check against abstracts, not as independent human validation or full-text adjudication.

Taxonomy normalisation inevitably involves trade-offs between over-splitting (treating genuine synonyms as distinct entities) and over-merging (treating distinct factors as equivalent). Systematic evaluation of normalisation accuracy was beyond the scope of the present work.

The evidence-tiering model assigns weights based on the explicit hierarchy of clinical evidence, but these weights embody value judgements that may not be universally shared.

A further methodological consideration is the temporal dimension of the evidence base. The mean evidence score of ACUs showed no large sustained upward shift over publication year, suggesting that growth in publication volume has not been accompanied by a proportional improvement in average study quality (Fig. 24). This finding is consistent with the observation that the expansion of the longevity literature has been driven in large part by small exploratory trials that, while adding breadth, do not substantially shift the evidence tier distribution. The score distribution by study type underscores the design quality gap: interventional ACUs had a higher median score than observational ACUs (74.5 versus 69.3; Fig. 25). This difference reflects the higher design and comparison-structure subscores assigned to controlled interventional evidence, although both groups showed broad score distributions.

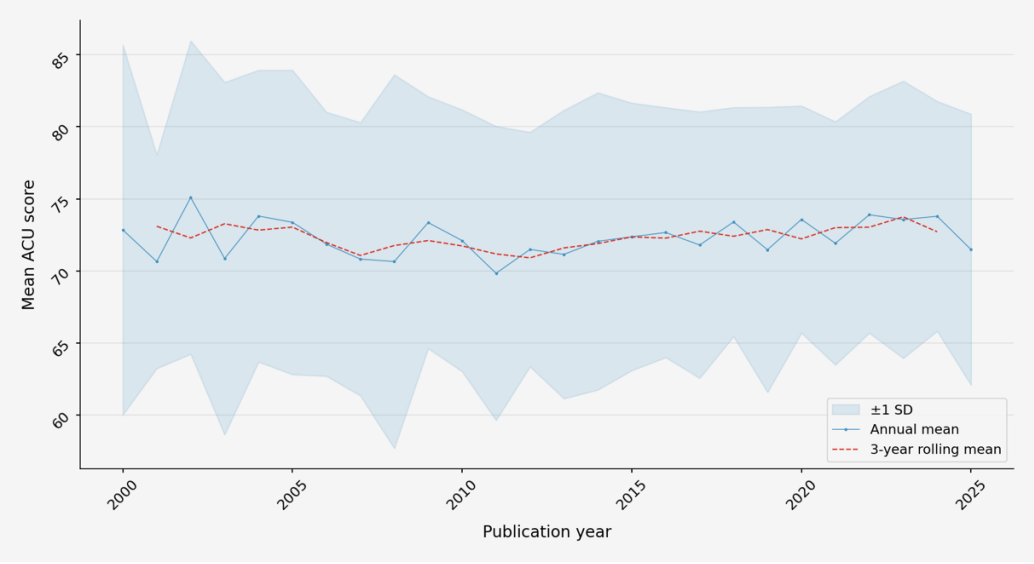


*Fig. 24. Mean evidence score by publication year (2000–2025).*

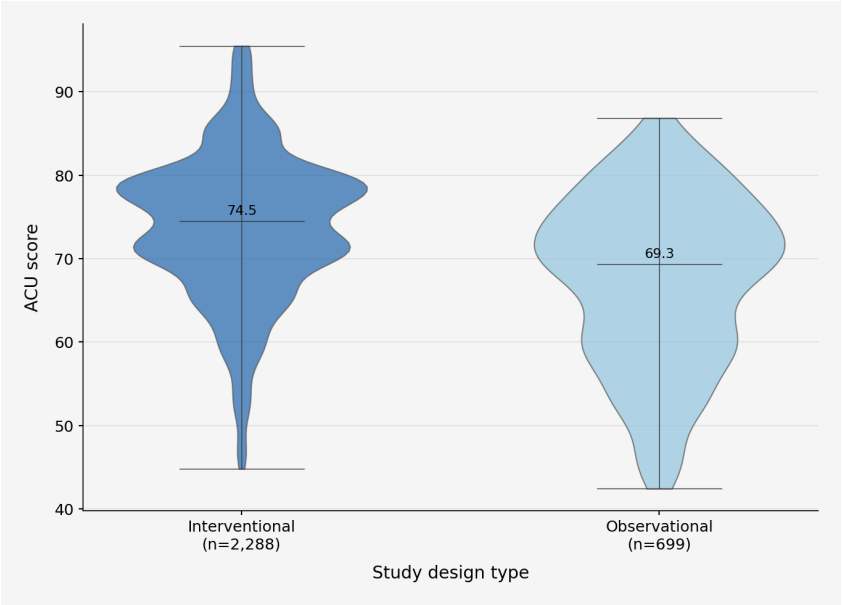


*Fig. 25. Evidence score distribution (violin plot) stratified by ACU type (interventional versus observational).*

### 5.6 Future Directions

Several extensions of the present work would substantially increase its value for the research community. Extension to full-text processing using open-access articles accessible through PubMed Central would enable extraction of more detailed information about study methodology, secondary outcomes, and subgroup analyses.

Multi-language retrieval would address the current English-language bias. Integration with OpenAlex's citation network data would enable citation-based evidence weighting.

Formal meta-analytic pooling of effect sizes across ACUs sharing the same (factor, outcome) pair – where standardised effect sizes can be computed and combined – would transform the evidence database from a qualitative ranking resource into a quantitative synthesis tool.

A further extension would be to convert validated ACUs into predicate-logic or knowledge-graph statements. This would allow rule-based inference across factors, endpoints, populations, evidence tiers, and hallmark mappings. The present thesis stops at extraction, normalisation, polarity correction, abstract-level validation, and descriptive evidence structuring; it does not attempt logical inference over the resulting claim graph.

The public web interface at longevityevidence.org enables querying of the evidence database by factor, outcome, evidence quality tier, and hallmark, substantially increasing the accessibility and utility of the resource. Pre-registration of the systematic update protocol would further increase the rigour and transparency of the living evidence database.

## CONCLUSIONS

1. Claim-level decomposition is an appropriate representation for the longevity evidence problem because it converts compound biomedical abstract statements into comparable factor–endpoint units while preserving study provenance, population context, study design, and effect direction.
2. Large language model-based screening can make high-recall surveillance of more than one hundred thousand biomedical records operationally feasible, but it should be interpreted as a recall-oriented filtering layer whose outputs require structured extraction, validation, and quality grading before evidence claims are used.
3. The resulting corpus indicates that the published human longevity literature is broad but shallow: after filtering and validation, 2,987 quality-graded claims from 1,797 publications covered 793 factor nodes and 769 endpoint nodes, yet most factor–endpoint pairs had little independent replication.
4. Evidence quality in the retained corpus is limited mainly by endpoint indirectness and sparse replication. Only 7.6% of claims reached the highest quality tier, and only 1.1% directly targeted survival or longevity outcomes, showing that most current evidence remains clinically indirect.
5. The strongest direct evidence in the corpus is concentrated in exercise and lifestyle factors, whereas many supplement, biomarker-centred, and mechanistic longevity claims remain less direct. The public resource at longevityevidence.org therefore provides a transparent starting point for prioritising expert review and future synthesis rather than a substitute for systematic review or meta-analysis.

## RECOMMENDATION

Future longevity intervention studies should prioritise clinically direct outcomes, particularly survival, disability, frailty, major disease events, and validated functional measures, instead of continuing to accumulate mainly surrogate endpoint evidence. Where surrogate biological age or molecular measures are used, they should be included alongside primary clinical outcomes so that their long-term interpretability can be tested rather than assumed.

Researchers conducting evidence synthesis should consider claim-level decomposition as a complement to traditional study-level review methods. For this evidence database specifically, the highest-value next step is independent replication and expert review of the strongest direct-survival and clinical-function claims, because the current corpus remains broad but sparsely replicated across most factor-outcome pairs.

## ACKNOWLEDGEMENTS

The author wishes to thank Prof. PhD Audronė Jakaitienė for her supervision, guidance, and constructive feedback throughout the development of this thesis. Gratitude is also extended to the faculty and staff of the Systems Biology Master's Programme at the Faculty of Medicine, Vilnius University, for providing an intellectually rigorous environment in which this work was developed. The author also thanks Doc. Dr. Erinija Pranckevičienė for her careful review of the thesis and constructive comments.

AI-assisted tools were used during the thesis work. ChatGPT, GPT-5.2 and GPT-5-mini by OpenAI were used for LLM-based pipeline extraction and validation tasks, language polishing, formatting support, and generation of some code snippets during development. The author reviewed, verified, and takes responsibility for all thesis content, analyses, code, and conclusions.

## REFERENCES


1. Aging Atlas Consortium. Aging Atlas: a multi-omics database for aging biology. Nucleic Acids Research. 2021;49(D1):D825–D830.
2. Balasubramanian JB, Adams D, Roxanis I, et al. Leveraging large language models for structured information extraction from pathology reports. Journal of Pathology Informatics. 2025;19:100523.
3. Barardo D, Thornton D, Thoppil H, et al. The DrugAge database of aging-related drugs. Aging Cell. 2017;16(3):594–597.
4. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research. 2004;32(Database issue):D267–D270.
5. Budovsky A, Craig T, Wang J, et al. LongevityMap: a database of human genetic variants associated with longevity. Trends in Genetics. 2013;29(10):559–560.
6. Cummings SR, Kritchevsky SB. Endpoints for geroscience clinical trials: health outcomes, biomarkers, and biologic age. GeroScience. 2022;44:2925–2931.
7. Dent E, Morley JE, Cruz-Jentoft AJ, et al. Physical frailty: ICFSR international clinical practice guidelines for identification and management. The Journal of Nutrition, Health & Aging. 2019;23(9):771–787.
8. DeYoung J, Lehman E, Nye B, Marshall IJ, Wallace BC. Evidence Inference 2.0: More Data, Better Models. Proceedings of the BioNLP 2020 Workshop. 2020:123–132.
9. Garda S, Weber-Genzel L, Martin R, Leser U. BELB: a Biomedical Entity Linking Benchmark. arXiv preprint. 2023:arXiv:2308.11537.
10. Hu Q, Long Q, Wang W. Decomposition Dilemmas: Does Claim Decomposition Boost or Burden Fact-Checking Performance? Proceedings of NAACL-HLT 2025. 2025:6313–6336.
11. Hühne R, Thalheim T, Sühnel J. AgeFactDB–the JenAge Ageing Factor Database–towards data integration in ageing research. Nucleic Acids Research. 2014;42(D1):D892–D896.
12. Ioannidis JPA. Why most published research findings are false. PLoS Medicine. 2005;2(8):e124.
13. Justice JN, Ferrucci L, Newman AB, et al. A framework for selection of blood-based biomarkers for geroscience-guided clinical trials. GeroScience. 2018;40(5–6):419–436.
14. Kartchner D, Deng J, Lohiya S, et al. A comprehensive evaluation of biomedical entity linking models. Proceedings of EMNLP 2023. 2023:14462–14478.
15. Kartchner D, Turner H, Ye C, et al. TrialSieve: A Comprehensive Biomedical Information Extraction Framework for PICO, Meta-Analysis, and Drug Repurposing. Bioengineering. 2025;12:486.
16. Kennedy BK, Berger SL, Brunet A, Campisi J, Cuervo AM, Epstein ES, et al. Geroscience: Linking Aging to Chronic Disease. Cell. 2014;159(4):709–713.
17. Kriukov D, Efimov E, Gelfand MS, Moskalev A, Khrameeva EE. Do we actually need aging clocks? npj Aging. 2026;12:15.

18. Li G, Cheng L, Wong IN, et al. Predicting healthspan and disease risks through biological age. Trends in Molecular Medicine. 2025 (Article in Press).

19. Lipscomb CE. Medical Subject Headings (MeSH). Bulletin of the Medical Library Association. 2000;88(3):265–266.

20. Liu F, Shareghi E, Meng Z, Basaldella M, Collier N. Self-alignment pretraining for biomedical entity representations. Proceedings of NAACL-HLT 2021. 2021:4228–4238.

21. López-Otín C, Blasco MA, Partridge L, Serrano M, Kroemer G. Hallmarks of aging: An expanding universe. Cell. 2023;186(2):243–278.

22. Marshall IJ, Nye B, Kuiper J, et al. Trialstreamer: A living, automatically updated database of clinical trial reports. Journal of the American Medical Informatics Association. 2020;27(12):1903–1912.

23. Metropolitansky D, Larson J. Towards Effective Extraction and Evaluation of Factual Claims. Proceedings of ACL 2025. 2025:6996–7045.

24. Min M, Egli C, Dulai AS, Sivamani RK. Critical review of aging clocks and factors that may influence the pace of aging. Frontiers in Aging. 2024;5:1487260.

25. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: Fast and robust models for biomedical natural language processing. Proceedings of the BioNLP 2019 Workshop. 2019:319–327.

26. Newman JC, Milman S, Hashmi SK, et al. Strategies and challenges in clinical trials targeting human aging. The Journals of Gerontology Series A. 2016;71(11):1424–1434.

27. Ntinopoulos V, Rodriguez Cetina Biefer H, Tudorache I, et al. Large language models for data extraction from unstructured and semi-structured electronic health records: a multiple model evaluation. npj Digital Medicine. 2025;8:182.

28. Nye B, Li JJ, Patel R, et al. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. Proceedings of ACL 2018. 2018:197–207.

29. OpenAI. ChatGPT, GPT-5.2 and GPT-5-mini. OpenAI. 2026. Available at: https://chatgpt.com/ (accessed 10 May 2026).

30. Parish A, Ioannidis JPA, Zhang K, Barardo D, Swindell WR, de Magalhães JP. Reporting quality, effect sizes, and biases for aging interventions: a methodological appraisal of the DrugAge database. npj Aging. 2025;11:96.

31. Peng H, Xiong Y, Xiang Y, Wang H, Lu Y, Liu J. Biomedical named entity normalization via interaction-based synonym marginalization. Journal of Biomedical Informatics. 2022;136:104238.

32. Pradeep R, Ma X, Nogueira R, Lin J. Scientific Claim Verification with VERT5ERINI. arXiv preprint. 2020:arXiv:2010.11930.

33. Reason T, Langham J, Gimblett A. Automated Mass Extraction of Over 680,000 PICOs from Clinical Study Abstracts Using Generative AI: A Proof-of-Concept Study. Pharmaceutical Medicine. 2024;38:283–296.

34. Rolland Y, Sierra F, Ferrucci L, et al. Challenges in developing Geroscience trials. Nature Communications. 2023;14:5038.

35. Sanders J. Biomarker development for translational geroscience: Considerations for a strategic framework focusing on early clinical development. Aging Cell. 2023;22:e13837.

36. Sarrouti M, Ben Abacha A, Mrabet Y, Demner-Fushman D. Evidence-based Fact-Checking of Health-related Claims. Findings of EMNLP 2021. 2021:3499–3512.

37. Scheibye-Knudsen M. A Grand Challenge in Aging Interventions: From Mice to Humans. Frontiers in Aging. 2020;1:566651.

38. Sung M, Jeon H, Lee J, Kang J. Biomedical entity representations with synonym marginalization. Proceedings of ACL 2020. 2020:3641–3650.

39. Sung M, Jeong M, Choi Y, Kim D, Lee J, Kang J. BERN2: an advanced neural biomedical named entity recognition and normalization tool. Bioinformatics. 2022;38(20):4837–4839.

40. Tacutu R, Thornton D, Johnson E, et al. Human Ageing Genomic Resources: new and updated databases. Nucleic Acids Research. 2018;46(D1):D1083–D1090.

41. Thorne J, Vlachos A, Christodoulopoulos C, Mittal A. FEVER: a large-scale dataset for Fact Extraction and VERification. Proceedings of NAACL-HLT 2018. 2018:809–819.

42. United Nations, Department of Economic and Social Affairs, Population Division. World Population Ageing 2023: Challenges and opportunities of population ageing in the least developed countries. New York: United Nations; 2024.

43. Vladika J, Schneider P, Matthes F. HealthFC: Verifying Health Claims with Evidence-Based Medical Fact-Checking. Proceedings of LREC-COLING 2024. 2024:8095–8107.

44. Wadden D, Lin S, Lo K, et al. Fact or Fiction: Verifying Scientific Claims. Proceedings of EMNLP 2020. 2020:7534–7550.

45. Wanner M, Ebner S, Jiang Z, Dredze M, Van Durme B. A Closer Look at Claim Decomposition. arXiv preprint. 2024:arXiv:2403.11903.

46. Wu Z, Feng C, Hu Y, et al. HALD, a human aging and longevity knowledge graph for precision gerontology and geroscience analyses. Scientific Data. 2023;10:851.

47. Yang A, Yang B, Zhang B, et al. Qwen2.5 Technical Report. arXiv preprint. 2024:arXiv:2412.15115.

48. Zheng L, Li C, Liu Z, Huang F, Jia H, Ye Z, Zhang X. Fact in fragments: Deconstructing complex claims via LLM-based atomic fact extraction and verification. Expert Systems with Applications. 2026;302:130572.

## SUMMARY

Vilnius University, Faculty of Medicine, Systems Biology study programme.

Arnas Stučinskas.

A Large-Scale LLM-Driven System for Extracting Claims About Longevity Factors and Assessing Evidence Quality from Scientific Literature.

This thesis addresses the difficulty of comparing scientific claims about longevity factors across a rapidly growing and methodologically heterogeneous biomedical literature. The aim was to develop and evaluate a reproducible computational system for retrieving longevity-related publications, extracting individual factor-outcome claims, checking claim support against source abstracts, and assigning evidence quality grades. A broad PubMed retrieval identified 108,431 records, of which 28,721 were retained after relevance screening. Structured extraction, claim eligibility filtering, entity normalisation, polarity correction, claim validation, hallmark mapping, and evidence tiering produced 2,987 quality-graded claim units from 1,797 publications. The final corpus covered 793 normalised factor nodes and 769 normalised endpoint nodes. The results showed that human longevity research is broad but shallow: only 7.6% of claims reached the highest evidence tier, only 1.1% directly targeted survival or longevity outcomes, and 91.8% of merged factor-outcome claim pairs were supported by a single study. The thesis concludes that claim-level decomposition combined with evidence tiering can make longevity evidence more transparent, comparable, and reusable, while also exposing major gaps in direct and replicated human evidence.



Keywords: longevity; large language models; evidence quality; natural language processing; biomedical literature; claim extraction.

## SUMMARY IN LITHUANIAN

Vilniaus universitetas, Medicinos fakultetas, Sistemų biologijos studijų programa.

Arnas Stučinskas.

Dideliais kalbos modeliais paremta sistema teiginiams apie ilgaamžiškumo veiksnius ir jų įrodymų kokybei vertinti mokslinėje literatūroje.

Šiame darbe nagrinėjama problema, kad mokslinius teiginius apie ilgaamžiškumo veiksnius sunku palyginti dėl sparčiai augančios ir metodologiškai nevienalytės biomedicininės literatūros. Darbo tikslas buvo sukurti ir įvertinti atkuriamą sistemą, kuri surenka su ilgaamžiškumu susijusias publikacijas, išskiria atskirus veiksnio ir baigties teiginius, patikrina jų atitikimą šaltinio santraukai ir priskiria įrodymų kokybės įvertinimus. Plati PubMed paieška rado 108 431 įrašą, iš kurių po tinkamumo atrankos paliktas 28 721 įrašas. Struktūruotas išgavimas, teiginių tinkamumo filtravimas, esybių normalizavimas, poliariškumo korekcija, teiginių validavimas, senėjimo požymių susiejimas ir įrodymų pakopų nustatymas sudarė 2 987 kokybiškai įvertintus teiginių vienetus iš 1 797 publikacijų. Galutinis rinkinys apėmė 793 normalizuotus veiksnių mazgus ir 769 normalizuotus baigčių mazgus. Rezultatai parodė, kad žmogaus ilgaamžiškumo tyrimų bazė yra plati, bet menkai replikuota: tik 7,6 procento teiginių pateko į aukščiausią įrodymų pakopą, tik 1,1 procento tiesiogiai nagrinėjo išgyvenamumą arba ilgaamžiškumą, o 91,8 procento sujungtų veiksnio ir baigties teiginių porų rėmėsi viena studija. Daroma išvada, kad teiginių skaidymas ir įrodymų kokybės pakopos gali padaryti ilgaamžiškumo įrodymus skaidresnius, lengviau palyginamus ir pakartotinai naudojamus, kartu išryškindami tiesioginių ir replikuotų žmogaus tyrimų spragas.

Raktažodžiai: ilgaamžiškumas; didieji kalbos modeliai; įrodymų kokybė; natūralios kalbos apdorojimas; biomedicininė literatūra; teiginių išgavimas.

# APPENDICES

## Appendix A. Exact PubMed high-recall query

The following query is copied from `retrievers/pubmed.py` and was used as the base PubMed query for retrieval.

```
(
  aging[Title/Abstract]
  OR ageing[Title/Abstract]
  OR "age-related"[Title/Abstract]
  OR longevity[Title/Abstract]
  OR healthspan[Title/Abstract]
  OR "health span"[Title/Abstract]
  OR lifespan[Title/Abstract]
  OR "life span"[Title/Abstract]
  OR geroscience[Title/Abstract]
  OR "biological age"[Title/Abstract]
  OR "age acceleration"[Title/Abstract]
  OR "epigenetic age"[Title/Abstract]
  OR "epigenetic clock"[Title/Abstract]
  OR "DNA methylation age"[Title/Abstract]
  OR frailty[Title/Abstract]
  OR frail[Title/Abstract]
  OR sarcopenia[Title/Abstract]
  OR "functional decline"[Title/Abstract]
  OR "loss of independence"[Title/Abstract]
  OR "activities of daily living"[Title/Abstract]
  OR ADL[Title/Abstract]
  OR IADL[Title/Abstract]
  OR falls[Title/Abstract]
  OR "fall risk"[Title/Abstract]
  OR "mobility decline"[Title/Abstract]
  OR "grip strength"[Title/Abstract]
  OR "cognitive decline"[Title/Abstract]
  OR dementia[Title/Abstract]
  OR multimorbidity[Title/Abstract]
  OR cachexia[Title/Abstract]
  OR senescence[Title/Abstract]
  OR senolytic[Title/Abstract]
  OR inflammaging[Title/Abstract]
  OR immunosenescence[Title/Abstract]
  OR "telomere length"[Title/Abstract]
  OR centenarian[Title/Abstract]
  OR centenarians[Title/Abstract]
  OR "intrinsic capacity"[Title/Abstract]
  OR Aging[MeSH Terms]
  OR "Life Expectancy"[MeSH Terms]
  OR Frailty[MeSH Terms]
  OR Sarcopenia[MeSH Terms]
)
AND English[Language]
```

```
AND
(
  "Clinical Trial"[Publication Type]
  OR "Controlled Clinical Trial"[Publication Type]
  OR "Randomized Controlled Trial"[Publication Type]
  OR "Pragmatic Clinical Trial"[Publication Type]
  OR "Equivalence Trial"[Publication Type]
  OR "Adaptive Clinical Trial"[Publication Type]
  OR "Clinical Trial, Phase I"[Publication Type]
  OR "Clinical Trial, Phase II"[Publication Type]
  OR "Clinical Trial, Phase III"[Publication Type]
  OR "Clinical Trial, Phase IV"[Publication Type]
  OR "Observational Study"[Publication Type]
  OR "Comparative Study"[Publication Type]
  OR "Multicenter Study"[Publication Type]
  OR "Twin Study"[Publication Type]
)
NOT
(
  "Clinical Trial Protocol"[Publication Type]
  OR Review[ptyp]
  OR "Systematic Review"[ptyp]
  OR "Scoping Review"[ptyp]
  OR "Meta-Analysis"[ptyp]
  OR "Network Meta-Analysis"[ptyp]
  OR Comment[ptyp]
  OR Letter[ptyp]
  OR Editorial[ptyp]
  OR "Practice Guideline"[ptyp]
  OR Guideline[ptyp]
  OR "Consensus Development Conference"[ptyp]
  OR "Consensus Development Conference, NIH"[ptyp]
  OR Case Reports[ptyp]
  OR "Retracted Publication"[ptyp]
  OR "Expression of Concern"[ptyp]
)
```

## Appendix B. Structured extraction and ACU schemas

Appendix B provides JSON-format schema summaries corresponding to the SEF and ACU field figures shown in Figs. 2 and 3. The SEF schemas are implemented in `schemas/interventional.py` and `schemas/observational.py`; they capture the full per-abstract extraction before atomisation. The downstream ACU schema is constructed by `s02_04_build_acus_interventional.py` and `s02_05_build_acus_observational.py` and retains one factor, one endpoint, effect direction or stance, effect size, significance, study design, population descriptors, sample size, species, raw-string anchors, and the deterministic ACU identifier.

The following JSON listings provide the machine-readable schema summaries. They are abbreviated only by representing field types and enum sets as strings; the object hierarchy and required fields follow the repository schemas and ACU construction code.

Interventional SEF schema summary in JSON format

This schema is the LLM extraction target for interventional studies. It preserves arm structure and comparison-level endpoint findings before ACU construction.

```
{
  "interventional_sef": {
    "source": "schemas/interventional.py::SEF_SCHEMA",
    "required": [
      "study_design",
      "species_group",
      "species_name",
      "health_status",
      "population_summary",
      "population_gender",
      "population_age_group",
      "total_n",
      "trial_registration_id",
      "regimen_text",
      "shared_base",
      "arms",
      "comparisons"
    ],
    "fields": {
      "study_design": "rct | nonrandomized_controlled | single_arm | crossover | factorial | other |
unknown",
      "species_group": "human | animal | in_vitro | mixed | other | unknown",
      "species_name": "string | null",
      "health_status": "healthy | diseased | mixed | unknown",
      "population_summary": "string | null",
      "population_gender": "male | female | mixed | unknown",
      "population_age_group": "older_adults | younger_or_midlife | mixed | unknown",
      "total_n": "integer | null",
      "trial_registration_id": "string | null",
      "regimen_text": "string | null",
      "shared_base": {
        "summary": "string | null"
      },
      "arms": [
        {
          "arm_id": "string",
          "arm_role": "treatment | comparator | other | unknown",
          "arm_name": "string | null",
          "arm_summary": "string | null",
          "arm_type": "active | placebo | sham | usual_care | no_treatment | other | unknown",
          "intervention_components": [
            {
```

```
              "categories": [
                {
                  "label": "drug | supplement | diet | exercise | behavioral | procedure | device |
other | unknown",
                  "rank": "1 | 2"
                }
              ],
              "modification": "string | null",
              "name": {
                "raw_string": "string",
                "canonical": "string | null"
              }
            }
          ]
        }
      ],
      "comparisons": [
        {
          "comparison_scope": "between_arms | within_subject | unknown",
          "treatment_arm_id": "string",
          "comparator_arm_id": "string",
          "endpoints": [
            {
              "categories": [
                {
                  "label": "mortality | lifespan | clinical_event | frailty | patient_reported |
functional | physiology | biomarker | imaging | safety | other | unknown",
                  "rank": "1 | 2"
                }
              ],
              "timepoint": "string | null",
              "endpoint_variant": "string | null",
              "name": {
                "raw_string": "string",
                "canonical": "string | null"
              },
              "effect_direction": "increase | decrease | no_change | unknown",
              "effect_polarity": "more_is_better | less_is_better | unknown",
              "significance": "significant | not_significant | unknown",
              "effect_size": "none | small | medium | large | unknown",
              "stance": "benefit | harm | no_difference | unknown",
              "longevity_relation": "promotes | demotes | direct | not_related | unknown"
            }
          ]
        }
      ]
    }
  }
}
```

## Field descriptions:

`source` – Repository object from which the summary schema was derived.

`required` – Top-level fields that must be present in every interventional SEF.

`fields` – Top-level object describing the accepted field structure and value types.

`study_design` – Interventional design class used later for evidence-quality scoring.

`species_group` – Broad species category of the study population.

`species_name` – More specific species label when available.

`health_status` – Baseline health category of the study population.

`population_summary` – Free-text summary of the participants described in the abstract.

`population_gender` – Participant sex or gender composition when reported.

`population_age_group` – Age category used for population relevance scoring.

`total_n` – Total sample size extracted from the abstract.

`trial_registration_id` – Clinical trial registration identifier when present.

`regimen_text` – Free-text description of the intervention regimen or dosing schedule.

`shared_base.summary` – Information common to study arms, such as background treatment.

arms[] – List of intervention and comparator arms described by the abstract.

arms[].arm_id – Internal identifier used to link comparisons to arms.

arms[].arm_role – Whether the arm functions as treatment, comparator, or other.

arms[].arm_name – Short arm label reported or inferred from the abstract.

arms[].arm_summary – Free-text description of the arm.

arms[].arm_type – Whether the arm is active, placebo, sham, usual care, no treatment, or other.

arms[].intervention_components[] – Specific intervention components contained in the treatment arm.

intervention_components[].categories[] – One or two broad intervention categories.

intervention_components[].categories[].label – Intervention category label used for later grouping.

intervention_components[].categories[].rank – Priority rank for primary versus secondary category assignment.

intervention_components[].modification – Dose, intensity, duration, or other modifier when available.

intervention_components[].name.raw_string – Verbatim intervention text span from the abstract.

intervention_components[].name.canonical – Normalised intervention label proposed by the model.

comparisons[] – List of analysed treatment-comparator relationships.

comparisons[].comparison_scope – Whether the comparison is between arms, within subject, or unknown.

comparisons[].treatment_arm_id – Arm identifier for the active treatment side of the comparison.

comparisons[].comparator_arm_id – Arm identifier for the comparator side of the comparison.

comparisons[].endpoints[] – Endpoint findings attached to the comparison.

endpoints[].categories[] – One or two endpoint categories used for endpoint grouping.

endpoints[].categories[].label – Endpoint category label such as mortality, functional, biomarker, or imaging.

endpoints[].categories[].rank – Priority rank for primary versus secondary endpoint category assignment.

endpoints[].timepoint – Follow-up time or measurement timepoint when reported.

endpoints[].endpoint_variant – Variant or subtype of the endpoint when reported.

endpoints[].name.raw_string – Verbatim endpoint text span from the abstract.

endpoints[].name.canonical – Normalised endpoint label proposed by the model.

endpoints[].effect_direction – Observe direction of change in the endpoint.

endpoints[].effect_polarity – Whether a higher or lower endpoint value is beneficial.

endpoints[].significance – Statistical significance category reported or inferred from the abstract.

endpoints[].effect_size – Categorical effect magnitude extracted from the abstract.

endpoints[].stance – Interpreted claim direction: benefit, harm, no difference, or unknown.

endpoints[].longevity_relation – Whether the endpoint relation is direct, promoting, demoting, not related, or unknown.

## Observational SEF schema summary in JSON format

This schema is the LLM extraction target for observational studies. Each association points to a specific exposure by index, which prevents an exposure-endpoint cross-product.

```
{
  "observational_sef": {
    "source": "schemas/observational.py::SEF_SCHEMA",
    "required": [
      "study_design",
      "species_group",
      "species_name",
      "health_status",
      "population_summary",
      "population_gender",
      "population_age_group",
      "total_n",
      "primary_exposures",
      "covariates",
      "associations"
    ],
    "fields": {
      "study_design": "cohort | case_control | cross_sectional | mendelian_randomization | unknown",
      "species_group": "human | animal | in_vitro | mixed | other | unknown",
      "species_name": "string | null",
      "health_status": "healthy | diseased | mixed | unknown",
      "population_summary": "string | null",
      "population_gender": "male | female | mixed | unknown",
      "population_age_group": "older_adults | younger_or_midlife | mixed | unknown",
      "total_n": "integer | null",
      "primary_exposures": [
        {
          "categories": [
            {
              "label": "behavioral | diet | exercise | biomarker | medication_use | genetic_variant
| environmental | socioeconomic | other | unknown",
              "rank": "1 | 2"
            }
          ],
          "modification": "string | null",
          "name": {
            "raw_string": "string",
            "canonical": "string | null"
          }
        }
      ],
      "covariates": [
        {
          "name": {
            "raw_string": "string",
            "canonical": "string | null"
          }
        }
      ],
      "associations": [
        {
          "association_exposure_index": "integer; zero-based index into primary_exposures",
          "comparison_scope": "between_groups | dose_response | unknown",
          "exposure_contrast": "string | null",
          "exposure_level_active": "string | null",
          "exposure_level_reference": "string | null",
          "model_adjustment": "adjusted | unadjusted | unknown",
          "analysis_scale": "individual_level | group_level | unknown",
          "endpoints": [
            {
              "categories": [
                {
                  "label": "mortality | lifespan | clinical_event | frailty | patient_reported |
functional | physiology | biomarker | imaging | safety | other | unknown",
                  "rank": "1 | 2"
                }
              ],
              "timepoint": "string | null",
              "endpoint_variant": "string | null",
              "name": {
                "raw_string": "string",
                "canonical": "string | null"
              },
              "effect_direction": "increase | decrease | no_change | unknown",
              "effect_polarity": "more_is_better | less_is_better | unknown",
              "significance": "significant | not_significant | unknown",
              "effect_size": "none | small | medium | large | unknown",
              "stance": "benefit | harm | no_difference | unknown",
              "longevity_relation": "promotes | demotes | direct | not_related | unknown"
            }
          ]
        }
      ]
    }
  }
}
```

## Field descriptions:

`source` – Repository object from which the summary schema was derived.

`Required` – Top-level fields that must be present in every observational SEF.

`fields` – Top-level object describing the accepted field structure and value types.

`study_design` – Observational design class used later for evidence-quality scoring.

`species_group` – Broad species category of the study population.

`species_name` – More specific species label when available.

`health_status` – Baseline health category of the study population.

`population_summary` – Free-text summary of the participants described in the abstract.

`population_gender` – Participant sex or gender composition when reported.

`population_age_group` – Age category used for population relevance scoring.

`total_n` – Total sample size extracted from the abstract.

`primary_exposures[]` – Exposure variables that may be linked to endpoint associations.

`primary_exposures[].categories[]` – One or two broad exposure categories.

`primary_exposures[].categories[].label` – Exposure category label used for later grouping.

`primary_exposures[].categories[].rank` – Priority rank for primary versus secondary category assignment.

`primary_exposures[].modification` – Dose, level, frequency, or other exposure modifier when available.

`primary_exposures[].name.raw_string` – Verbatim exposure text span from the abstract.

`primary_exposures[].name.canonical` – Normalised exposure label proposed by the model.

`covariates[]` – Adjustment covariates explicitly represented in the abstract.

`covariates[].name.raw_string` – Verbatim covariate text span from the abstract.

`covariates[].name.canonical` – Normalised covariate label proposed by the model.

`associations[]` – Exposure-outcome associations reported by the abstract.

`associations[].association_exposure_index` – Zero-based pointer to the exposure in `primary_exposures[].`

`associations[].comparison_scope` – Whether the association compares groups, dose response, or unknown.

`associations[].exposure_contrast` – Textual contrast used to compare exposure levels.

`associations[].exposure_level_active` – Higher, active, or exposed level in the association.

`associations[].exposure_level_reference` – Reference or lower exposure level in the association.

`associations[].model_adjustment` – Whether the statistical model was adjusted, unadjusted, or unclear.

`associations[].analysis_scale` – Whether analysis is individual-level, group-level, or unknown.

`associations[].endpoints[]` – Endpoint findings attached to the exposure association.

`endpoints[].categories[]` – One or two endpoint categories used for endpoint grouping.

`endpoints[].categories[].label` – Endpoint category label such as mortality, functional, biomarker, or imaging.

`endpoints[].categories[].rank` – Priority rank for primary versus secondary endpoint category assignment.

`endpoints[].timepoint` – Follow-up time or measurement timepoint when reported.

`endpoints[].endpoint_variant` – Variant or subtype of the endpoint when reported.

`endpoints[].name.raw_string` – Verbatim endpoint text span from the abstract.

`endpoints[].name.canonical` – Normalised endpoint label proposed by the model.

`endpoints[].effect_direction` – Observed direction of association with the endpoint.

`endpoints[].effect_polarity` – Whether a higher or lower endpoint value is beneficial.

`endpoints[].significance` – Statistical significance category reported or inferred from the abstract.

`endpoints[].effect_size` – Categorical effect magnitude extracted from the abstract.

`endpoints[].stance` – Interpreted claim direction: benefit, harm, no difference, or unknown.

`endpoints[].longevity_relation` – Whether the endpoint relation is direct, promoting, demoting, not related, or unknown.

## ACU schema summary in JSON format

The ACU object is not submitted directly to the LLM; it is produced deterministically from SEFs by the ACU builder scripts and used for filtering, validation, tiering, and merging.

```
{
  "acu": {
    "source": [
      "s02_04_build_acus_interventional.py",
      "s02_05_build_acus_observational.py"
    ],
    "description": "One analysis-ready factor-endpoint-direction claim produced from one SEF
comparison or association.",
    "fields": {
      "acu_id": "string; deterministic identifier",
      "record_id": "string",
      "pipeline_version": "acu_interventional_v2 | acu_observational_v2",
      "health_status": "healthy | diseased | mixed | unknown",
      "population_summary": "string | null",
      "population_gender": "male | female | mixed | unknown",
      "population_age_group": "older_adults | younger_or_midlife | mixed | unknown",
      "total_n": "integer | null",
      "species": {
        "group": "human | animal | in_vitro | mixed | other | unknown",
        "name": "string | null"
      },
      "effector": {
        "categories": [
          "string"
        ],
        "canonical": "string",
        "modification": "string | null",
        "raw_string": "string | null"
      },
      "endpoint": {
        "categories": [
          "string"
        ],
        "canonical": "string",
        "timepoint": "string | null",
        "variant": "string | null",
        "raw_string": "string | null"
      },
      "effect": {
        "direction": "increase | decrease | no_change | unknown",
        "polarity": "more_is_better | less_is_better | unknown",
        "significance": "significant | not_significant | unknown",
        "effect_size": "none | small | medium | large | unknown",
        "stance": "benefit | harm | no_difference | unknown",
        "longevity_relation": "promotes | demotes | direct | not_related | unknown"
      },
      "details": {
        "study_design": "string",
        "comparison_scope": "string",
        "interventional_only": {
          "comparator_arm_id": "string",
          "arm_id": "string",
          "arm_role": "string",
          "arm_type": "string"
        },
        "observational_only": {
          "model_adjustment": "adjusted | unadjusted | unknown",
          "analysis_scale": "individual_level | group_level | unknown",
          "exposure_contrast": "string | null",
          "exposure_level_active": "string | null",
          "exposure_level_reference": "string | null",
          "association_exposure_index": "integer"
        }
      }
```

```
    }
  }
}
```

## Field descriptions:

`source` – Builder scripts from which the ACU object structure was derived.

`description` – One-sentence definition of the claim-level ACU object.

`fields` – Top-level object describing fields stored in each ACU.

`acu_id` – Deterministic unique identifier for the claim unit.

`record_id` – Identifier of the source publication record.

`pipeline_version` – ACU construction version and study stream.

`health_status` – Baseline health category inherited from the parent SEF.

`population_summary` – Free-text population summary inherited from the parent SEF.

`population_gender` – Participant sex or gender composition inherited from the parent SEF.

`population_age_group` – Age category inherited from the parent SEF.

`total_n` – Sample size inherited from the parent SEF.

`species.group` – Broad species category inherited from the parent SEF.

`species.name` – Specific species label when available.

`effector.categories[]` – Normalised factor categories for the intervention or exposure.

`effector.canonical` – Normalised factor node used for matching and merging claims.

`effector.modification` – Dose, level, frequency, or other modifier when available.

`effector.raw_string` – Verbatim factor text span from the source abstract.

`endpoint.categories[]` – Normalised endpoint categories for grouping outcomes.

`endpoint.canonical` – Normalised endpoint node used for matching and merging claims.

`endpoint.timepoint` – Follow-up or measurement timepoint when available.

`endpoint.variant` – Endpoint subtype or variant when available.

`endpoint.raw_string` – Verbatim endpoint text span from the source abstract.

`effect.direction` – Observed increase, decrease, no change, or unknown direction.

`effect.polarity` – Clinical interpretation of whether more or less is better.

`effect.significance` – Statistical significance category.

`effect.effect_size` – Categorical effect magnitude.

`effect.stance` – Final interpreted claim stance used for filtering and merging.

`effect.longevity_relation` – Relationship of the endpoint to longevity relevance.

`details.study_design` – Study design inherited from the SEF and used in scoring.

`details.comparison_scope` – Type of comparison or association from which the ACU was built.

`details.interventional_only.comparator_arm_id` – Comparator arm identifier for interventional ACUs.

`details.interventional_only.arm_id` – Treatment arm identifier for interventional ACUs.

`details.interventional_only.arm_role` – Role of the treatment arm in the parent SEF.

`details.interventional_only.arm_type` – Type of treatment arm in the parent SEF.

`details.observational_only.model_adjustment` – Adjustment status of the observational model.

`details.observational_only.analysis_scale` – Scale at which the observational association was analysed.

`details.observational_only.exposure_contrast` – Textual exposure contrast used in the association.

`details.observational_only.exposure_level_active` – Active or higher exposure level.

`details.observational_only.exposure_level_reference` – Reference or lower exposure level.

`details.observational_only.association_exposure_index` – Pointer back to the source exposure in the observational SEF.

## Appendix C. Evidence scoring thresholds and sensitivity scenarios

Individual ACU evidence tiers use score thresholds A >=85, B 71-84, C 55-70, and D <55. The base weight vector is design 0.25, comparison 0.10, adjustment 0.08, endpoint proximity 0.30, effect credibility 0.18, sample size 0.05, and population relevance 0.04. Sensitivity analyses compared this base model with two alternative weighting scenarios: a design-heavy scenario, which increased the weight of study design and reduced endpoint proximity, and an effect-heavy scenario, which increased the weight of effect credibility and reduced study design and endpoint proximity. In the consolidated corpus, the combined Tier A/B proportion varied from 50.6% in the effect-heavy scenario to 72.0% in the design-heavy scenario, compared with 60.6% under the base model. Thus, absolute tier labels should be interpreted as model-dependent, but the main qualitative conclusion was stable: Tier A evidence remained a small minority under all scenarios, and most retained claims were graded as moderate rather than high-certainty evidence. The exact weight vectors used in the base and sensitivity scenarios are shown in Table C1.

*Table C1. Evidence-scoring weight scenarios used in sensitivity analysis.*

| Scenario | Design | Comparison | Adjustment | Endpoint | Effect | N | Population |
|---|---|---|---|---|---|---|---|
| Base model | 0.25 | 0.10 | 0.08 | 0.30 | 0.18 | 0.05 | 0.04 |
| Design heavy | 0.35 | 0.10 | 0.08 | 0.20 | 0.18 | 0.05 | 0.04 |
| Effect heavy | 0.20 | 0.10 | 0.08 | 0.25 | 0.28 | 0.05 | 0.04 |