Mapping Claims to Evidence in Ageing Research: An Automated NLP Pipeline for Claim-Level Literature Synthesis

Arnas Stučinskas[1], Audronė Jakaitienė[2]

[1] Vilnius University, Faculty of Medicine, Lithuania
[2] Vilnius University, Faculty of Mathematics and Informatics, Institute of Data Science and Digital Technologies, Lithuania
arnas.stucinskas@mf.stud.vu.lt

Abstract. Longevity research spans tens of thousands of clinical and observational publications, yet no systematic, claim-level, quality-graded synthesis of the human literature exists. We present an end-to-end natural language processing pipeline that retrieves, screens, structures, normalises, validates, and quality-grades evidence claims from PubMed at scale. The pipeline uses a local large language model (LLM) for relevance screening and record splitting, and frontier LLMs for structured extraction, entity filtering, taxonomy normalisation, polarity correction, claim validation, and hallmark mapping. Applied to 108,431 retrieved records, the pipeline produced a final dataset of 2,987 quality-graded claims from 1,797 publications, merged into 2,641 factor–outcome claim pairs. The results reveal a broad but shallow evidence landscape: exercise and physical training account for 33.8% of the final corpus, only 1.1% of claims target direct survival or longevity outcomes, and 91.9% of claim pairs are supported by a single study. The main contribution is a modular, updatable NLP system for large-scale claim-level evidence synthesis, together with a public database available at longevityevidence.org.

Keywords: natural language processing, evidence extraction, large language models, biomedical text mining, longevity research, evidence synthesis.

1 Introduction

The longevity literature is large, heterogeneous, and difficult to synthesise manually. PubMed indexes tens of thousands of clinical and observational publications on diet, exercise, supplements, medications, and other factors affecting ageing-related outcomes. Traditional reviews are valuable but slow, narrow in scope, and difficult to keep up to date. Recent advances in large language models (LLMs) create an opportunity to automate evidence extraction and structuring at scale [1, 2].

DOI: https://doi.org/10.15388/LMITT.2026.28

Existing longevity resources such as DrugAge [3], HAGR [4], and LongevityMap [5] are useful, but they are manually curated, limited in coverage, and do not decompose findings into minimal comparable units. This work addresses that gap with a claim-level NLP pipeline for the human longevity literature. The main novelty is not only automated retrieval and extraction, but also the integration of extraction, normalisation, hallmark mapping, validation, and evidence tiering into a single reproducible pipeline.

In this work, the basic evidence record is an Atomic Claim Unit (ACU): a structured representation of exactly one factor–outcome association from one study. For example, if a study reports that resistance training improved grip strength and gait speed, these are represented as two separate ACUs rather than one combined statement: "resistance training improved grip strength" and "resistance training improved gait speed."

2 System Design

The proposed pipeline consists of nine stages that progressively transform raw PubMed records into a structured, quality-graded evidence corpus: literature retrieval, relevance screening and stream split, type-matched structured extraction, ACU construction and filtering, entity filtering and taxonomy normalisation, polarity correction, claim validation, hallmark mapping, and evidence tiering (Fig. 1).

(1) Literature retrieval. A high-recall PubMed query retrieved 108,431 English-language records on human ageing and longevity. Included publication types covered clinical trials and observational studies; reviews, editorials, letters, case reports, and protocols were excluded.

(2) Relevance screening and stream split. A local LLM (Qwen 2.5-7B via Ollama) screened each abstract using the core question: *Does this report a human study relevant to ageing or longevity?* Retained abstracts were then assigned to interventional and/or observational streams: interventional studies apply a treatment or program and measure its effects, whereas observational studies analyse naturally occurring exposures or behaviours without assigning an intervention. Some abstracts were assigned to both streams when the model identified both study types in the same record. This step retained 28,721 abstracts (26.5%), while 1,797 unique publications ultimately contributed at least one ACU to the final corpus (Table 1).

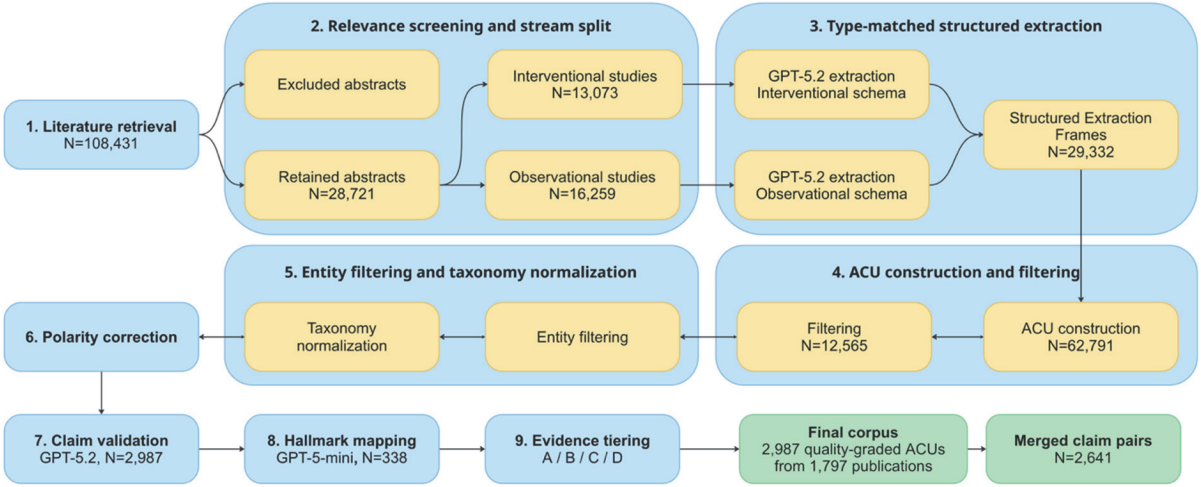

Fig. 1. End-to-end pipeline for claim-level literature synthesis in ageing research. Numbered groups correspond to the nine stages described in Section 2.

Table 1. Publication screening yield.

Stage	N	% of previous row
PubMed retrieval	108,431	–
After LLM screening (qwen2.5-7b-gate)	28,721	26.5%
Unique publications contributing ≥1 ACU	1,797	6.3%

(3) Type-matched structured extraction. Separate JSON schemas were defined for the two streams: one schema for interventional abstracts and one for observational abstracts. Retained abstracts were processed with GPT-5.2 via the OpenAI Batch API to produce one Structured Extraction Frame (SEF) per abstract: a structured JSON representation of the study design and key findings, including factors, outcomes, and effect information. Matching the abstract type to the schema reduced schema mismatches and helped stabilise extraction behaviour, because the model was only asked to fill structures appropriate to the study design identified during screening.

(4) ACU construction and filtering. SEFs were decomposed into 62,791 raw ACUs, each describing exactly one factor affecting one outcome in one study. Multi-criteria filtering retained only human-relevant, directional, statistically supported, and longevity-related claims, yielding 12,565 ACUs. After subsequent entity and hallmark filtering, the final corpus contained 2,987 ACUs, which were merged into 2,641 canonical factor–endpoint claim pairs (Table 2).

Table 2. ACU extraction and filtering yield.

Stage	N	% of previous row
Raw ACUs constructed (gpt-5.2 extraction)	62,791	–
After filtering	12,565	20.0%
After entity and hallmark filtering (final corpus)	2,987	23.8%
Merged (factor × endpoint pairs)	2,641	–

(5) Entity filtering and taxonomy normalisation. Unique factor and endpoint strings were reviewed by GPT-5.2 to remove underspecified, control-label, or irrelevant terms, then mapped to a fixed group–node taxonomy. Here, a group is a broad semantic category (e.g., Exercise & physical training), while a node is a specific normalised concept within that group (e.g., Resistance training). Exact-match lookup tables were used first; unresolved terms were assigned by a two-stage GPT-5-mini classifier. This step produced 890 unique factor labels and 879 endpoint labels.

(6) Polarity correction. After normalisation, endpoint-node mappings were checked for semantic polarity consistency (e.g., whether the normalised node preserved or reversed the direction of the original endpoint term). Rows labelled as *flipped* triggered correction of effect direction and polarity fields. This additional fixing step reduced directionality errors introduced by normalisation.

(7) Claim validation. Each grouped ACU was re-checked against the original abstract with a claim-validation prompt asking whether the normalised claim was *supported*, *not supported*, or *unclear*. Rows labelled *not supported* were dropped. This stage served as a post-extraction consistency check between the normalised claim and the original source text.

(8) Hallmark mapping. A subset of endpoints describing biological mechanisms rather than direct clinical outcomes—for example, molecular, cellular, or physiological indicators—was mapped to the 12 hallmarks of ageing [6]. Deterministic keyword rules were applied first (e.g., *telomere length* → telomere attrition; *mitochondrial respiration* → mitochondrial dysfunction), followed by GPT-5-mini classification for unresolved cases.

(9) Evidence tiering. Each ACU was assigned a 0–100 score and then placed in a quality tier: A (≥85), B (71–84), C (55–70), or D (<55). "Quality-graded ACUs" refers to ACUs with an explicit evidence score and tier. The weighting scheme prioritised endpoint proximity (0.30) and study design quality (0.25), followed by effect credibility (0.18), comparison structure (0.10), confounding adjustment (0.08), sample size (0.05), and population context (0.04). Higher weights were assigned to dimensions with the greatest direct impact on clinical interpretability: whether the outcome is close to survival or functional health, and whether the study design supports stronger causal inference [7]. Tier boundaries were selected manually after inspection of the score distribution to preserve meaningful separation between evidence groups. ACUs sharing the same canonical factor–endpoint pair were then merged into 2,641 claim records for replication and convergence analysis (Table 2).

3 Results

The final corpus comprised 2,987 quality-graded ACUs drawn from 1,797 unique publications (Tables 1 and 2). The evidence-tier distribution was Tier A, 227 ACUs (7.6%); Tier B, 1,583 (53.0%); Tier C, 1,012 (33.9%); and Tier

D, 166 (5.6%). Tier A claims were relatively rare, whereas Tier B contained more than half of the final corpus. This pattern reflects the pipeline design: successive filtering stages retain human-relevant, directional, statistically supported, and longevity-related claims, enriching the retained corpus for moderate-to-strong evidence rather than weak or ambiguous findings.

Of all ACUs, 2,288 (76.6%) came from interventional studies, of which 1,593 (53.3% of all ACUs) came from randomised controlled trials. The dataset covered 890 distinct factors and 879 distinct outcomes. This is broad coverage for a filtered human-only corpus: it indicates that the pipeline captures a wide range of interventions and outcomes rather than a narrow set of repeated claims.

At the same time, the corpus revealed two strong structural patterns. First, only 32 ACUs (1.1%) targeted direct survival or longevity outcomes, while 1,956 (65.5%) measured clinical or functional outcomes, 728 (24.4%) physiological or imaging outcomes, and 272 (9.1%) molecular or ageing-clock outcomes. This "proximity gap" is illustrated in Fig. 2 and shows that most of the field relies on outcomes that are easier to measure than survival but are less closely related to the question of whether an intervention extends life.

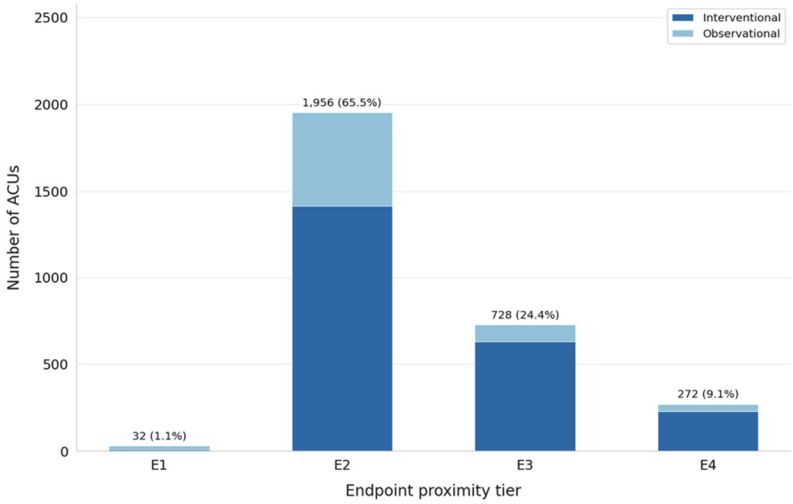

Fig. 2. Number of ACUs by endpoint proximity tier, stacked by study type. E1 = survival/longevity outcomes; E2 = clinical or functional outcomes; E3 = physiological or imaging outcomes; E4 = molecular or ageing-clock outcomes.

Second, the evidence base was broad but shallow. Of 2,641 merged factor–outcome claim pairs, 2,427 (91.9%) were supported by only one study. This distribution is shown in Fig. 3 and indicates that, although the corpus spans many factors and outcomes, most specific claims have not yet been independently replicated.

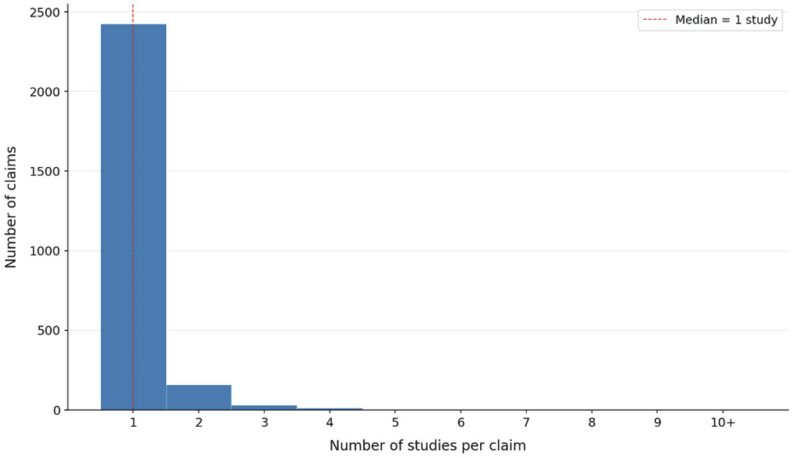

Fig. 3. Replication distribution of merged claim pairs.

The largest factor category was Exercise & physical training, contributing 1,009 ACUs (33.8% of the corpus). The most frequent normalised factors were *resistance training* (214 ACUs), *physical activity/exercise* (141), and *aerobic endurance training* (65). This indicates that exercise dominates the human evidence base not only in popular discourse but also in the structured literature extracted by the pipeline.

4 Conclusions

This work presents a modular NLP pipeline for claim-level synthesis of the human longevity literature. Its main technical contribution is the integration of retrieval, LLM-based screening, stream splitting, type-matched structured extraction, normalisation, polarity correction, claim validation, hallmark mapping, and evidence tiering into a single reproducible system. The results show that the field is broad in scope but limited by sparse replication and

a strong reliance on non-survival endpoints. The full database is publicly accessible at longevityevidence.org.

A limitation of the current study is that formal stepwise evaluation against a human-annotated gold standard has not yet been completed for every critical pipeline stage. Additional limitations include abstract-only processing, potential LLM extraction errors, approximate normalisation of difficult biomedical terms, and the possibility that some relevant studies were missed during screening due to a smaller local model used for the initial relevance filter. Future work will extend the system to full-text processing, contradiction detection, formal module-level evaluation, inclusion of non-human species, and improved screening with stronger models.

References

[1] Wei, J., Wang, X., Schuurmans, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824–24837.

[2] Singhal, K., Azizi, S., Tu, T., et al. (2023). Large language models encode clinical knowledge. Nature, 620, 172–180.

[3] Barardo, D., Thornton, D., Thoppil, H., et al. (2017). The DrugAge database of aging-related drugs. Aging Cell, 16(3), 594–597.

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

[5] Budovsky, A., Craig, T., Wang, J., et al. (2013). LongevityMap: a database of human genetic variants associated with longevity. Aging, 5(1), 15–19.

[6] López-Otín, C., Blasco, M. A., Partridge, L., et al. (2023). Hallmarks of aging: An expanding universe. Cell, 186(2), 243–278.

[7] Cummings, S. R., Kritchevsky, S. B. (2022). Endpoints for geroscience clinical trials: health outcomes, biomarkers, and biologic age. GeroScience, 44, 2925–2931.