

Core document, version 11



A measurement-discipline protocol for longevity practice

Or Hever

Founder and CEO, EVOLVE Health Hub, Warsaw

September 2026

Written for health and exercise professionals. Not medical or legal advice. All worked examples are hypothetical. Local regulation and clinical judgement take precedence.

Companion document: SIIM Evidence Annex, version 2.

Notice

This document is written for health and exercise professionals. It sets out a measurement protocol and the reasoning behind it, and it is intended for professional information and discussion.

It is not medical advice, and it is not legal advice. Reading or applying it creates no professional relationship and no duty of care between the author, EVOLVE Health Hub and any reader, operator or client.

All worked examples, including the clients, values and scenarios described, are hypothetical and constructed for illustration. They do not describe real individuals.

Local law and regulation, professional scope of practice and the clinical judgement of a qualified practitioner take precedence over anything in this document. Operators applying the protocol remain responsible for the care they provide.

Abstract

Longevity practice now has several published frameworks. They describe what a clinic should contain, how it should be governed, which interventions each grade of evidence supports, and which care pathway a person should follow. None of them specifies how large a measured difference must be before it counts as a result.

This paper sets out SIIM, a four-step operating protocol covering Screen, Interpretation, Implementation and Monitoring, designed to sit beneath the existing frameworks. It makes two specific contributions. It separates Implementation from Monitoring, which current pathways combine into a single step. It also anchors the monitoring verdict to the measurement error of each test, where current frameworks rely on consensus thresholds or coordinator judgement.

Each step is described with its inputs, decision rules, failure modes, minimum viable form and documentation requirement, so that an operator can apply the protocol to their own battery. SIIM cites the frameworks discussed here and adapts none of them.

How to read this

SIIM is a foundation layer. A clinical practice and a business model are both built on top of it, using other tools. This paper is written for clinicians deciding whether to build a practice on this foundation.

Two things sit outside its scope. The first is a clinical protocol: what to prescribe and at what grade of evidence is covered by the Delphi consensus [4], and the governance a licensed facility needs is covered by the Swiss framework [3]. The second is a business model: staffing budgets, pricing and unit economics. Section 12 lists these boundaries in full.

On the clinical side, sections 5 to 9 carry the argument, with each rule supported by published sources and a worked example. Section 4 (Guardrails) sets the floor a clinical protocol is built on. It is written as a minimum, and it requires jurisdiction-specific legal review before any public use.

On the business side, two points matter most. On capital, Level 1 compliance across all four steps requires documentation and no specialist equipment (section 13). On staffing, interpretation of blood panels requires a physician in most jurisdictions (section 7, Rule 5, and section 11), and no other step in the protocol makes that demand.

Section 13 also describes the trade that comes with the protocol. An operator using it reports fewer improvements, because part of what would otherwise be called progress sits inside measurement noise. The same discipline is what allows that operator to state, with evidence, which changes are real.

How this document stays current

The paper has two parts. The Core, which you are reading, sets out the protocol, its rules, its failure modes and its limits. It is written to hold whichever test, device or dataset an operator uses, and it should change rarely.

Every specific value the protocol relies on in practice, including reliability figures, retest windows, reference cut-points and the studies behind them, is kept in a companion Evidence Annex with its own date and version. When a better source replaces a figure, the Annex is updated and the Core stays the same, unless the change affects the logic of the protocol. References of the form "Annex A1" point to that document.

The Annex is reviewed at least once a year and each revision is logged. Section 12 lists the conditions under which the Core itself would be revised.

1. What the field already has

Several bodies of work bracket this problem. Each one settled something specific, and the remaining gap is narrower than it first appears.

Facility architecture. Mironov and colleagues set out a structure for a healthy longevity clinic: an analytical centre at the core, clinical units, three-level patient routing from wearable monitoring through in-facility assessment to individualised study, and a classification of technologies by validity and accessibility [1]. Bonnes and colleagues describe the same problem from inside publicly funded hospitals, including the constraints that institutional policy, billing and electronic record interoperability place on what a longevity service can offer [2].

Governance. The Swiss framework proposes hospital-grade governance for longevity medicine clinics: a named responsible physician, an internal ethics board with at least one external member, public quality dashboards, an annual external audit with corrective actions closed within 90 days, and a therapeutic ladder that maps interventions to evidence tiers and regulatory status [3]. It is the most complete governance treatment available, and it is written for licensed facilities.

Evidence grading. A 2026 Delphi consensus of 21 experts adapts the AHA and ESC classes of recommendation to healthspan-oriented care and adds a Class IIc for interventions with no proven benefit yet [4]. Class IIb covers interventions with some signal of benefit, and Class III covers those where the evidence points to no benefit or to harm. The new class solves a real classification problem with an established grading vocabulary.

Care pathway. The second edition of the WHO ICOPE handbook sets out a four-step pathway for primary care: basic assessment with community-level interventions, in-depth assessment, a personalised care plan, and implementation and monitoring [5]. It is the most widely adopted person-centred pathway in this space.

Measurement selection. The TAME Biomarkers Workgroup derived four criteria for selecting biomarkers in geroscience trials: reliability and feasibility of measurement, relevance to the biology of ageing, robust and consistent association with clinical and functional outcomes, and responsiveness to intervention [6]. Applying them reduced 258 candidate biomarkers to 8.

Terminology. Johnson and Shokhirev argue that the output of an ageing clock should be named after its input data, so that an epigenetic clock produces epigenetic age, and that "biological age" should remain an abstract concept [7]. They cite a survey in which 86% of respondents, all ageing researchers, agreed that ageing cannot and should not be measured by a single metric. Their illustration is grip strength: a person with above-average grip may still be ageing poorly in other systems.

Any new framework in this field has to state what it adds.

2. What is still missing

The field's own reviewers describe its problems directly. Forman and Pignolo write that longevity clinics at their best encourage healthy lifestyles, appropriate screening and evidence-based interventions, and at their worst over-diagnose benign problems, endorse unnecessary procedures and expose clients to unproven therapies, in a sector with no formal regulatory oversight and no evidence-based guidelines [8].

Those are intervention problems, and the frameworks above address them well. A second problem remains open: none of the published pathways specifies how large a change must be before it counts.

ICOPE's follow-up process includes repeating the basic or in-depth assessment and documenting any changes. The frequency of follow-up is set according to the intensity of care, the type of intervention and the person's circumstances, and agreed with the person [5]. This is a sound coordination decision. It leaves open what a difference between two measurements means.

The Swiss framework comes closest, from a different direction. It sets analytic validity thresholds: a test-retest ICC of at least 0.9 for functional and frailty indices, an inter-laboratory coefficient of variation of 5% or less for key assays, and, for clinical use of ageing clocks, an ICC of at least 0.90 or a technical error within one year, with decisions based on a single clock avoided [3]. For escalation it requires two visits at least six months apart showing stability or deceleration on at least two validated clocks, plus a pre-specified functional gain such as one additional MET or a half-point fall in a frailty index [3].

These rules answer a different question. An ICC describes how consistently an instrument agrees with itself across a population. A threshold of one MET reflects how much change an expert panel judges clinically meaningful. An operator holding two numbers from one person needs a third answer: whether the difference between them exceeds what the instrument would produce by chance. That quantity is the minimum detectable change (MDC), computed for each test from test-retest data. No framework in this space puts it at the centre of the monitoring verdict.

A second gap

ICOPE combines implementation and monitoring into one step [5], and the Swiss pathway's Optimize phase does the same in practice [3].

Separating them matters. Implementation asks whether the plan was delivered. Monitoring asks whether anything changed. A fully adherent person with no detectable change needs the plan re-interpreted. A partially adherent person with no detectable change has a plan that was never tested. In a combined step the two cases look identical, and they call for opposite responses.

The scope gap

Every framework above assumes a licensed facility. The Swiss model requires a named responsible physician and excludes wellness operators from its designation [3]. ICOPE is written for primary care, and Bonnes and colleagues write from inside hospitals [2].

Many clinicians also oversee or refer to services outside that setting: exercise programmes, corporate health programmes and physiotherapy practices. A measurement discipline that holds in a licensed clinic and carries unchanged into those settings lets a clinician apply one standard across everything they supervise.

What this paper argues

SIIM is a protocol for the measurement discipline beneath the existing frameworks. It is facility-independent and domain-general, and it cites the frameworks above without adapting them. Its two contributions are the separation of Implementation from Monitoring, and a monitoring verdict anchored to the measurement error of each test.

A note on terminology. ICOPE's second edition replaced the term "screening" with "basic assessment", to widen the first step from losses of intrinsic capacity to social support and carer needs [5]. This paper keeps "Screen" for a narrower reason: in a wellness or non-licensed setting, the word signals measurement without implying diagnosis. The two uses are separate and should be read that way.

3. SIIM

The SIIM cycle

Each completed cycle returns to Interpretation; the battery is revised only by deliberate decision

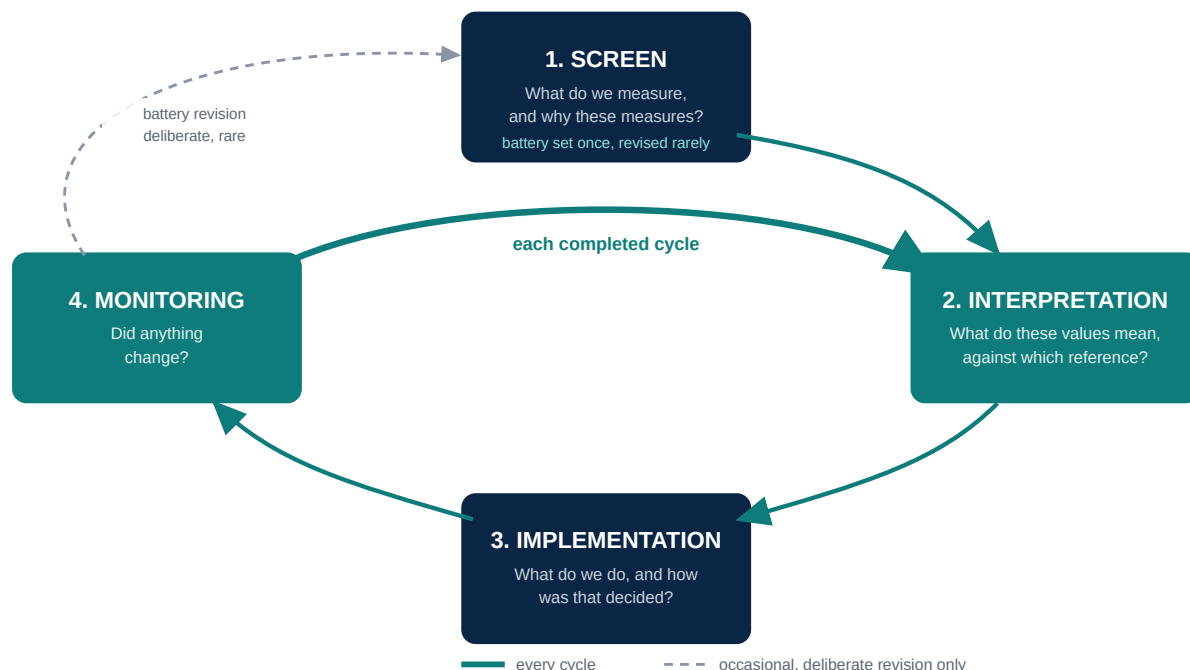


Figure 1. The SIIM cycle. Each completed cycle returns to Interpretation. The battery itself is revised only by deliberate decision.

Table 1. The four steps.

Step	Question	Input	Output
Screen	What do we measure, and why these measures?	Candidate measures	A bounded battery with documented admission decisions
Interpretation	What do these values mean, and against which reference?	Raw values	An ordered problem list with stated uncertainty
Implementation	What do we do, and how was that decided?	Problem list plus client constraints	A prescription with an audit trail
Monitoring	Did anything change?	Retest values plus the prior assessment	A verdict for each measure and an action for the next cycle

Table 2. SIIM alongside published pathways.

	ICOPE [5]	Swiss framework [3]	SIIM
Step 1	Basic assessment	Measure	Screen
Step 2	In-depth assessment	Measure	Interpretation
Step 3	Personalised care plan	Optimize	Implementation
Step 4	Implementation and monitoring, combined	Enhance	Monitoring, as a separate step
Threshold for change	Document any changes	Consensus and analytic validity	Measurement error of each test
Setting	Primary care	Licensed clinic	Any setting

Three properties define the protocol.

It is facility-independent. Nothing in SIIM requires imaging, genomics or a laboratory. Capital extends what can be screened. Compliance depends on the rules being followed.

It is domain-general. The same four steps govern a functional battery, a blood panel or a cognitive assessment. Section 11 sets out what changes between these instances.

It closes to Interpretation. A completed cycle returns to Interpretation. The battery is set once and revised by deliberate decision, because changing measures between cycles breaks comparability, and comparability is what the protocol exists to protect.

4. Guardrails

Every step assumes a person is present. This section sets out what protects that person, whichever step is running.

The published frameworks handle this through licensure: a responsible physician, emergency-equipped premises and a transfer agreement with a hospital. Where the operating model is a licensed facility, those arrangements apply in full. Where it is not, the rules below are the minimum discipline, and they make no claim to equal licensure.

Pre-test screening and exclusion

Before any test in the battery is administered, the operator needs an answer to a specific question: is this test safe for this person today. The answer comes from self-report at intake.

At minimum, intake covers a history of cardiovascular events or symptoms, uncontrolled blood pressure, recent injury or surgery relevant to the tests planned, bone health where a test carries a fall or loading risk, and any condition the operator's scope of practice requires them to recognise as an exclusion. Each admitted test carries its own exclusion list, written at the same time as its decision rule.

A positive flag removes that specific test from the session. Other tests may still be safe and useful. Blanket exclusion and blanket inclusion share the same weakness: both replace a specific decision with a rule of thumb.

Stop criteria during a test

Every test involving meaningful exertion or loading needs stop criteria decided in advance. Chest pain, dizziness, unusual breathlessness and the client's own request to stop are standard triggers, written down for each test alongside its decision rule. A test stopped on a valid trigger is recorded as stopped. Recording it as a low score would corrupt Interpretation, which needs to tell "performed poorly" apart from "did not complete".

Incidental findings and referral

A functional battery will sometimes surface something outside its purpose: an abnormal resting heart rate or blood pressure, a symptom mentioned at intake, or a sign the operator is not qualified to assess.

Two rules apply to every such finding. The finding is recorded, and the client is told that it fell outside the expected range and should be discussed with an appropriate clinician, without the operator characterising, diagnosing or minimising it. The wording of this referral is fixed in advance, under the same communication rule described in Implementation, Rule 5, so that no one improvises language under pressure.

Competency by step

Who is qualified to run SIIM depends on jurisdiction, on the test and on scope of practice, so the protocol asks for that question to be answered separately for each step. The competency needed to administer a test differs from the competency needed to choose a reference frame, and both differ from the standing needed to communicate a finding at the edge of scope.

At minimum, whoever administers a test is trained on that test's protocol. Whoever sets a reference frame or a weighting understands population match and protocol match (Interpretation, Rules 1 and 2). Whoever approves client-facing wording for a gated finding holds the standing that Implementation, Rule 5 requires. An operator should be able to name who currently holds each of these three roles.

Data handling

Functional, blood and self-report data collected under this protocol is personal health data, and in the European Union it falls under the special categories of the GDPR. Three minimums apply. Access is limited to the people who need the data for the roles above. A client can request their own data and its deletion. Any sharing outside the operator's practice, whether with a referral partner, for research or with an employer in a workplace programme, requires specific, informed consent for that sharing, separate from the general consent given at intake.

Missing and incomplete data

A client who cannot complete a test produces a gap. The gap is recorded as missing. It is never scored as the worst possible result, and it is never silently dropped from the record. Where a composite depends on that measure, the coverage rule in Interpretation, Rule 3 applies: the

composite is either withheld or computed and clearly flagged as partial.

Regulatory positioning

A protocol like this one, and any document describing it, sits close to language that regulators treat as medical device or medical advertising language, including when the operator is not a licensed medical facility. Three boundaries follow.

Wellness framing. Mortality-adjacent language, named-disease language and diagnostic thresholds belong inside a licensed clinical relationship. A functional battery run under SIIM measures capacity, and its outputs, including client reports, describe capacity consistently. The same discipline applies in Interpretation, Rule 4: a derivable number is only stated where stating it is appropriate.

Plain language. Where a protocol description or a client report cites research findings or specialist terms, it is written so that a reader without clinical training can follow it. Several jurisdictions treat public-facing material that cites clinical evidence in terms a lay reader cannot follow as a distinct problem, whatever the accuracy of the underlying claim.

Credentials. Where the author of public-facing material holds a clinical credential, the argument stands on its evidence, and the credential belongs in the biography. This is a conservative default, and its treatment in any specific jurisdiction should be confirmed with counsel before publication.

None of this replaces jurisdiction-specific legal review. It is the minimum a self-published protocol document should get right beforehand, so that the review can focus on detail.

5. Screen

What it is

Screen is the deliberate selection of what will be measured, together with an equally deliberate record of what will be left out.

The admission criterion

The TAME Workgroup's four criteria, developed for geroscience trials, transfer to practice with one addition [6]. A measure is admitted if it satisfies all five:

1. **Reliable and feasible measurement.** Published test-retest data exist. Without them, the operator cannot later separate change from noise, and the measure can be collected but cannot be scored.
2. **Relevance** to the construct being assessed.
3. **Robust and consistent association** with the outcomes that matter, holding across datasets and populations.
4. **Responsiveness to intervention** within the timeframe the operator actually works in.
5. **A decision rule written in advance**, together with scope of practice and feasibility within a single visit.

The fifth criterion is SIIM's addition, and in practice it does most of the work. Before administering a measure, the operator writes down what will be done differently depending on the result. If no action changes, the measure produces data without producing information.

The TAME process shows how demanding this is. Of 258 candidates, 8 remained [6]. Telomere length was excluded, despite its wide use, because its association with clinical and functional outcomes is inconsistent. Epigenetic clocks were excluded on cost and on weak evidence of responsiveness to the intervention under study. These were popular measures, set aside by researchers with no commercial interest in the outcome.

Three tiers

Table 3. Measure tiering.

Tier	Definition	Requirements	In the score?	In the report?
Scored	Contributes to the composite and drives prescription	All five criteria met	Yes	Yes
Context	Informs the practitioner, captured when useful	Decision rule may be informal; norms may be immature	No	Yes, labelled as context
Excluded	Considered and set aside, with reasons recorded	Recorded rationale	No	No

The Delphi consensus offers a useful parallel. Its Class IIc holds interventions with no proven benefit yet, separate from Class IIb, where some signal exists, and Class III, where the evidence points to no benefit [4]. The context tier plays the same role for measures. It holds what is plausible and unproven, and keeps it out of a composite where it would gain an authority the evidence does not yet support.

A third data class: self-report

Instrumented tests are one data class. An intake history and a lifestyle questionnaire form a second, with its own strengths and its own rules.

Self-report captures what no instrument can reach: history, symptoms, sleep and lived experience. The published pathways treat it as foundational. ICOPE's basic assessment rests largely on self-report [5], and the Swiss framework opens its clinical pathway with a structured life-course history and standardised screening instruments before any instrumented test [3]. A battery with a thorough set of tests and a thin intake covers only part of what these pathways assess.

Two rules govern admission for self-report, in addition to the general criterion.

Validated instruments, in the language of administration. A translated questionnaire without a validation study in that language is a new instrument with unknown properties. Validation establishes that a given wording, in a given language, measures what it claims and against which threshold. An unvalidated translation belongs in the context tier until it has that evidence.

Self-report gates testing first. The exclusion criteria in section 4 come from intake. This is the first function of Screen, and intake should be built with that function in mind.

A validated instrument, correctly administered and scored, can enter the scored tier on the same terms as any test. For a questionnaire, the known measurement property is usually a published minimal clinically important difference (MCID), described in section 8, Rule 3. Most informal lifestyle

questions fall below this bar and belong in the context tier, where they inform the practitioner without driving prescription on their own.

Failure modes

Panel creep. Each additional measure looks like added value and costs little once the client is present. This is how the over-testing Forman and Pignolo describe enters an otherwise careful practice [8].

Equipment-led selection. A measure enters because equipment was bought. The test is whether the decision rule was written before the invoice.

Silent promotion. A context measure appears in enough reports that it starts to influence prescription without ever meeting the scored-tier criteria.

Construct sprawl. A functional battery gradually accumulates metabolic markers and risk scores until it becomes a general health report with no clear boundary.

Worked example: a candidate measure

An operator considers adding a joint range-of-motion battery. It is quick, clients find it meaningful, and the equipment is inexpensive.

Decision rule. Restricted hip internal rotation would change exercise selection within a block. It would leave the choice of block, its priority and its dose unchanged. The action that differs is a detail of selection.

Reliability. Goniometric range of motion shows inter-rater variability that is large relative to the differences an operator would act on. Published repeatability data exist, and they are weak.

Responsiveness. Good. Range of motion changes within the operator's working timeframe.

Scope and feasibility. Both acceptable, if limited to a few joints.

Verdict: context tier. The measure informs the practitioner and appears in the practitioner-facing record. It stays out of the composite and out of prescription. The reasoning is recorded, so the question does not reopen each time a colleague reads a new mobility paper.

This is the common case. Most candidates fall between clearly in and clearly out, and a two-way choice would either discard something useful or promote something immature.

Minimum viable Screen

Three to five scored measures with published repeatability, covering distinct capacities, each with a written decision rule, plus a log of rejected measures.

What to document

The battery, the decision rule for each scored measure, the repeatability source for each scored measure, the tier of every measure, and the rejection log.

6. Interpretation

What it is

Interpretation converts raw values into severity, ordering and stated confidence.

Why errors here spread

An error in Screen produces a missing measure, which is visible. An error in Monitoring produces a wrong verdict, which the next cycle can catch.

An error in Interpretation sets the target that Implementation prescribes toward and that Monitoring evaluates against. Both later steps then run correctly on a wrong premise, and the resulting report is internally consistent. Nothing in the cycle exposes the error, which is why this step deserves the most care.

Rule 1: name the reference frame

A value has meaning only against a comparison, and the choice of comparison is stated in the output.

Three properties of any reference frame are checked before it is adopted.

Population match. The age, sex distribution, baseline activity and health status of the reference cohort, compared with the operator's own clients. A mismatch can still be used, provided it is declared and the conclusions drawn are weighted accordingly.

Protocol match. Reference data belong to the protocol that produced them. Two datasets for the same measure, collected under different protocols, answer different questions.

The literature shows this directly. The Swiss framework cites grip strength below 26 kg in men and 18 kg in women as associated with a twofold increase in all-cause hospital admissions [3]. The Delphi consensus, published in the same year, gives normative cut-offs of 27 kg in men and 16 kg in women within its assessment of intrinsic capacity [4]. Both are defensible. They come from different sources and serve different purposes, and a number taken without knowing which source it came from is being interpreted against an unnamed frame.

Regional fit. Where reference data exist for the operator's own region, they are preferred over more familiar data from elsewhere.

Rule 2: import constructs, re-derive norms

Much of the functional reference data in wide circulation comes from community-dwelling older adults. Applied to active adults in midlife, these norms produce ceiling effects: the instrument stops discriminating at the point where discrimination is needed. The signs are visible in reports, with scores clustered at the top of the scale, timed tests capped at protocol maxima, and grades that barely vary between clients.

When a measure is adopted from another assessment system, the construct can be imported, and the reference frame is rebuilt for the intended population.

Rule 3: carry uncertainty forward

Every reported score carries a confidence band, and every composite propagates the uncertainty of its components through to the final figure.

A composite computed from partial data is a different quantity from the full composite. Each composite therefore has a minimum coverage threshold, below which it is withheld.

Rule 4: keep constructs separate

Functional capacity and biological age are separate constructs, measured by different methods, and a result in one says nothing about the other.

Johnson and Shokhirev make the same point: name a clock's output after its input data, treat biological age as an abstract concept, and recognise the broad agreement among ageing researchers that no single metric captures ageing [7].

The same reasoning applies to any composite expressed as an age. Such a figure is persuasive and carries less information than the percentiles behind it, because it compresses several domains into one number that invites comparison with chronological age and with unrelated methods. An operator may reasonably choose to withhold it. That choice belongs in the protocol.

Rule 5: weights encode a claim that ages

Where a composite combines several measures, its weights express a view of their relative evidence at a point in time. New cohort data, particularly fully adjusted analyses, can change that view after the weights are set.

The protocol requires three things: weights are treated as revisable, the evidence behind them is reviewed when significant new data appear, and any unresolved gap between current weights and current evidence is documented. Specific weightings can legitimately remain proprietary. The obligation to keep them under review applies to everyone.

Worked example: one value, two reference frames

A 52-year-old woman records a dominant-hand grip strength of 24 kg.

Against reference data from community-dwelling older adults, she sits comfortably above the median. The report reads as reassuring, and no action follows.

Against a distribution for active women in her own age band, she sits below the middle. A strength block enters the prescription.

The measurement is identical in both cases. The interpretation reverses because of a choice that most reports never disclose.

The operator's task is to name the frame used, state its population and protocol, and keep it constant across a client's assessments. Several defensible frames usually exist. Changing frame between two assessments turns a follow-up into a comparison between two different questions.

Minimum viable Interpretation

Each scored measure expressed against a named, population-matched reference distribution, in severity bands, with an interval, and with an explicit order of priority. A composite is optional.

What to document

The reference source for each measure, its population and protocol, the banding thresholds, the uncertainty method, the coverage threshold for any composite, and the review date for any weighting.

7. Implementation

What it is

Implementation converts an ordered problem list into a prescription. The requirement is that the prescription be reproducible and open to inspection: two operators with the same inputs reach the same output, and a third can reconstruct how.

The decision sequence

Table 4. Decision sequence.

Order	Factor	Determines
1	Safety gates	What is excluded outright
2	Severity band	Urgency, and which problem is primary
3	Time budget	Distribution across the week
4	Training experience	Rate of introduction
5	Recovery state	Ceiling on intensity
6	Client preference	Choice between equivalent options

The prescriptions themselves draw on current guidance and syntheses, including the 2026 ACSM position stand on resistance training prescription (Annex A12) and evidence on minimal-dose resistance training (Annex A13).

Each factor draws on a fixed vocabulary, so the same client state produces the same routing. The order is strict. Preference comes after every safety gate, and severity yields to a time budget that would otherwise make the prescription undeliverable.

Rule 1: the population gate

Every intervention study was carried out in a population, and that population is part of its finding. Before evidence informs a prescription, the source cohort is compared with the operator's clients on age, sex distribution, baseline activity and health status. Where they differ materially, the evidence is used at reduced weight and the mismatch is stated.

Much of the strongest balance and mobility evidence comes from cohorts a generation older and frailer than a typical midlife client (Annex A10). Much of the strongest strength evidence comes from cohorts with a mean age in the seventies, where declining capacity may reflect underlying illness as much as a modifiable exposure. These are good studies of a different population.

A study can be strong enough to publish and still too distant from a given cohort to prescribe from. The two thresholds are separate.

Rule 2: proxies prescribe the upstream driver

Some measures are system-wide markers, and training them directly misses the point. Grip strength is the clearest case. Its associations with outcomes are robust across large cohorts (Annex A7), and it reflects whole-body capacity. That association does not transfer to training the grip in isolation.

Where a measure acts as a proxy, the prescription addresses the underlying capacity, and the report makes no claim that improving the marker changes what the marker predicts.

Rule 3: labels match what was measured

If a test captures one condition, the domain label and the prescription language stay within that condition. A single static test supports a statement about that test. Where an existing label is broader than the measurement behind it, the gap is recorded and the label is scheduled for correction.

Rule 4: effect sizes are population-level

Every published effect is an expected value across a group. Each report says so consistently, and the discipline holds best when it is built into the report templates themselves.

Rule 5: governance is a system property

Scope of practice managed through policy documents and training is a weak control in any operation with more than one practitioner.

The stronger form is a gate inside the system. Findings that need clinical judgement to communicate appear in client-facing output as a referral line until a named clinician with the relevant standing approves specific wording. Practitioner-facing output keeps the full detail throughout. While the question is open, the default is restraint.

This achieves, for an operator outside a licensed facility, the purpose the Swiss framework secures through licensure and a named responsible physician [3]. The boundary becomes a state of the system that anyone can inspect.

Worked example: the sequence at work

A client presents with aerobic capacity in the most severe band, strength in the middle range and mobility on track. She has under two hours a week available, no structured training history, and a suppressed recovery indicator.

Safety gates. Nothing excluded.

Severity. Aerobic capacity is primary, strength secondary, mobility on maintenance.

Time budget. Under two hours is the binding constraint, and it applies before any dose is set. The plan becomes two sessions a week, with the secondary strength work built into them.

Experience. A novice. The first block builds tolerance ahead of the aerobic target itself.

Recovery. Suppressed. Intensity is capped for this cycle, whatever the severity band would justify.

Preference. Steady-state work is preferred to intervals. Both are defensible under the cap, so preference decides.

The order is what matters. When urgency overrides time budget and recovery, the result is a technically correct prescription the client will not complete, and at the next assessment non-completion looks exactly like non-response.

Minimum viable Implementation

A written decision sequence with safety gates first, a fixed vocabulary for each input, a recorded population-gate check for every evidence source in use, and a stated rule on who may communicate which findings. A one-page decision table meets this standard.

8. Monitoring

This is where SIIM departs from the published pathways, and it carries the paper's central argument.

Rule 1: separate process from outcome

ICOPE combines implementation and monitoring into one step [5]. SIIM separates them, because they answer different questions and their answers can diverge.

Was the plan delivered? This is a process question, answered by adherence data.

Did anything change? This is an outcome question, answered by retesting.

A client with full adherence and no detectable change has a plan that was delivered and did not work, which calls for re-interpretation. A client with partial adherence and no detectable change has a plan that was never tested, which calls for a conversation about feasibility. Combined into one step, the two cases look identical.

Rule 2: respect the adaptation window

A retest carried out before the relevant adaptation can occur measures noise, however carefully it is administered.

The protocol sets a minimum interval for each type of measure. A retest inside that interval is reported as timing-invalid, which is a different finding from a null result. Treating the two as the same leads an operator to conclude that a plan failed when it never had time to work.

ICOPE sets goal horizons of around three months for short-term goals and six to twelve months for longer-term goals [5], and the Swiss framework requires two visits at least six months apart before escalation [3]. Both cadences are defensible on practical grounds. SIIM asks that each interval be justified by the adaptation physiology of the specific measure. In practice this often produces similar intervals, with a stated reason behind each one (Annex Table 1).

Rule 3: compare against measurement error

Each measure is compared against its own minimum detectable change at 95% confidence, computed from test-retest data for that test.

Table 5. Verdicts.

Condition	Verdict	Action
Change smaller than the MDC	Inconclusive	Hold the plan. Confirm adherence. Retest after a full window.
Favourable change larger than the MDC	Confirmed improvement	Progress the dose, or move to maintenance if the target band is reached.
Unfavourable change larger than the MDC	Confirmed regression	Stop. Re-interpret before progressing. Check adherence, illness and load.

Current reliability sources for each measure, with their review dates, are kept in the Evidence Annex (Table 1), because these are exactly the values that should be replaced when better data appear.

Three different thresholds. The Swiss framework sets analytic validity thresholds such as a test-retest ICC of at least 0.9, and escalation thresholds such as a gain of one MET [3]. Each answers its own question. An ICC is a gate on whether a measure should be used at all. A one-MET threshold is a judgement about importance. An MDC is a statement about detectability: the smallest difference in one person that exceeds the instrument's own variability. Detectability comes first. A change below the MDC has not been observed, whatever its clinical meaning would have been.

Inconclusive is a first-class verdict. It is reported as a result in its own right, with no rounding toward improvement and no omission from the report. A substantial share of what the field reports as progress belongs in this category.

Validated self-report uses its own threshold. A questionnaire is held to Rule 3 through a different statistic. A test's threshold is the MDC, derived from test-retest data, and it answers whether a difference was observed at all. A validated instrument's threshold is usually the MCID, derived from validation studies anchored in patient experience, and it answers whether an observed difference matters. The two are complementary. A raw change in questionnaire points, read against zero, repeats the same error as reading a gait speed retest against zero. Every validated instrument in the scored tier has its own stated MCID, sourced in the Evidence Annex.

Rule 4: repeatability governs change, accuracy governs the absolute value

A device has an accuracy, meaning how close it lands to a criterion measurement, and a repeatability, meaning how consistently it reproduces its own reading. The two properties are independent, and a device can do well on one and poorly on the other.

Submaximal aerobic estimation illustrates this. A validation study of a seismocardiography-based estimate of maximal oxygen uptake reported a mean absolute percentage error above 10% against laboratory measurement, and its authors concluded that the method needs further development before clinical use (Annex A5). Much of that error is systematic, so it largely cancels when the same device measures the same person twice.

The operator's obligation is disclosure. The absolute figure is reported as an estimate, with its accuracy limitation stated. The change between two figures is judged against the repeatability threshold.

A related rule follows. Different estimation equations applied to the same raw test can give different absolute values, so the estimation method is held constant across a client's assessments. A change of method between two assessments makes their difference uninterpretable.

Rule 5: predictive strength and training response are separate questions

A measure can be the strongest predictor in a battery and still a poor marker of progress.

Grip strength is again the clearest case. Its outcome associations are robust across large cohorts (Annex A7). A meta-analysis of exercise training in older adults found only small transfer effects on grip strength, and its authors concluded that grip cannot be recommended as a general marker of functional change for all types of exercise programme (Annex A8). Grip therefore earns its place in the battery for interpretation, and stays out of the account of what improved this cycle.

The TAME criteria list association with outcomes and responsiveness to intervention as two separate requirements [6]. Practice often merges them.

Rule 6: the loop closes to Interpretation

A completed cycle returns to Interpretation. The battery is revised by documented decision, on evidence, and never in response to a single disappointing result.

Worked example: a change inside the noise

A client is retested at fourteen weeks. Maximal gait speed has moved from 1.52 to 1.61 m/s, a gain of about 6%. Many operators would report this as progress.

The largest available reliability study of four-metre gait speed, from a norming sample of 1,320 adults, reports only fair test-retest ICCs and high minimum detectable changes, and its authors conclude that the limited reliability of the measure restricts its use in judging change (Annex A1). A difference of 0.09 m/s sits well inside that noise.

Verdict: inconclusive. Hold the plan, confirm adherence and retest after a further full window. There is no evidence of a dose response to justify progression, and no evidence of failure to justify a change of plan.

Change is judged against measurement error

Worked example: maximal gait speed, 1.52 to 1.61 m/s at fourteen weeks

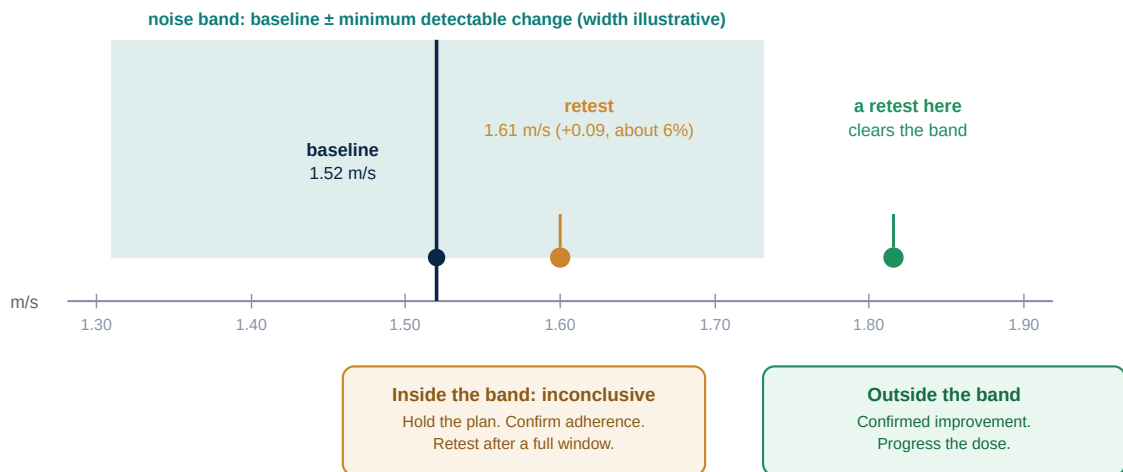


Figure 2. The retest value falls inside the band set by the measure's minimum detectable change, so the verdict is inconclusive. The band width is illustrative; the source value is kept in the Evidence Annex.

Three points follow from this example. The honest report is the more useful one, because progressing load on a change inside the noise is how an operator escalates a plan that was never working and then attributes the plateau to the client. The same client may well show a confirmed improvement elsewhere in the battery, and a cycle with one confirmed improvement, one confirmed regression and three inconclusive results is an ordinary outcome. The source study also illustrates Rule 5: gait speed has strong normative data for Interpretation and weak reliability for Monitoring at short retest intervals.

Minimum viable Monitoring

A stated retest window for each measure, a stated MDC for each measure with its source, all three verdicts reported including inconclusive, a stated next action for each verdict, and process reported separately from outcome.

9. Wearables and continuous data

Continuous data is now the largest input stream available to most operators. The published position on it is candid: Bonnes and colleagues note that evidence-based protocols have yet to be established for the use of such data, the frequency of assessment and its incorporation into the medical record [2]. Mironov and colleagues place remote monitoring and wearables at the base of a three-level routing pyramid, with the emphasis on continuity and volume [1].

That is a sound deployment decision. Measurement governance is a separate question, and SIIM treats continuous data as its own data class with rules at each step.

Table 6. Three data classes.

Property	Standardised test	Self-report instrument	Continuous stream
Measures well	Capacity under defined conditions	History, symptoms, experience	Behaviour in daily life
Measures poorly	Daily behaviour	Actual capacity	Capacity
Change threshold	MDC, from test-retest data	MCID, from validation studies	Trend against the person's own baseline
Main threat	Inconsistent administration	Recall bias, unvalidated wording	Unannounced algorithm changes
Admission requirement	Published repeatability	Published validation in the language used	Raw quantity with documented derivation

A standardised test shows what a person can do under defined conditions and says nothing about the weeks between assessments. A continuous stream shows what the person did, and is a poor measure of capacity. Self-report reaches experience that neither instrument can. The three classes complement each other, and each one fills a gap the others leave.

Screen

The admission criteria apply unchanged, and most derived wearable metrics fail on reliability. A proprietary readiness or recovery score computed by an undisclosed algorithm has no published repeatability, cannot be reproduced outside the vendor, and can change without notice, so it stays out of the scored tier. Raw or near-raw quantities with an established measurement literature, such as step count, sleep duration and resting heart rate, are better candidates, and even these rarely meet the scored-tier criteria.

For most wearable data, the context tier is the permanent home.

Interpretation

Single-day values carry noise that swamps most differences an operator would act on. The unit of interpretation is a multi-week aggregate or trend.

The reference frame rules apply with one addition. Wearable reference distributions usually come from the vendor's own user base, which is undocumented and unrepresentative. Vendor percentiles are therefore an unnamed frame, which Interpretation, Rule 1 excludes. Wearable data is best interpreted within the person over time, against their own multi-week baseline, a frame the operator can name.

Implementation

Two of the six decision factors are exactly what continuous data measures well.

Recovery state, which caps intensity, is a legitimate wearable input: a short-horizon state, best judged against the person's own baseline.

Delivered dose. A prescription assumes a week that may not happen. Continuous data shows whether the prescribed dose was delivered, which is the most useful thing to know before interpreting a disappointing retest. It also captures activity accumulated outside structured sessions, which only continuous capture can measure (Annex A11).

Monitoring

This applies Rule 1 of section 8 to a data class.

Wearable data answers whether the plan was delivered. Standardised tests answer whether capacity changed. Both are outcomes, and each belongs in its own line of the report. Without the continuous stream, an adherent non-responder and a non-adherent, untested client cannot be told apart.

Unannounced method changes. Consumer manufacturers update firmware and derived-metric algorithms without notice and without a version history available to the user. A value that changes between assessments may reflect an algorithm revision. The problem is harder than its laboratory equivalent, where the Swiss framework can at least set inter-laboratory limits [3], because there is usually no changelog, no notice and no way to recompute past values.

Two mitigations apply. Record the device model and firmware version at every assessment, and rely on raw or near-raw quantities for any longitudinal claim.

Worked example: a continuous stream through all four steps

Screen. The vendor's composite readiness score is excluded, because its algorithm is undisclosed and it has no repeatability data. Step count, sleep duration and resting heart rate enter the context tier, each with a decision rule: activity informs the delivered-dose check, sleep informs the recovery input, and resting heart rate is followed as a trend within the person.

Interpretation. Nothing is interpreted below a two-week aggregate, and vendor percentiles are excluded. Each client's first four weeks set their own baseline, which is recorded as the reference frame.

Implementation. The recovery input, drawn from the aggregate, feeds factor five of the decision sequence and caps intensity for the cycle. Delivered dose is checked against prescribed dose at the midpoint, and a shortfall prompts a conversation about feasibility.

Monitoring. The continuous data gives a process verdict: the plan was delivered at about 70% of the prescribed dose. The functional battery gives the capacity verdicts. Both are reported separately, and the process verdict frames how the capacity verdicts are read. Device model and firmware are recorded alongside.

10. Maturity levels

Table 7. Maturity levels.

	Level 1: Compliant	Level 2: Developing	Level 3: Mature
Screen	3 to 5 scored measures, written decision rules, rejection log	Three-tier system in use, repeatability sourced for every scored measure	Battery fixed with a formal revision process, context tier actively managed
Interpretation	Named reference frame, severity bands, stated priority	Uncertainty intervals reported, population and protocol match documented	Composite with coverage threshold, propagated uncertainty, scheduled weighting review
Implementation	Written decision sequence, safety gates first, communication rule stated	Fixed vocabularies, population gate applied and recorded	Governance built into the system, full audit trail for each prescription
Monitoring	Windows and thresholds stated, three verdicts reported, process separate from outcome	Thresholds sourced for each measure, method changes tracked	Thresholds calibrated on the operator's own clients and equipment

For most operators, the single most valuable step is Level 1 Monitoring. It needs no equipment, takes an afternoon of documentation, and changes what the operator can claim more than any addition to the battery.

11. Domain instances

SIIM governs conduct. Between instances, the battery, the reference frames, the adaptation windows and the scope boundary change, and the four steps stay the same.

Functional

Measures of physical capacity across aerobic, strength and mobility domains. Reference frames come from active-adult distributions, with European normative data available for grip strength and body composition (Annex A6, A9). Windows run from weeks to months, and the scope boundary sits at instrumented testing. This instance has the most developed repeatability literature, which makes it the natural first implementation.

Blood biomarkers

Blood panels attract most of the field's commercial attention, and the admission discipline of section 5 has already been applied to them once, in public, by researchers with no commercial stake. The TAME Workgroup reduced 258 candidate biomarkers to 8 using the same criteria SIIM uses: reliability and feasibility, relevance to the underlying biology, robust association with outcomes across populations, and responsiveness to intervention [6]. An operator building a blood panel under SIIM is applying a discipline with a citable precedent, and the two failure modes SIIM warns about, panel creep and equipment-led selection, are exactly what that selection process was designed to resist.

Interpretation faces a sharper form of a familiar problem. Section 6 shows two defensible sources giving different grip cut-offs for the same construct. Blood carries the same problem more acutely, because two laboratories running the same analyte on different assay platforms can report meaningfully different values for the same sample. This follows from the assay itself. A reference range is complete only when the assay platform is named.

A second problem is specific to this domain: some analytes change clinical direction with age. The TAME Workgroup's own example is LDL cholesterol, where higher values are associated with greater risk in middle age and, at advanced ages, lower values may carry the higher risk [6]. A reference frame that omits the age band can therefore point interpretation in the wrong direction entirely.

Pre-analytical variability is a failure mode of its own. Fasting status, time of day, recent exercise, hydration and acute illness at the time of the draw can move a value by more than the change a panel is meant to detect. The analyser may be perfectly precise. The weakness lies in how the draw was conducted, so draw conditions are specified and checked with the same care as a functional test's administration protocol.

Monitoring can use an established tool from clinical chemistry. The reference change value (RCV) combines published within-person biological variation with the analytical variation of the assay to give the smallest difference between two results from the same person that exceeds their combined noise (Annex A14, A15). It answers the question section 8, Rule 3 asks of functional measures, using the same logic, developed independently in another discipline. Where an RCV is published for an analyte, it is the Monitoring threshold for that analyte, standing to blood panels as the MDC stands to functional tests.

Retest windows follow the biology of each analyte. HbA1c has a floor of around twelve weeks set by red cell lifespan. Lipids usually stabilise within six to eight weeks of a change in diet or training. Inflammatory markers such as hs-CRP vary so much within a person that a single reading, however precise, may not be interpretable without a repeat draw (Annex Table 1).

Worked example: a value that needs its age band. A panel returns similar LDL cholesterol values for two clients, one aged 45 and one aged 82. Against a single all-ages range, both receive the same flag. Read correctly, the 45-year-old's result may sit in a range associated with raised risk, while for the 82-year-old a similar value may sit in a range where, in some populations, lower values carry the greater concern [6]. The same number at two ages points in two directions. Interpretation, Rule 1 does more work in this domain than anywhere else in the protocol, because the frame can reverse a finding as well as grade it.

The scope boundary sits with ordering and interpretation authority, which in most jurisdictions means a physician. For this instance, Implementation, Rule 5 is binding.

Cognitive

Domain-specific measures with well-documented practice effects, so the adaptation window rule gains a second requirement: an interval long enough for adaptation, structured to keep learning effects out of the comparison. Reference frames depend heavily on education. The scope boundary generally sits at screening. The Delphi consensus lists standard instruments and thresholds for the cognitive domain of intrinsic capacity [4].

Maturity across instances

The instances differ in maturity. Functional measurement has the best-developed repeatability literature. Blood biomarkers have the best-developed reference ranges, a workable Monitoring threshold from clinical chemistry, and the sharpest age-dependent problem in Interpretation. Cognitive measurement has the hardest retest problem. An operator building a multi-domain service should expect the protocol to expose these differences.

Outside this version

Two domains are left for a later version. Genetic testing behaves differently under SIIM: a germline result stays constant on retest, and the object of periodic review becomes the interpretation of each variant as classifications are updated. Ageing clocks raise a direct question for Interpretation, Rule 4, and their responsiveness to intervention is often weaker than that of the functional measures in this paper. Each will be treated as its own instance.

12. What SIIM does not do

Clinic governance. Leadership, ethics review, audit and transparency are treated comprehensively by the Swiss framework [3], and SIIM leaves them there.

Intervention grading. Evidence tiers for therapeutics are covered by the Delphi consensus [4] and by the Swiss therapeutic ladder [3].

The care pathway. ICOPE remains the reference person-centred pathway [5]. SIIM addresses the measurement discipline within a pathway.

Validation of a measure. The protocol makes reasoning open to inspection. A poorly chosen measure stays uninformative however carefully it is monitored.

Thresholds. Published reliability values reflect other populations and other equipment. Until they are calibrated locally, every verdict inherits that uncertainty. Operators should expect at least one threshold in any battery to be a working convention, and should label it as one.

Evidence gaps. Where intervention evidence for a population does not exist, the population gate exposes the gap and leaves it open.

Individual prediction. Every association behind any battery is population-level. The protocol constrains what may be claimed, and population evidence remains population evidence.

When this Core should be revised

Four conditions would call for a substantive revision of this Core, beyond a routine Annex update.

First, if thresholds calibrated on operators' own clients diverge materially and consistently from the published values the protocol currently relies on, section 8 should be revised to reflect what local calibration shows.

Second, if a domain beyond functional measurement develops a mature literature on repeatability and adaptation windows, its treatment in section 11 should be rewritten by someone working in that domain.

Third, if a future published framework adopts a detectability-based monitoring verdict of the kind argued for here, this document should say so and cite it.

Fourth, if separating Implementation from Monitoring turns out not to change what operators do in practice, the four-step structure should be reconsidered.

None of these conditions has been met in this version. Naming them now makes it possible to notice when one is.

13. Adoption

Three things hold for any operator considering this protocol.

It reduces the number of improvements reported. Part of what would otherwise be reported as improvement falls inside measurement noise. This is a real cost, and it should be stated openly.

It is inexpensive. Level 1 compliance across all four steps requires documentation and no specialist equipment. The expensive part of longevity practice is the capital stack that facility frameworks describe [1], and SIIM requires none of it.

It answers the question the field currently cannot. The Delphi consensus makes the point in another context: reaching guideline targets defines entry to its well-controlled tier and does not in itself constitute evidence of preserved healthspan, since people who meet conventional targets can still show functional decline that standard biomarkers miss [4]. Longevity operators work in that gap, and closing it requires knowing whether a measured change exceeded what the instrument would have produced anyway.

Better infrastructure for these questions is being built. Foundation models trained on longitudinal functional data, and clocks anchored to each person's own peak, are described in current literature, and none is yet available [9]. The gap exists today, in the clients who present this week. A protocol that runs now, with established instruments and honest thresholds, serves those clients in the meantime.

Competing interests

The author is the founder and chief executive of EVOLVE Health Hub, Warsaw, which applies this protocol in its services. The paper received no external funding.

Frameworks referenced

This Core cites the published frameworks that shape its position and argument. Specific measurement values, thresholds and study-level findings used to operate the protocol are cited in the companion Evidence Annex.

- 1 . Mironov S, Borysova O, Morgunov I, Zhou Z, Moskalev A. A framework for an effective healthy longevity clinic. *Aging and Disease*. 2025;16(4):1971-1986. doi:10.14336/AD.2024.0328-1

- 2 . Bonnes SLR, Strauss T, Palmer AK, Hurt RT, Island L, Goshen A, et al. Establishing healthy longevity clinics in publicly funded hospitals. *GeroScience*. 2024;46:4217-4223. doi:10.1007/s11357-024-01132-0
- 3 . Pagani A, Addor G, Axmann G, Bode L, Erat A, Gartenbach A, et al. Toward responsible longevity medicine: Swiss framework for healthy longevity medicine clinics. *Longevity*. 2026;6(1):2583907. doi:10.1080/30653495.2025.2583907
- 4 . Postula M, Kistkins S, Nowak-Szwed A, DeRosa S, Kaminski K, Erat A, et al. Delphi-derived framework for healthspan-oriented care: the cardiometabolic health perspective. *Aging and Disease*. 2026; early access. doi:10.14336/AD.2026.0279
- 5 . World Health Organization. *Integrated care for older people (ICOPE): guidance for person-centred assessment and pathways in primary care*. 2nd ed. Geneva: WHO; 2024. ISBN 978-92-4-010372-6
- 6 . Justice JN, Ferrucci L, Newman AB, Aroda VR, Bahnson JL, Divers J, et al. A framework for selection of blood-based biomarkers for geroscience-guided clinical trials: report from the TAME Biomarkers Workgroup. *GeroScience*. 2018;40:419-436. doi:10.1007/s11357-018-0042-y
- 7 . Johnson AA, Shokhirev MN. Contextualizing aging clocks and properly describing biological age. *Aging Cell*. 2024;23:e14377. doi:10.1111/accel.14377
- 8 . Forman DE, Pignolo RJ. A pragmatic approach to introducing translational geroscience into the clinic. *The Journals of Gerontology, Series A*. 2024;79(9):glae062. doi:10.1093/gerona/glae062
- 9 . Zhavoronkov A, Ying K, Wilczok D. Peakspan: defining, quantifying and extending the boundaries of peak productive lifespan. *Aging and Disease*. 2026; early access. doi:10.14336/AD.2026.0080