Intrinsic Capacity Framework for Geroscience and Healthspan Trials:

A Consensus

Prepared by:

ARPA-H Convened Working Group

(Full participant listed in document)

May 18, 2026

Intrinsic Capacity Consensus Statement to FDA

1. Purpose and Scope

1.1. Purpose of this document

The purpose of this document is to provide FDA with expert consensus recommendations on how intrinsic capacity (IC) and its domains might be measured and used as clinical endpoints and enrichment tools in geroscience/healthspan trials. This document will also outline a stepwise validation plan toward potential future use of IC as a broader gerotherapeutic endpoint.

1.2. Scope and limitations

IC as discussed here is a multi-domain clinical outcome framework (locomotor, cognitive, psychological, sensory, vitality), not a single biomarker. The context of use considered in this document is initially older, community-dwelling adults with early decline in intrinsic capacities, but who remain largely independent in activities of daily living, in geroscience and healthspan trials. In this document, we distinguish between decline in intrinsic capacity and its domains, and later ADL/IADL disability, loss of independence, and major clinical outcomes, which IC may precede and predict.

This document does not claim that IC is a validated surrogate endpoint for mortality or disability nor that a single global IC composite is currently ready as a universal primary endpoint for drug registration. IC measurement may initially function as a clinical outcome framework for detecting early decline in capacities, while also serving as a risk stratifier for subsequent disability, loss of independence, and other major outcomes.

For the purposes of this statement, gerotherapeutics refers to interventions intended to modify biological or physiological processes of aging in ways expected to preserve or improve healthspan. Healthspan refers to the period of life during which

individuals maintain function, independence, and freedom from major chronic disease or disability.

2. Conceptual Framework and Rationale

2.1. Definition of intrinsic capacity

Intrinsic capacity is defined, according to the World Health Organization (WHO), as the composite of an individual's physical and mental capacities that underpin functional ability and healthy aging. IC is conceived as an individual-level construct, distinct from environmental factors and social circumstances, that captures the internal capacities that allow a person to see, move, think, feel, and adapt in daily life.

2.1.1. Domain structure and data-driven development

The IC construct was originally proposed on conceptual grounds and then empirically examined using large longitudinal datasets such as the English Longitudinal Study of Ageing (ELSA), WHO's Study on Global Ageing and Adult Health (SAGE), and other European and Chinese cohorts. In these analyses a broad set of candidate variables, covering mobility, strength, balance, cognition, mood, sensory function, nutrition, respiratory function, and related measures, was entered into exploratory and confirmatory factor analyses. Across multiple datasets and countries, these analyses consistently yielded a five-domain structure:

- **Locomotor / Physical Function:** the capacity to move and perform physical tasks in daily life, reflected by measures such as gait speed, chair stands, balance tests, and related performance measures.
- **Cognitive Function:** the capacity to remember, attend, process information, and make decisions, reflected by performance on memory tests, orientation, executive function, and processing speed measures.
- **Psychological / Mental Health:** the capacity to maintain mood, motivation, and psychological wellbeing, reflected by measures of depressive symptoms, anxiety, distress, affect, and related constructs.

- **Sensory Function:** the capacity to sense the environment well enough to navigate and interact with it, reflected by measures of visual acuity, contrast sensitivity, hearing thresholds, smell, taste, and related sensory indicators.
- **Vitality:** A deeper, more biologic domain thought to represent the capacity to retain capacity or the intrinsic ability to generate and sustain energy, maintain physiological balance, and support other capacities over time. In factor analyses, vitality has been reflected by measures such as grip strength, respiratory function, nutritional status, and elements of fatigue or low energy. Conceptually, vitality captures the underlying physiological reserve that allows locomotor, cognitive, psychological, and sensory functions to be maintained.

These factor analyses supported the view that the five domains are not arbitrary groupings, but empirically coherent dimensions that replicate across cohorts and cultural contexts.

Recent work using the UK Biobank has further quantified the prognostic value of IC domains. Fuentealba et al. (2026) constructed domain-specific mortality risk estimators ("IC age" clocks) using 63 IC-related clinical features and Cox models with up to 18.3 years of follow-up. Each domain's "IC age" strongly correlated with chronological age (e.g. sensory R= 0.96, cognitive R= 0.93, vitality R= 0.91) and predicted 18-year mortality with C-indices in the range of approximately 0.73-0.76. These results reinforce earlier findings that intrinsic capacity, when operationalized at the domain level, is a robust predictor of future mortality and healthspan outcomes in large, independent populations.

2.1.2. Construction of IC scores in prior work

In the initial WHO-related and cohort analyses, each domain was operationalized using a small set of observed variables. Domain scores were typically derived by standardizing each component measure and summing or

averaging within domains or using factor-analytic loadings to form weighted domain scores.

Overall IC scores were then constructed as an average or sum of the domain scores, sometimes with equal weighting of domains, sometimes with weights informed by factor loadings or statistical models. In some implementations, domains were equally weighted for simplicity, in others, loadings from factor analyses provided implicit weighting, although this was not standardized as a formal 'trial endpoint' algorithm.

In practice, this means that IC, as used in ELSA and similar cohorts, was a composite index derived empirically from multiple measures across the five domains, with the exact composition and scaling varying by dataset and analysis. *No single, universally accepted trial-grade scoring algorithm has been established to date, which is one of the gaps this consensus aims to address*.

Recent work also suggests that IC and domain trajectories may differ systematically by sex. For example, UK Biobank analyses of domain-specific IC age found higher IC age acceleration in males than females across all five domains, with the largest difference in locomotor IC age. These findings suggest that sex-specific norms, thresholds, and context-specific analyses may be warranted in future validation work and, in some settings, in trial design.

2.1.3. Associations of IC with mortality and other outcomes

Across multiple longitudinal cohorts, both domain-specific IC measures and overall IC scores have been shown to predict important health outcomes. Lower IC scores, and steeper declines in IC over time, are associated with increased risk of incident disability, higher probability of dependence and care-home admission, greater risk of hospitalization, and increased all-cause mortality. These associations hold even after adjusting for age, sex, and

traditional disease-based measures, indicating that IC captures aspects of underlying vulnerability and reserve that are not fully explained by disease counts alone.

While effect sizes and specific metrics differ across studies, the consistent pattern is that individuals with higher IC (or more favorable IC trajectories) have better survival, lower rates of functional decline, and lower rates of transition to disability and dependence.

This observational evidence provides face validity and prognostic validity for IC as a meaningful healthspan construct and underpins its inclusion in ICD-11 as a "decline in intrinsic capacity" in WHO's ICOPE and related care models.

At the same time, prior IC work has largely focused on population-level prediction and classification, not on the precise trial-grade choices of instruments, scoring rules, weighting, or definitions of meaningful change that are needed for FDA-facing clinical endpoints. **The present consensus builds on this global foundation to propose trial-ready domain definitions and measurement principles, while recognizing that more work is needed to standardize and validate scoring approaches in interventional settings.**

2.2. Intrinsic capacity vs. frailty measures and disease-specific endpoints

2.2.1. IC vs. frailty measures

Frailty measures, particularly the deficit accumulation and phenotype-based frailty measures, have been extremely valuable in aging research and clinical practice. They summarize an individual's vulnerability by counting deficits (symptoms, signs, diseases, disabilities, etc.) or criteria thought to reflect a syndromic loss of reserve and have proven highly predictive of adverse outcomes such as disability, hospitalization, and mortality.

However, IC differs from frailty in several important ways:

- Capacity vs. deficits: Frailty measures emphasize the presence and accumulation of deficits while IC emphasizes underlying capacities that support function, even in the presence of disease.
- Conceptual framing: Frailty is often framed as a late manifestation of vulnerability, while IC is framed as a multi-domain reserve construct that can be tracked and modified earlier in the aging trajectory, before frailty or disability are established.
- Regulatory implications: Both frailty and IC are multi-component constructs, and both can be heterogeneous in their content and interpretation. The distinction is not that IC is 'simple' while frailty is 'complex', but rather that many frailty measures summarize accumulated deficits (often including diseases, symptoms, and disabilities), whereas IC is organized around capacity domains intended to reflect how people move, think, feel, sense, and maintain physiologic reserve. For this reason, we see IC as having stronger potential to function as a multi-domain clinical outcome framework, while frailty measures remain especially valuable as risk stratifiers and prognostic tools.

We see IC and frailty measures as complementary rather than competing constructs. Frailty measures will continue to be essential for risk prediction and stratification of older or less functional individuals. IC and its domains offer a structured way to measure capacity-based outcomes that may respond to interventions and help explain changes in frailty, disability, and healthspan.

2.2.2. IC vs. disease-specific endpoints

Traditional clinical trials in older adults have largely relied on disease-specific endpoints, such as major adverse cardiovascular events (MACE) in cardiometabolic trials, incident dementia or cognitive decline in neurology

trials, or disease-specific scales in rheumatology, oncology, and other specialties. These endpoints are appropriate when the intent is to treat or prevent a particular disease. However, geroscience interventions, those that target aging biology with the aim of slowing multi-system decline and preserving healthspan, raise additional questions.

A single-domain improvement (e.g. only gait speed, or only one cognitive test) can be achieved by interventions that primarily act on one organ system or disease pathway; such effects may be best interpreted as disease- or organ-specific, not gerotherapeutic, even if the measure itself is age-sensitive. To credibly claim that an intervention is acting on aging processes per se, rather than on a single disease, we ultimately expect to see multi-domain benefits across several IC domains, and demonstrated relationships between changes in IC (or its domains) and broader healthspan outcomes, such as disability/dementia-free survival, time to incident age-related multimorbidity, hospitalization, institutionalization, falls, and related clinically meaningful outcomes.

In this sense, IC provides a multi-domain healthspan lens that complements disease-specific endpoints. In the near term, IC domains (especially locomotor and cognitive) could serve as clinically meaningful endpoints in their own right (e.g. age-associate decline in the locomotor domain of IC) and as pre-specified secondary outcomes in trials centered on disability, disease events, or other established endpoints. Over time, as evidence accumulates that IC/domain changes consistently track with major healthspan outcomes across interventions, IC may become a bridge between disease-level effects and broader gerotherapeutic claims.

2.3. Single-domain effects vs. gerotherapeutic effects

Participants in our consensus emphasized that single domain improvements are important and can support meaningful indications (e.g. improved mobility, better executive function), and they may well be the first approvals we see in this space.

Improvements in a single IC domain, particularly domains such as locomotor and cognition, which are themselves multidimensional and clinically meaningful, may still represent important and valuable therapeutic effects. However, in the absence of evidence of benefit in additional domains or on broader healthspan outcomes, such effects are more naturally interpreted as domain-focused rather than definitive evidence of a broad gerotherapeutic effect on aging biology.

Domain-specific modeling using the UK Biobank (Fuentealba et al. 2026) also supports the view that IC domains capture partially distinct aspects of biological aging. IC age acceleration (IC age adjusted for chronological age and sex) showed only modest correlations between domains (no pairwise correlations above ~0.28), with psychological, locomotor, and vitality domains clustering together and cognitive and sensory domains forming a second cluster. Different diseases were preferentially associated with acceleration in specific domains, while most diseases had little effect on cognitive or sensory IC age. These findings support the notion that single-domain improvements, by themselves, are more naturally interpreted as disease- or system-targeted effects, whereas multi-domain IC benefits would be more consistent with a gerotherapeutic effect on aging processes.

We also recognize that some interventions may have both domain-specific and broader geroscience-relevant effects. For example, an intervention may act through a disease-specific mechanism yet also confer more general benefits on functional aging. In such cases, demonstrating a geroscience benefit that is separable from the domain-specific benefit may require showing that the intervention improves IC domains or broader IC measures beyond what would be expected from the single domain effect alone.

Our recommendations therefore adopt a stepwise view:

- Near term:

    - Use IC domains as FDA-interpretable, patient-relevant endpoints (e.g. locomotor IC composite, cognitive IC composite), and as multi-domain secondary outcomes.
    - Accept that many early interventions will show effects in one or a small number of domains, this is still valuable for approvals and for building the IC evidence base.
- Longer-term (gerotherapeutic claims)
    - For broader claims about slowing aging or age-associated decline in IC, it will be important to build strong evidence that changes in IC and its domains are associated with clinically important outcomes such as disability, loss of independence, hospitalization, and mortality. Over time, evidence across multiple cohorts and interventional studies may help determine whether IC can support broader gerotherapeutic claims.

Within this framework, IC is not intended to replace disease-specific endpoints where clear indications exist. Rather, IC offers a structured, multi-domain capacity framework that can be used alongside existing endpoints to characterize geroscience interventions, provide clinically interpretable, patient-relevant outcomes in multiple domains, and can, with appropriate validation, evolve into a broader gerotherapeutic endpoint framework.

3. IC domains and trial-ready measures (Core Trial Battery v1.0)

The Core IC Trial Battery version 1.0 intentionally balances clinical interpretability, scalability, and regulatory familiarity. In some contexts, more objective or physiologic measures may be preferable for initial regulatory acceptance, even if they are less scalable or more resource intensive. Accordingly, version 1.0 distinguishes between a broadly feasible battery and enhanced/context-dependent measures that may be particularly valuable in FDA-facing development programs.

3.1. Target population for version 1.0

The recommendations in this section are intended for older, community-dwelling adults with early decline in intrinsic capacities and at risk for subsequent ADL/IADL

disability, loss of independence, hospitalization, or institutionalization; such as individuals aged approximately > 60 years with mild impairments but not yet severely disabled or institutionalized. This is the population most relevant to near-term geroscience and healthspan trials that aim to preserve or improve function rather than treat single, late-stage diseases.

We anticipate that specific measures, thresholds, and timing may need to be adapted for fitter "younger-old" adults (e.g. 60-65 with few impairments), where ceiling effects require more challenging tests or stress-type protocols or for middle-aged adults (e.g. 50-65), where IC measurement may initially be more useful as a risk stratifier and early marker than as a primary clinical endpoint. However, for version 1.0 we focus on the older at-risk population, while noting where adaptations may be needed.

A major challenge for IC measurement in fitter "younger-old" adults is that several commonly used geriatric instruments show ceiling effects, while in very impaired groups some measures show floor effects. This is one reason version 1.0 is framed as a starting point for older adults with early decline in intrinsic capacities, with future refinement needed for younger and higher-functioning populations.

3.2. Locomotor / physical function domain

3.2.1. Rationale

Locomotor capacity is central to intrinsic capacity. Measurements of mobility and lower-extremity performance, such as gait speed and the Short Physical Performance Battery (SPPB), are among the most robust predictors of disability, falls, hospitalization, institutionalization, and mortality in older adults. They are also widely used and well-understood in clinical research practice.

3.2.2. Recommended Tier-1 measures

For the older at-risk population, we recommend:

- Gait speed (e.g. usual-pace gait speed over 4-6 meters), and

- Short Physical Performance Battery (SPPB), ideally using a continuous or expanded scoring approach to reduce ceiling effects

Key points:

- Gait speed is simple, quick, and strongly predictive of adverse outcomes
- SPPB combines gait speed, chair rises, and balance tasks into a composite score that captures a broader range of locomotor capacity. Several participants emphasized that the chair rise component of SPPB may be among the most predictive and age-sensitive locomotor measures and may deserve particular attention in future refinement of the locomotor battery.
- Standard SPPB scores (0-12) can show ceiling effects in higher-functioning populations; modified scoring (e.g. continuous SPPB or extended scales) has been proposed and appears more sensitive to change and should be preferred where feasible.

3.2.3. Tier-2 / context-dependent/enhanced measures

- 6-minute walk test (6MWT):
    - Provides an endurance and integrated functional capacity measure with established use in several disease areas and strong physiologic relevance.
    - Because it requires more space, training, and repeated standardization than gait speed or SPPB, we treat it in version 1.0 as a context-dependent enhanced measure, rather than a universal requirement. In trials where endurance, cardiometabolic reserve, or integrated physical capacity are central, 6MWT may appropriately function as a Tier-1-equivalent measure.
- Chair rise test (e.g. 5-times sit-to-stand, if SPPB is not used)

    - Captures lower-extremity strength and power; may be more sensitive than gait speed in some contexts
    - Can be more variable; best used within SPPB or, where SPPB is not feasible, as a Tier-2 standalone measure.
- Additional strength/power tests (e.g. 30-second chair stands, knee extensor strength):
    - Highly informative in selected trials, but not required for all IC applications.

3.2.4. Key considerations

- Ceiling effects: In fitter, younger-old participants, both gait speed and SPPB can saturate. In these contexts, trials should consider more challenging tasks, stress-type assessments (e.g. fast gait, endurance walking), and more sensitive strength/power measures.
- Repeat baselines: For endurance or more variable tests, repeated baseline measurements spaced over several days may be considered to improve reliability and reduce learning effects.

3.3. Cognitive domain

3.3.1. Rationale

Age-related cognitive decline, particularly in processing speed and executive function, emerges well before overt dementia and is strongly associated with subsequent functional decline, loss of independence, and health utilization. Traditional global screens like MMSE were designed as dementia screening tools and have significant ceiling effects in higher-functioning older adults, making them poorly suited to detect subtle, aging-related cognitive changes in typical geroscience trials.

3.3.2. Recommended Tier-1 measures

For version 1.0, we recommend a two-component cognitive battery:

- A global cognitive screen:

    - Montreal Cognitive Assessment (MoCA), which provides a brief assessment of multiple domains and is more sensitive than MMSE to mild impairment and early decline.
- A processing speed/ executive function measure:
    - Digit symbol substitution/coding, or
    - Trail Making Test (A and/or B) or similar timed executive/attention tasks

This combination reflects both overall cognitive status and the most age-sensitive aspects of cognition (processing speed and executive function).

Several participants noted that global cognitive screens such as MoCA were developed primarily for detecting clinical impairment rather than subtle aging-related change. For this reason, version 1.0 pairs a global screen with a processing speed/executive function measure, and future work should examine whether more sensitive digital or connectivity-oriented cognitive assessments should replace or supplement MoCA in younger or higher-functioning populations.

3.3.3. Tier-2 / optional measures

- Brief computerized batteries (e.g. NIH Toolbox Cognition, validated tablet-based tasks):
    - May offer advantages in sensitivity, repeatability, and scalability
    - Recommended where infrastructure and participant familiarity allow, recognizing that not all trials can implement digital platforms
- Additional domain-specific tests (e.g. list learning, working memory tasks) may be appropriate in more intensive cognitive trials.

3.3.4. Key considerations

- Practice and anxiety effects: Cognitive tests are vulnerable to practice effects and performance anxiety. Trials should use alternative forms

where available, space assessments appropriately, and consider the impact of test environment and instructions on performance.

- Sensory confounding: Many cognitive tests rely on adequate vision and motor function. Sensory and motor deficits should be measured (as part of the IC sensory and locomotor domains) and considered when interpreting cognitive results.
- MMSE: Given its limited sensitivity and substantial ceiling effects, MMSE is not recommended as a core IC cognitive measure in version 1.0 for general geroscience trials, though it may still be used in specific dementia-focused or very frail populations.

3.4. Psychological domain

3.4.1. Rationale

Psychological capacity, encompassing mood, motivation, anxiety/distress, and psychological wellbeing, is a key component of intrinsic capacity. Depressive symptoms, anxiety, distress, apathy, and related constructs are common in older adults, influence physical activity, adherence, sleep, cognition, and social engagement, and are associated with higher risk of disability, cognitive decline, hospitalization, and mortality.

There is also emerging evidence of biologic links between mood and aging processes. Psychological capacity is therefore important both as a direct determinant of healthspan and as a potential marker of underlying aging biology.

3.4.2. Principle for version 1.0

The psychological domain should not be limited to depressive symptoms alone. For version 1.0, we emphasize trial-ready instruments that are short and feasible in large multi-site trials, well validated in older adults, and together capture both depressive symptoms and anxiety/distress as core aspects of psychological capacity, while acknowledging that broader constructs may be incorporated in future versions.

3.4.3. Recommended Tier-1 measures

For a pragmatic, trial-friendly assessment of psychological capacity, we recommend:

- A brief, depression measure that either has explicit anxiety/distress items (e.g. PHQ-9-based variant that includes anxiety items or distress scale that captures both domains) or is paired with a short anxiety/distress measure so that both mood and anxiety/distress are represented in the psychological domain (e.g. PHQ-9 + GAD-7).
- Optional within Tier-1: Wellbeing/positive affect: Where space allows, we encourage inclusion of a very brief wellbeing/positive affect instrument (e.g. WHO-5 or a PROMIS mental health short form). This adds a positive valence component to psychological capacity, which is conceptually important.

The group did not endorse a single proprietary combined scale at this time but agreed that depression-only coverage is insufficient for the IC psychological domain; at minimum, anxiety/distress should be represented.

3.4.4. Tier-2 / optional measures

- Broader distress measures (e.g. Kessler K-10), particularly in settings with ongoing population mental health surveillance
- Apathy/motivation scales, especially in trials where apathy is a prominent concern
- More comprehensive wellbeing/quality-of-life instruments where deeper psychological profiling is feasible.

These Tier-2 instruments can enrich understanding of psychological capacity and may inform future refinements of the IC psychological domain.

3.4.5. Key considerations

- Regulatory context

    - For drugs primarily indicated for depression or anxiety, these measures are complements, not replacements, for established CNS trial endpoints
    - In geroscience/healthspan trials, psychological domain measures will typically serve as pre-specified secondary outcomes reflecting mood and mental wellbeing, and components of a broader multi-domain IC profile, rather than stand-alone primary endpoints.
- Overlap with vitality and social domains
    - Fatigue, apathy, and social loneliness can straddle psychological, vitality, and social concepts. For version 1.0, we place mood/anxiety/distress primary in the psychological domain, place energy/fatigue predominantly in vitality, and treat social engagement as a social or contextual factor that interacts with IC but is not a separate IC domain.

The version 1.0 configuration ensures that the psychological domain of IC extends beyond depression alone and includes at least one additional dimension (anxiety/distress, and preferably some wellbeing) while remaining implementable in multi-site trials.

3.5. Sensory domain

3.5.1. Rationale

Sensory capacity is critical for functional ability, social engagement, and quality of life. Sensory loss contributes to isolation, loneliness, depression, increases risk of falls, mobility limitations, and cognitive decline, and is associated with mortality in some cohorts. Despite this, sensory measures are often underused or inconsistently collected in aging trials.

3.5.2. Recommended Tier-1 measures

For version 1.0, we recommend a minimal sensory package:

- Visual acuity, measured with standard charts

- A basic hearing assessment (pure-tone audiometry where feasible, or a validated, brief alternative in settings where audiometry is not feasible)

These measures are familiar to clinicians, objectively defined (acuity, thresholds), and reasonably scalable in large trials, especially when audiometry equipment is available.

3.5.3. Tier-2 / enhanced sensory measures

- Contrast sensitivity
    - More sensitive than acuity for functional vision; associated with falls and all-cause mortality
    - Recommended as Tier-2 for trials where visual function is central or where infrastructure allows
- Speech-in-noise tests
    - More functionally relevant than pure tones for hearing; may pick up earlier or more subtle auditory deficits
    - Recommended where feasible and relevant
- Olfactory tests
    - Strongly predictive of mortality and neurodegenerative risk; conceptually attractive as a sensory IC component
    - Currently recommended as Tier-2 due to proprietary test costs, feasibility concerns, and limited trial-grade responsiveness data in general aging populations.

3.5.4. Key considerations

- Treatable vs. intrinsic sensory aging
    - Visual acuity can often be improved with lenses or cataract surgery, which can complicate interpretation if the trial is not targeting eye disease

    - Olfaction and some hearing deficits may more directly reflect intrinsic aging processes, but practical implementation is more complex
    - Gustatory and touch sensation could also be considered in future IC versions when more trial-grade responsiveness data is available, as these are also known to decline with aging.
- Mechanistic vs. clinical roles
    - Tier-1 measures represent the most feasible minimum clinical sensory outcomes for broad version 1.0 implementation. Tier-2 measures such as contrast sensitivity, speech-in-noise, and olfactory testing are also clinically meaningful and may in some settings be more sensitive and more specific to gradual aging-related change than gross acuity or basic hearing thresholds. We classify them as Tier-2 primarily because of current implementation and standardization considerations, not because they are less clinically relevant.

3.6. Vitality domain

3.6.1. Rationale and conceptual stance

Vitality is perhaps the least straightforward but arguably the most central IC domain from a geroscience perspective. Although several candidate components of the vitality domain have been identified, more data are needed to fully define the domain and determine how best to measure it in trials. For now, we conceptualize vitality as: the intrinsic capacity of an individual to generate and sustain energy, maintain physiological balance, and support other capacities over time, as reflected in nutritional status, energy/fatigue, cardiorespiratory, and muscular reserve. Vitality captures the underlying physiological reserve that allows locomotor, cognitive, psychological, and sensory functions to be maintained. Vitality may also be understood clinically as the capacity to remain functionally effective despite physiologic challenges

and obstacles, reflecting both underlying reserve and the ability to sustain function in the face of stressors. Vitality appears to be conceptually related to resilience, understood as a system's behavior in response to stress. That said, there is insufficient evidence to confidently define vitality as resilience or something similar. The group recognized that vitality may overlap conceptually with certain psychological constructs, particularly motivation, drive, and perceived energy, and that these intersections should be explored in future refinement of the IC framework.

In factor analyses of cohorts like ELSA and WHO SAGE, vitality has been represented by variables such as grip strength, respiratory function, nutritional indicators, and fatigue-related measures. The recent UK Biobank analysis by Fuentealba et al. similarly defined a vitality IC age using spirometry, grip strength, hemoglobin, IGF-1, and related features, and showed that vitality IC age predicts mortality and is associated with disease and environmental exposures.

3.6.2. Recommended components

Given current evidence and the lack of a single accepted set of vitality measures, we recommend that vitality be represented by a small set of candidate measures spanning physiologic reserve, energy/fatigability, and nutritional/metabolic state, rather than by any single instrument. Candidate vitality measures discussed by the group included:

- Grip strength, which has repeatedly loaded with vitality in factor analyses and is strongly associated with mortality and disability, although there was substantial debate in the expert group about the specificity of grip strength and its sensitivity to change in shorter trials. For version 1.0, we recommend measuring grip strength when feasible, with its role in composites clarified through empirical analyses in future studies.

- Respiratory function, which has been used in prior IC operationalizations and reflects physiologic reserve;
- Selected biologic markers of reserve or anabolic/metabolic state (e.g. hemoglobin, IGF-1, inflammatory or metabolic measures), particularly for mechanistic and validation studies;
- Nutritional and appetite-related measures, including weight trajectory and brief appetite or nutritional screens, which may provide clinically relevant signals in some contexts but are highly dependent on trial population and mechanism of action.

  The group emphasized that body weight and nutritional measures are highly context dependent and should not be treated as universally good or bad indicators of vitality. For example, intentional weight loss induced by a therapy such as a GLP-1 receptor agonist may be beneficial rather than harmful. Similarly, multidomain nutritional instruments such as the MNA overlap with other IC domains and may be insufficiently specific or sensitive for all trial contexts.

  Simple 1-2 item questions or visual analog scales assessing appetite may provide an early, individualized signal of change in vitality, particularly in interventions expected to affect appetite, weight change, or nutritional intake. Appetite is highly context dependent and subject to ceiling/floor effects, so it should be interpreted as a supportive vitality measure, not a universal stand-alone endpoint.
- VO_2max: Although VO_2-related measures are highly informative regarding cardiorespiratory reserve, interpretation may depend on intervention mechanism. Some agents may alter measured VO_2max without changing real-world functional capacity; therefore, VO_2-related endpoints should be interpreted in the context of complementary functional outcomes.

- Selected inflammatory and immune-related markers, where mechanistically relevant. Vitality may also reflect immune and inflammatory reserve, particularly in the context of resilience to stressors and infections.

3.6.3. Key considerations

- Avoiding circularity: vitality measures should not simply duplicate locomotor measures; we aim for indicators that reflect underlying reserve, not just performance
- Resilience and stress tests: In the longer term, vitality measurement should be expanded to include provocative or stress-response paradigms that more directly assess homeostatic reserve and resilience. Candidate approaches include physiological stress tests, contrasts between usual and maximal performance (e.g. maximal to usual walking speed), and validated fatigability measures that quantify decline in performance or effort under sustained demand. These approaches may help quantify the gap between routine function and reserve capacity, help decrease floor and ceiling effects, and were considered highly promising for inclusion if future tests confirm their ability to accurately represent the vitality domain.

3.7. Summary of core IC trial battery v1.0

This version 1.0 battery is intended to be feasible in multi-site trials, clinically interpretable, and sufficiently aligned with existing evidence to justify its use as a common IC measurement framework in upcoming geroscience and healthspan trials.

We anticipate that this battery will be refined and expanded as empirical data accumulate on measurement properties, domain responsiveness to interventions, and relationships between IC changes and healthspan outcomes. Specifically, many of the Tier-2 could become Tier-1 measures as more data become available.

4. Recommended contexts of use for IC and its domains

4.1. Overview

Intrinsic capacity and its domains can play several roles in geroscience and healthspan trials. Based on our consensus discussions and current evidence, we recommend near-term uses in which IC/domain measures are scientifically justified and likely interpretable to regulators, and we identify future uses that require additional validation.

We emphasize that these recommendations apply initially to older, community-dwelling adults with early decline intrinsic capacities, but not yet severe disability, with notes on extensions to other populations.

4.2. Populations in which IC is appropriate to measure now

- Older at-risk adults (>65, community dwelling)
    - Older adults with mild to moderate mobility or cognitive limitations, or early multimorbidity or frailty risk, but not advanced disability or dementia.
    - This is the primary target population for near-term IC-informed geroscience trials.
- Fitter "younger-old" adults (60-70, higher function)
    - Individuals aged 60-70 with relatively preserved function, but with disease risk factors or early markers of decline.
- Middle-aged adults
    - Individuals without overt disease but with risk factors or subclinical changes.
    - Here, IC may be more suitable as a risk stratifier and longitudinal marker than as a primary clinical endpoint in the near term.

For version 1.0, we recommend that full IC measurement be prioritized in older at-risk adults, while acknowledging that specific domains may be selectively applied in younger populations.

4.3. Roles for IC and its domains in current trials

The usefulness of IC/domain endpoints depends strongly on the expected time horizon of intervention effects and the plausibility that the intervention's mechanism of action can produce measurable change in the selected domain(s) within a practical trial period. If age-related trajectories are slow and the intervention is not expected to alter IC/domain measures detectably in the near term, harder clinical outcomes may be appropriate as primary endpoints.

4.3.1. Enrichment and stratification

Recommended use:

- Use baseline IC domain scores to:
    - Enrich trials for individuals at higher risk of functional decline (e.g. low locomotor IC, low cognitive IC)
    - Stratify randomization and analyses by baseline capacity (e.g. higher vs. lower IC strata)
- Examples:
    - Trials of metabolic or anti-inflammatory agents in older adults enriched for low locomotor IC
    - Trials of neuroprotective or vascular interventions enriched for lower-than expected cognitive IC given age and education

Rationale

- Frailty measures and simple gait speed are already used for enrichment. IC domain scores provide a more structured, multi-domain view of reserve and risk.
- IC-based stratification can help identify subgroups more likely to benefit or be at risk, ensure balance of capacity across arms, and facilitate post-hoc analyses of who responds.

Caveats

- Enrichment strategies should be linked to the intervention's mechanism of action to avoid unnecessarily restricting eligibility
- Use of enrichment strategies based on IC or its components should balance the scientific value of identifying higher-risk individuals against the operational burden of additional screening, exclusion, and stratification requirements. In some contexts, adding IC-based eligibility criteria may increase screen failures, complicate recruitment, and reduce generalizability. Because the operational performance of IC-based screening in large trials has not yet been fully established, version 1.0 recommends that enrichment strategies be piloted and evaluated pragmatically, rather than assumed to improve trial efficiency in all settings. Experience from prior prevention and functional trials suggests that adding complex physiologic or body composition criteria can substantially impede enrollment; similar risks should be considered when implementing IC-based eligibility rules.

4.3.2. Multi-domain secondary clinical outcomes

Recommended use

- In trials that use established primary endpoints (e.g. disease-specific outcomes, disability/dementia-free survival), we recommend including IC global and domain scores as pre-specified secondary outcomes.
- IC secondary outcomes would include changes in domain scores, and change in a global IC composite, if measured.

Rationale

- This role allows IC to:
    - Capture multi-domain effects of geroscience interventions
    - Provide clinically interpretable information about how participants feel and function
    - Build the evidence base linking IC changes to hard outcomes

- IC as a secondary outcome aligns with current regulatory expectations:
    - Primary endpoints remain established clinical outcomes
    - IC provides additional evidence of broad functional impact

4.3.3. Domain-specific primary endpoints

Recommended use

- Selection of domain-specific primary endpoints should be guided not only by conceptual relevance but also by the likelihood that the intervention can induce observable, clinically meaningful change in that domain over the planned follow-up.
- For selective regulatory-facing programs, sponsors may reasonably choose more objective or intensive domain measures where this improves interpretability and acceptance.
- For selected interventions where there is a plausible mechanistic rationale and preliminary data, we recommend using domain-specific IC composites as primary endpoints in older at-risk adults. Examples include:
    - Locomotor IC composite as a primary endpoint for interventions targeting age-associated decline in the locomotor domain of IC (e.g. anabolic, metabolic, or exercise-mimetic agents)
    - Cognitive IC composite as a primary endpoint for interventions targeting age-associated decline in the cognitive domain of IC.

Rationale

- These domain-specific IC endpoints are
    - Clinically meaningful
    - Supported by strong observational evidence linking them to disability, hospitalization, and mortality

    - More interpretable and mechanism-consistent than a single global IC composite at this stage.
- From a regulatory perspective, domain-specific endpoints can be framed as “age-associated decline in the locomotor domain of IC” or “age-associated decline in the cognitive domain of IC”, which are more familiar and actionable than IC as an undifferentiated construct.

Caveats

- For drugs with existing disease indications, IC domain endpoints should supplement, not replace, accepted indication-specific endpoints.
- For gerotherapeutic claims, domain-specific benefits will likely need to be complemented by multi-domain and healthspan outcomes before broader indications can be considered.

4.3.4. Global IC as an endpoint

Near-term stance

- At the present stage of evidence, the group does not endorse a single global IC composite as the default standalone primary endpoint for registration-quality trials across all contexts. However, the group does not exclude the possibility that a global IC score could serve as a primary endpoint in a well-justified program, particularly if accompanied by clear evidence of responsiveness, prospectively defined domain-level analyses, and longitudinal or confirmatory data linking global IC change to clinically important outcomes.

Recommended role

- At this time, global IC composites are recommended as:
    - Secondary or exploratory endpoints
    - Part of multivariate or global tests that examine joint movement across domains

- Tools for validating whether multi-domain IC changes are realistic in trial-time horizons and predictive of healthspan outcomes.

Rationale

- Current evidence demonstrates that domain-level IC measures are strong predictors of outcomes
- However, there is insufficient trial-grade evidence that interventions produce consistent, clinically meaningful changes across all domains or changes in a global IC composite per se, or reliably predict major outcomes across interventions.
- FDA and clinicians will require clear interpretability of which domains moved and assurance that the composite does not mask domain-specific harms.

4.3.5. Surrogate or accelerated-approval uses (future)

Current position

- IC should not yet be considered validated surrogate endpoints for mortality, disability/dementia-free survival, time to incident age-related multimorbidity, hospitalization, institutionalization, falls, or other hard outcomes. At the same time, IC and its domains are intended to function as clinically meaningful outcomes in their own right, not only as proxies for later disability or mortality.

Future potential

- With sufficient evidence that changes in IC domains or composite:
  - Are consistently induced by interventions that improve healthspan and predict improved healthspan outcomes across multiple trials and mechanisms

IC components or composite scores could be considered as candidate surrogate endpoints or endpoints suitable for accelerated approval in specific

contexts, with confirmatory outcome trials. This will require the validation and regulatory roadmap described in Section 6. Although IC is not yet established as a surrogate endpoint, the long-term goal is not only to evaluate IC as a predictor of other outcomes, but also to determine whether decline in IC can itself be treated as a clinically meaningful outcome in defined contexts of use.

5. Composite scoring and reporting principles

5.1. Domain vs. global scores

Intrinsic capacity is inherently multi-domain. For trial use, our consensus is that:

- Each IC domain (locomotor, cognitive, psychological, sensory, vitality) should be measured with domain-appropriate instruments and analyzed and reported separately, because domain-specific effects are clinically meaningful and often mechanistically aligned with interventions.
- A global IC score can be constructed from domain scores, but should serve initially as a secondary or exploratory endpoint and should not replace domain-level analyses.

This approach reflects both scientific reality and regulatory expectations:

- Clinicians and regulators will want to know which domains an intervention affects, not only whether a single index changes.
- For many interventions, especially early gerotherapeutics, it is plausible that one or two domains may respond more strongly than others.

Accordingly, we recommend that:

- Domain scores are treated as the primary objects of interpretation and often as trial primary or key secondary endpoints.
- Global IC scores are reported alongside domains, primarily to characterize multi-domain effects and to inform future surrogate/indication development.

5.2. Composite options for IC v1.0

We agree that domain scores must be analyzed and reported separately, and that many early trials should focus their primary endpoints on domain-specific composites. However, we also see value in constructing a simple global IC score in parallel as a secondary or exploratory endpoint, to begin evaluating whether multi-domain changes emerge and how they relate to healthspan outcomes. We considered three broad composite strategies:

5.2.1. IC-A: Unweighted normalized composite

For trials that choose to construct a global IC score, version 1.0 recommends:

- Normalize each domain score (e.g. into z-scores or 0-100 scales) and compute a global IC composite as the simple average or sum of normalized domain scores.
- Characteristics:
    - Equal weight for all included domains
    - Simple, transparent, and easy to implement and interpret
    - Domain scores always reported and analyzed separately

5.2.2. IC-B: Weighted composite

- Domains (and/or measures within domains) are assigned non-equal weights based on prognostic importance (e.g. strength of association with disability/dementia-free survival or mortality), patient priorities or clinical importance, or underlying biology.
- Characteristics:
    - Potentially more aligned with prognosis or patient-valued outcomes
    - Requires a clear rationale and robust evidence for weights
    - More complex to communicate, particularly as a regulatory endpoint

5.2.3. IC-C: Data-driven composite (PCA/latent/index)

- Use methods such as principal component analysis (PCA) or latent variable modeling to derive a global IC score

- Characteristics:
    - Can more efficiently capture shared variance across domains
    - Useful for research, risk modeling, and mechanistic studies
    - Less directly interpretable to clinicians and regulators as a clinical endpoint, particularly in early stages

5.2.4. Multi-variate/global tests of multiple domains

In addition to single-score composites, we discussed multivariate test procedures (e.g. global tests of multiple domain endpoints, global rank approaches) that treat several domain endpoints jointly and allow a single inferential test of multi-domain improvement while still preserving domain-specific estimates.

We view these as promising for gerotherapeutic hypothesis testing. For version 1.0 we recommend that investigators consider pre-specified multi-variate tests as key secondary analyses, particularly in trials explicitly aiming for multi-domain effects. These approaches can help quantify whether an intervention produces a broader IC signature beyond single domains, without requiring a single composite to serve as the sole primary endpoint.

Version 1.0 recommendation

For version 1.0, and in the absence of sufficient trial data to justify more complex alternatives, the group provisionally favors IC-A (an unweighted normalized composite) when a global IC score is used. This is intended as a pragmatic starting point, not a definitive conclusion that equal weighting is optimal. The group recognizes that intervention-specific effects, variable time horizons, and uneven responsiveness across domains may ultimately support other approaches in future versions. At this time, the group recommends that IC-B and IC-C approaches be treated as research and validation tools, not as the default basis for primary clinical endpoints.

The scoring approaches described here define how IC/domain scores are calculated; they do not by themselves define what magnitude of change is clinically meaningful. Regardless of the global scoring method, domain-specific scores must be reported and analyzed, and any composite should not obscure domain-level harms or heterogeneity. This approach reflects both the scientific uncertainties about how IC domains will respond to interventions and the regulatory need for clarity and interpretability, while still enabling progress toward richer, multi-domain IC endpoints in subsequent versions.

5.3. Weighting

Weighting is a critical decision, and FDA will likely ask why any weighting scheme, including equal weights, is appropriate. For version 1.0:

- We recommend equal weights across domains when constructing a global IC composite for descriptive or secondary endpoint use, unless there is strong, reproducible evidence that specific domains should be weighted differently for a particular indication, and there is a prospectively defined rationale.
- Non-equal weighting (within or across domains) should be explicitly justified and flagged as program-specific or exploratory in early trials.
- Data-driven weighting approaches can be highly useful for risk prediction and understanding biology, but they are not yet appropriate as the primary basis for clinical or regulatory decision-making in version 1.0.

Over time, as evidence accumulates on which domains are most consistently predictive of healthspan outcomes and how different weighting schemes affect interpretability and performance, it may become appropriate to define indication-specific or context-specific weights. For now, simplicity and transparency are prioritized.

5.4. Meaningful change (MCIDs)

The methods used to construct IC and domain scores (e.g. z-score normalization, 0-100 scaling, averaging or summing domains) are conceptually distinct from the

method used to determine whether a change in those scores is clinically meaningful. In other words, scoring defines how raw measurements are converted into domain scores or global IC scores and meaningful change defines how much change in those scores, or in their underlying component measures, corresponds to a difference that matters clinically, or to patients.

5.4.1. IC as a clinical outcome, not a surrogate for now

At present, IC and its domains are positioned as clinical outcome measures (how people move, think, feel, and function), not as validated surrogates for mortality or major disease events. Therefore, meaningful change should be defined primarily in terms of patient-important differences, not solely statistical criteria.

5.4.2. Preferred approach: anchor-based, supplemented by distribution-based methods

We recommend that:

- IC/domain scores be calculated using transparent and standardized scoring procedures
- Interpretation of change in those scores be based on anchor-based approaches, using links to outcomes such as disability, loss of independence, falls, hospitalization, or patient-reported change.
- Distribution-based methods be used as supportive rather than standalone criteria

MCIDs may therefore play several roles, including defining what magnitude of change in a domain measure is clinically meaningful, defining responder thresholds in trials, and informing interpretation of change in a global IC score, once sufficient empirical data are available.

At present, MCIDs are best established for certain component measures and domains, and are less well established for global IC composites.

5.4.3. Domain vs global MCIDs

Given the current evidence base:

- Domain-level MCIDs (especially in locomotor and cognitive domains) should be prioritized for version 1.0:
    - For example, existing literature suggests MCIDs of about 0.5-1.0 points on SPPB, 0.05-0.1m/s in gait speed, or domain specific thresholds in cognitive and psychological scales.
- Global IC MCIDs are more complex and should be developed more cautiously, leveraging:
    - Longitudinal cohort data mapping IC or IC composite changes to risk of disability/dementia-free survival, incident age-related multimorbidity, hospitalization, falls, and other outcomes, and
    - Trials where multi-domain IC changes can be linked to healthspan endpoints

We expect MCID definitions to vary by population, baseline risk, and trial context, and to evolve as empirical data from PROSPR and other trials accrue.

6. Validation and regulatory roadmap

6.1. Goals of validation

To move IC from a widely used clinical and public health framework, already embedded in WHO's ICOPE model and observational studies, into a regulatory-relevant endpoint framework, we need to:

- Demonstrate that IC domains and composites are measurable and responsive in interventional trials.
- Show that changes in IC domains and composites are induced by interventions that plausibly modify aging biology or healthspan, and predict improvements in clinically important outcomes.
- Determine when IC changes represent multi-domain gerotherapeutic effects versus domain-specific, disease-focused effects.

- Define contexts of use where IC/domain endpoints could be primary endpoints in Phase 2 or Phase 3 trials, and/or candidate surrogate endpoints or accelerated-approval endpoints in specific settings.

The roadmap below outlines the steps needed to achieve these goals.

6.2. Existing data sources for IC validation

An important strength of IC is that many of its components are already measured in large studies. Key datasets include:

- Cohorts and population studies:
  - Strong US-based candidates: Health and Retirement Study (HRS), National Health and Aging Trends Study (NHATS), Baltimore Longitudinal Study of Aging (BLSA), Atherosclerosis Risk in Communities (ARIC), Cardiovascular Health Study (CHS), Women's Health Initiative (WHI), Multi-Ethnic Study of Atherosclerosis (MESA), and Framingham Heart Study.
  - English Longitudinal Study of Ageing (ELSA)
  - WHO SAGE and related international cohorts
  - ELSI-Brazil and other regional IC implementations
  - Longitudinal Aging Study Amsterdam (LASA)
  - Singapore Longitudinal Ageing Study (SLAS)
- Clinical trials and large observational trials where IC or IC-like data exist:
  - ASPREE (aspirin and disability/dementia-free survival in older adults)
  - LIFE and Look AHEAD (physical activity and lifestyle interventions with functional outcomes)
  - US POINTER and other multimodal dementia prevention studies
  - SYNERGIC (synergistic effects of exercise, cognitive training and vitamin D with multiple IC-related outcomes)
  - Other geroscience-relevant trials with rich functional and cognitive measures

- UK Biobank and similar resources, where:
    - IC domains have been operationalized, and
    - Domain-specific IC age clocks have been developed that predict mortality and are linked to distinct disease and lifestyle patterns

These datasets can be used to retrospectively construct IC domain and composite scores and to examine their relationship with outcomes.

6.3. Short-term (1-3 years) validation priorities

6.3.1. Retrospective IC construction and outcomes analysis

Construct IC domain scores and simple global IC composites in existing datasets using the version 1.0 measurement framework where possible.

Analyze:

- Baseline IC/domain scores as predictors of healthspan outcomes (disability, dementia, hospitalization, mortality).
- Trajectories of IC domains and composite scores over time and their association with outcomes
- How IC/domain scores compare with frailty measures, gait speed alone, and multimorbidity scores in predicting these outcomes
- Validation studies should explicitly examine whether domain-specific thresholds, trajectories, and meaningful change definitions differ by sex and whether sex-specific scoring or normative references are needed in selected contexts.

6.3.2. Responsiveness and measurement behavior in trials

In trials where repeated measures exist:

- Characterize the responsiveness of IC domains and composite scores to interventions
- Assess ceiling/floor effects, variability, and learning effects for key instruments. Particular priority should be given to quantifying flooring and ceiling effects in proposed IC measures across age and function

strata, especially among “younger-old” adults where many traditional geriatric measures may not provide sufficient dynamic range.

Where possible, estimate effect sizes and MCIDs in real trial data, not just cohorts.

6.3.3. Prospective implementation in geroscience/healthspan trials

The core IC trial battery v1.0 should be incorporated into new trials as domain-specific primary endpoints in selected arms, and multi-domain secondary outcomes and enrichment tools. Trials should ensure prospective collection of all domains and pre-specification of IC/domain analyses.

Future studies should evaluate the value of informant or study partner-reported change, particularly for early cognitive decline and potentially vitality, where family members or close observers may detect subtle change before it is evident to the participant or captured objectively.

Prospective studies should evaluate whether provocative vitality/resilience tests improve sensitivity to early change and better distinguish geroscience-relevant effects from static baseline measures.

Trials implementing IC should, where feasible, also collect geroscience biomarker panels (e.g. molecular aging markers, proteomic or inflammatory panels, biologic age measures) as complementary data. These panels may strengthen the biologic plausibility of observed IC/domain effects, help distinguish domain-specific from broader geroscience-related effects, and support future work on surrogate endpoint development.

6.3.4. Methodological and statistical work

Develop and test multivariate/global test procedures for multiple IC domains, simulation-based power calculations for domain and composite endpoints, and strategies to handle missing data, baseline imbalances, and multicollinearity.

6.3.5. Resilience-related outcomes

Resilience-related outcomes, including acute stressor responses such as severe infection, hospitalization, or mortality following infectious exposures, may provide important complementary evidence that an intervention is affecting underlying physiologic reserve. Prior work with metformin and rapalogs suggests that reduced hospitalization or death from respiratory infections may be relevant to this framework and should be explored in future trial designs.

6.4. Medium-term (3-7 years) goals

With several years of retrospective analyses and prospective IC-informed trials, we expect to be able to:

- Harmonize and pool data across studies
    - Harmonize IC domains and measures across cohorts and trials (where measurement overlap permits).
    - Pool data to evaluate consistency of IC/domain-outcome relationships across populations, interventions, and healthcare systems. The extent to which interventions produce multi-domain vs. single-domain IC changes.
- Assess how intervention-induced changes in IC and its domains relate to clinically important outcomes across interventions, both to refine IC scoring/MCIDs and to inform possible future surrogate or indication development.
    - Examine whether intervention-induced changes in IC domains and global IC:
        - Predict long-term outcomes such as disability/dementia-free survival, independence, and institutionalization, beyond traditional risk factors.
        - Show reasonable consistent relationships across diverse geroscience interventions.

- Refine scoring, weighting, and MCIDs
    - Use cross-trial data to refine domain scoring, global IC composites, and MCID thresholds for domains and composite scores.
- Identify candidate surrogate contexts
    - In specific indications and trial settings where IC domain or composite changes are strongly and consistently linked to improved healthspan outcomes, and interventions repeatedly demonstrate concordant effects on IC and outcomes
    - Explore whether IC/domain measures may be suitable as candidate surrogate endpoints for accelerated approval, or components of composite endpoints in future registration trials.
    - As the field advances, validated biological age biomarkers and molecular surrogates of IC may become important complements to clinical IC endpoints. Version 1.0 does not treat such biomarkers as replacements for IC domains, but the group agrees that this possibility should remain open and should be explicitly pursued in future validation work.

Throughout this phase, early and frequent engagement with FDA will be critical to calibrate expectations and ensure that analytic work aligns with regulatory needs.

6.5. Regulatory engagement and iterative refinement

We propose that sponsors and investigators:

- Introduce IC/domain measures as secondary endpoints and enrichment tools in early trials and discuss them with FDA in pre-IND or Type C interactions.
- For selected trials, propose domain-specific IC composites as primary endpoints accompanied by clear endpoint definitions and MCID considerations, plans for multi-domain and composite IC analyses, and longitudinal follow-up for healthspan outcomes.

- Over time, as evidence accumulates, IC and its domains may be formally submitted for endpoint qualification in defined contexts of use and considered for inclusion in regulatory guidance on geroscience and healthspan endpoints.
- We emphasize that this is a stepwise process. IC version 1.0 is meant to provide a standardized measurement and analysis framework for IC in trials, enable the field to generate consistent, comparable evidence, and support a constructive ongoing dialogue with FDA and other regulators about the future role of IC and its domains in drug development, including, but not limited to, potential surrogate endpoint roles in carefully defined settings.
- Engagement with FDA should begin early, including around prospective uses of IC and its domains in current development programs, so that qualification, IND, or other regulatory pathways can be informed by the data being generated rather than deferred indefinitely.

7. Summary of consensus and sign-off

7.1. Methods and Process

7.1.1. Participants and balance of expertise

This consensus statement was developed by a multidisciplinary group of experts convened by ARPA-H. Participants were selected to provide a balanced representation of relevant perspectives, including geriatric medicine and gerontology (clinical and research), clinical trialists in drugs, biologics, and multimodal/lifestyle interventions, biostatistics and clinical trial methods, intrinsic capacity and WHO/IC/ICOPE subject-matter experts, public health/epidemiology and cohort studies, regulatory science and individuals with direct FDA interaction experience, and health-system and implementation perspectives. The group included both proponents and skeptics of intrinsic capacity as trial endpoints to ensure that a full range of views and concerns were represented.

Participant Names and Authors of this document

Peter Abadir, MD	Johns Hopkins University
David Allison, PhD	Baylor College of Medicine
Karen Bandeen-Roche, PhD	Johns Hopkins University
Nir Barzilai, MD	Albert Einstein College of Medicine
John Beard, MBBS, PhD	Columbia University
Andrew Brack, PhD	ARPA-H
Peggy Cawthon, PhD, MPH	Sutter Health, UCSF
Pinchas Cohen, MD	University of Southern California
Mark Espeland, PhD	Wake Forest University
Sara Espinoza, MD	Cedars-Sinai Medical Center
Luigi Ferrucci, MD, PhD	National Institute on Aging, NIH
Roger Fielding, PhD	Tufts University
Alexander Fleming, MD	Kitalys Institute
Jonathan Gelfond, MD, PhD	UT Health Science Center, San Antonio
Jack Guralnik, MD, PhD, MPH	Maryland School of Medicine
Jamie Justice, PhD	XPRIZE
Steve Kritchevsky, PhD	Wake Forest University
George Kuchel, MD	University of Connecticut
Nathan LeBrasseur, PhD	Mayo Clinic
Jonathan Levy	AstraZeneca
Andrea Maier, MD, PhD	National University of Singapore
Manuel Montero-Odasso, ME, PhD	Western University
John Newman, MD, PhD	The Buck Institute
Ariela Orkaby, MD, MPH	Harvard University
Nicholas Schork, PhD	HonorHealth Research Institute
Bruno Vellas, MD, PhD	IHU HealthAge
Heather Whitson, MD, MHS	Duke University

7.1.2. Overall approach

The process was designed to be consistent with general principles for voluntary consensus standards.

- Openness and balance: multiple disciplines and viewpoints were included, and participation was not limited to advocates of IC.
- Due process and transparency: the process and intermediate outputs (survey results, meeting notes) were shared with participants.
- Consensus, not unanimity: discussion focused on achieving substantial agreement while documenting areas of continuing disagreement.
- Documentation and availability: the final framework and its rationale will be documented in a written statement and made publicly accessible.

7.1.3. Pre-meeting survey

We first conducted a structured pre-meeting survey of invited participants to:

- Assess familiarity and experience with IC, frailty, and related constructs
- Identify which domains participants considered essential for a "core IC set" in healthspan/aging trials
- Collect preferences for specific measures in each domain (locomotor, cognitive, psychology, sensory, vitality)
- Elicit views on
    - Appropriate roles for IC (primary, secondary, composite, enrichment)
    - Timeframes and frequency of assessment
    - Perceived challenges (e.g. definition, measurement, regulatory acceptance)
    - Initial preferences for composite scoring and definitions of clinically meaningful change.

Survey results were anonymized, summarized quantitatively (e.g. frequencies of domain/measure preferences) and qualitatively (free-text themes), and circulated to participants as part of the pre-meeting materials.

7.1.4. Pre-meeting discussions

We then held two virtual pre-meeting discussions:

Pre-meeting 1 – Concept and Context of Use

- Focused on what IC is for and how it should and should not be used in trials at this stage.
- Themes included:
    - IC as a multi-domain capacity construct grounded in WHO'S framework and ICD-11
    - Distinctions between IC and frailty measures or disease-based endpoints
    - Concerns about using a global IC composite as a primary registration endpoint
    - Agreement that near-term uses should emphasize enrichment/stratification, and domain-specific and multi-domain clinical outcomes, rather than surrogate claims.

Pre-meeting 2 – Measurements and Composites

- Focused on the practical measurement of IC domains and potential composite scoring approaches
- Participants discussed candidate Tier-1 and Tier-2 measures for each domain issues of ceiling and floor effects, feasibility, and sensitivity to change, pros and cons of different composite strategies (unweighted, weighted, data-driven), and the need to prioritize domain-level outcomes in early regulatory-facing uses.

Notes from both pre-meetings were shared with participants before the main consensus meeting.

7.1.5. Regulatory input

To ensure that the framework would be interpretable to regulators, ARPA-H sought informal input from an FDA-experienced regulatory advisor (not a geroscientist). This advisor reviewed the draft materials and provided feedback on the likely acceptability of domain-specific IC composites as clinical endpoints, the importance of clear rationale for chosen tests, scoring, and weighting, and appropriate near-term roles for IC (secondary endpoints, enrichment; domain-specific primary endpoints in selected Phase 2 trials), and long-term surrogate aspirations.

This regulatory perspective was shared with participants and used to frame the April 13 discussion, but did not determine the scientific content of the recommendations.

7.1.6. April 13 hybrid consensus meeting

A hybrid (in-person + virtual) consensus meeting was held on April 13, 2026, with the following objectives:

- Define a core IC trial battery v1.0 (domains + trial-grade measures)
- Clarify contexts of use for IC and its domains in current geroscience/healthspan trials
- Agree on composite and reporting principles and a preliminary approach to meaningful change
- Outline a validation and regulatory roadmap to build the evidence base FDA will require for stronger future uses (including possible surrogate roles)

The meeting agenda included:

- A summary of survey results and pre-meeting discussions
- Session on domains and measures, using a draft ‘scorecard’ to guide selection of Tier-1 and Tier-2 measures by domain

- Session on contexts of use to identify appropriate and inappropriate uses for the near term
- Session on composite scoring and meaningful change, comparing unweighted, weighted, and data-driven approaches and discussing domain vs. global scores.
- Session on validation priorities and regulatory engagement, focusing on existing datasets and prospective trials.

Facilitation emphasized inclusion of divergent views and identifying where substantial agreement existed and where differences remained.

7.1.7. Definition of consensus and handling of disagreement

For this effort, consensus was defined in the standard sense used by voluntary consensus bodies: substantial agreement among a balanced group of experts, following review and discussion of the relevant evidence and perspectives, but not requiring unanimity.

Where substantial agreement was reached, this is stated explicitly in the recommendations. Where significant difference of opinion persisted, this statement documents the range of views, and clearly separates version 1.0 recommendations from topics identified as future research or validation priorities

7.1.8. Post-meeting review and quantification of agreement

Following the April 13 meeting:

- A draft of this consensus statement, including proposed Core IC Trial Battery v1.0, contexts of use, composite principles, and validation roadmap, was circulated to all participants for comment and revision.
- Upon sharing the final draft, participants were only included if they provided explicit agreement to be listed as an author. It should be noted that those listed as authors contributed to the discussions. It should not be implied that all authors agree with every item listed in this

document, rather that they agree that what is described is the general consensus of the group as a whole.

7.1.9. Documentation and public availability

ARPA-H intends to make this framework publicly available, so that it can serve as a voluntary consensus reference for sponsors, investigators, and regulators considering the use of intrinsic capacity in geroscience and healthspan trials, and so that it can evolve as new evidence accumulates.

7.2. Next steps

This version 1.0 framework for intrinsic capacity in clinical trials is intended as a starting point, not a final standard. We encourage professional societies, patient organizations, industry consortia, and international partners to review, comment on, and adopt IC v1.0. We also encourage alignment of IC measures across major healthspan-focused initiatives to maximize comparability and learning. Together, these next steps are intended to ensure that intrinsic capacity moves from concept to practice in a way that is scientifically rigorous, clinically meaningful, and ultimately useful to FDA and other regulators considering geroscience interventions.