

Digital Physical-Function Assessment vs. Conventional Manual Testing in Rehabilitation

Why Measurement Workflow Determines Consistency,
Clinical Capacity, and Longitudinal Insight

KINEGRADE

September 2026 | Version 1.0

Abstract

This white paper examines conventional clinician-administered physical-function testing and digitally supported, smartphone-based assessment workflows in rehabilitation and older-adult care. It focuses on a practical distinction that is often overlooked in digital health: the clinical value of a validated test is not determined only by the test itself, but also by how reliably the protocol is executed, recorded, repeated, and interpreted over time. The paper reviews evidence for established functional tests, summarizes what smartphones can and cannot measure, and discusses remote self-assessment, algorithmic interpretation, data governance, and European medical-device considerations. It adopts a technology-agnostic perspective and draws on peer-reviewed clinical and digital-medicine literature as well as current EU regulatory guidance.

1. Introduction

Physical function is routinely evaluated in rehabilitation, geriatrics, chronic-disease management, and research. Common tests such as the Timed Up and Go (TUG), five-times sit-to-stand (5xSTS), 30-second chair stand,

gait-speed tests, balance tasks, and the Short Physical Performance Battery (SPPB) translate observable movement into standardized measures that can be followed over time. The tests themselves are not new. TUG was introduced as a practical measure of basic functional mobility in frail older people, the 30-second chair stand was developed as a measure of lower-body functional strength, and SPPB performance has been associated with disability and other clinically important outcomes (Podsiadlo & Richardson, 1991; Jones et al., 1999; Guralnik et al., 1994).

The operational problem emerges when these measures need to be repeated across many patients, visits, locations, and clinicians. A single stopwatch test can be simple; a longitudinal assessment program is not. It requires protocol selection, instruction, setup, observation, timing or scoring, documentation, interpretation, follow-up scheduling, and comparison with prior results. Each step can introduce delay, inconsistency, or missing data.

Smartphones create a new measurement option because they combine inertial sensors, cameras, microphones, connectivity, and computing in a device that patients and clinicians already know how to use. However, the presence of sensors does not by itself make a clinical test valid, nor does automation remove the need for standardized protocols. The relevant question is therefore not whether a phone can “measure movement” in general, but which clinical constructs can be measured reliably in a defined population and context of use.

2. Conceptual Distinction Between a Test, a Measurement, and a Monitoring Workflow

A clinical test is a protocol: it specifies the task, environment, instructions, equipment, and scoring rules. A measurement is the observed result produced when that protocol is performed. Monitoring is the longitudinal process of repeating comparable measurements and using change over time to inform care. These three layers are related but not interchangeable.

Digitizing only the measurement step may improve timing precision while leaving the broader workflow unchanged. By contrast, a digital assessment system can also standardize instructions, capture protocol metadata, document results automatically, prompt reassessment when due, and present baseline-to-current change. The potential value therefore lies partly in measurement automation and partly in making reassessment a reproducible system rather than an isolated event.

3. Physical-Function Measurement as a Workflow System

Conventional physical-performance tests are deliberately simple. Their strength is that clinicians can administer them with minimal equipment and interpret them within a broader clinical examination. Their simplicity, however, can hide the number of operational steps required to produce a usable longitudinal data point.

For a repeated mobility assessment, the workflow typically includes selecting an appropriate test, confirming safety and eligibility, preparing the environment, giving standardized instructions, ensuring the correct chair or walking distance, observing the task, timing or counting performance, recording contextual details, interpreting the result, documenting it in the patient record, and deciding when the next reassessment should occur. The reliability of longitudinal monitoring depends on the chain, not only the stopwatch.

Protocol variation can matter because functional tests are sensitive to execution conditions. Chair height and arm use affect sit-to-stand performance; TUG depends on the specified walking distance, turn, pace instruction, and chair; gait-speed results depend on course length and timing method; balance measures depend on stance, support surface, vision, and test termination rules. When conditions change between assessments, part of the observed difference may reflect the protocol rather than the patient.

This distinction becomes more important when measurements are distributed across a service organization. A clinic may have a strong protocol on paper while still having incomplete reassessment data in practice because tests are missed, performed differently, or documented in unstructured notes. In that setting, the operational question is not only whether a test is valid, but whether the organization can execute the test consistently at the intended frequency.

3.1. Measurement Error, Biological Change, and Operational Variation

Any repeated measurement contains several sources of variation. There is true biological change in the patient, random measurement error, learning or familiarization effects, day-to-day variation, and operational variation caused by protocol differences. Digital systems can reduce some operational variation by enforcing instructions and capturing metadata, but they cannot eliminate biological variability or make an unsuitable test clinically meaningful.

The correct objective is therefore not maximal automation. It is controlled measurement: preserving the clinical meaning of an established protocol while reducing avoidable friction and making deviations visible. In digital-medicine evaluation, this logic is reflected in frameworks that distinguish verification of the sensing system from analytical validation of the derived measure and clinical validation in the intended context of use (Goldsack et al., 2020).

A digital tool should consequently be evaluated at multiple levels. First, does the sensor capture the underlying signal accurately? Second, does the algorithm convert that signal into the intended metric with acceptable error? Third, is the metric clinically valid for the target population and setting? Fourth, can patients and clinicians actually use the workflow reliably? A failure at any one of these layers can make an otherwise sophisticated system unsuitable for routine care.

3.2. Visual Comparison of Conventional and Digitally Supported Assessment Workflows

The difference between a manual and digitally supported assessment is easiest to see at the workflow level. Both may use the same underlying clinical test. The difference is where instructions, timing, documentation, scheduling, and longitudinal comparison occur.

Conventional clinician-administered workflow	Digitally supported workflow
1. Clinician selects test and prepares setting	1. Protocol assigned with eligibility and setup rules
2. Clinician gives instructions and observes task	2. App guides patient or clinician through standardized task
3. Stopwatch / manual count / visual scoring	3. Sensors assist timing, counting, or movement segmentation
4. Result documented in note or form	4. Result and protocol metadata stored automatically
5. Reassessment depends on follow-up workflow	5. Due date and reminders support reassessment cadence
6. Prior results reviewed manually when needed	6. Baseline-to-current trend is available at point of review

Source: Kinegrade, 2026. The figure compares workflow architecture rather than claiming equivalence of any specific digital implementation.

4. Characteristics and Limits of Conventional Clinician-Administered Testing

Conventional testing remains clinically appropriate in many settings. A trained clinician can observe compensatory movement, recognize pain or instability, adapt the assessment when safety concerns emerge, and integrate performance with history and examination. These are important capabilities that a phone does not replace.

The limitations arise primarily when standardized tests must be repeated at scale. Manual timing and counting consume staff attention; structured results may be copied into free-text documentation; protocol details may not be recorded; and reassessment can be deferred when appointments are busy. The result is not necessarily poor care, but a practical ceiling on how frequently structured measurement can be collected.

These limitations should not be exaggerated. Many performance tests take only seconds or minutes to administer. The operational cost is the complete process around the test, including preparation, documentation, interpretation, and repetition. The business and clinical value of digitization therefore depends on where the real bottleneck is in a specific service. If clinicians already collect and use repeated structured outcomes with high compliance, automation may offer only incremental benefit. If reassessment is inconsistent or documentation is fragmented, workflow support may be materially more valuable.

5. Technical Characteristics of Smartphone-Assisted Functional Assessment

A smartphone is a multi-sensor platform rather than a single measurement modality. Accelerometers and gyroscopes capture linear and rotational motion; cameras can estimate body pose or movement trajectories; microphones can support instruction or event detection; and the device can guide test execution, record patient-reported information, and transmit results. Which sensor should be used depends on the clinical endpoint and the required accuracy.

Evidence already shows that selected standardized tests can be instrumented with consumer mobile devices. Smartphone-based implementations of 5xSTS and TUG have demonstrated strong agreement with reference measurements in older adults, and recent work has reported high concurrent validity for a pocket-worn TUG self-assessment. At the same time, remote assessment studies show that agreement is test-specific: 5xSTS and TUG can perform well, while walking-speed and balance measures may be more sensitive to environment and administration conditions (Chan et al., 2016; Böttinger et al., 2025; van Schooten et al., 2025).

The practical implication is that a digital platform should not claim that “the phone measures physical function” as one undifferentiated capability. Each test requires its own protocol, sensing method, error model, and validation evidence.

Assessment	What a smartphone can plausibly support	Important boundary
5xSTS / 30-s chair stand	Automatic timing, repetition counting, movement-phase segmentation using IMU or video.	Chair height, arm use, phone placement and incomplete stands must be controlled.
Timed Up and Go	Total time and, with validated algorithms, sit-to-stand, walking, turning and sitting phases.	Course distance, turn geometry, pace instruction and device placement affect comparability.
Gait speed	Timing over a defined course; video or inertial sensing may estimate gait parameters.	A known distance and controlled setup are still required; home environments vary.
2-minute walk	Test guidance, elapsed time, movement/step capture and potentially distance when the method is validated.	Indoor distance estimation and route geometry are non-trivial; space requirements remain.
Standing balance	Sway or stability proxies with a worn device; video can assist duration/position scoring.	Sensor location, stance, footwear, surface and safety supervision can materially affect results.
Range of motion	Joint-angle estimation using camera pose or device orientation for selected movements.	Validity is joint-, view- and placement-specific; one generic ROM claim is insufficient.
Grip strength	Workflow, instructions, recording and trend display.	A phone alone does not directly measure force; a dynamometer or external sensor is required.
Pain / self-reported function	Structured digital questionnaires and repeated patient-reported outcomes.	These are self-reported constructs, not objective sensor measurements.

Table 1. Illustrative test-by-test boundaries. A test should be considered digitally measurable only after the specific implementation is validated in its intended population and context.

6. Algorithms, Automation, and Interpretation Layers

Sensor data are not clinical endpoints by themselves. An accelerometer produces acceleration; a gyroscope produces angular velocity; a camera produces image sequences. Software must transform these raw signals into events or metrics such as “stand completed,” “turn started,” “five repetitions completed,” or “TUG duration.” This transformation is where algorithms - including machine learning - enter the measurement chain.

Artificial intelligence can improve event detection or pose estimation, but it cannot remove the need for a defined reference standard. A model that counts chair rises accurately is useful only if the system also recognizes invalid repetitions, handles common variations, and performs acceptably in the target population. A highly accurate algorithm in healthy volunteers may not generalize to people using walking aids, people with severe mobility impairment, or patients who move in atypical patterns.

For early clinical products, interpretable outputs can be preferable to a composite “AI score.” Timing, repetition count, walking speed, task completion, and protocol adherence are clinically legible and can be compared against established test literature. More complex predictive models may eventually add value, but they create additional requirements for validation, explainability, bias assessment, and ongoing performance monitoring.

6.1. Validation Must Be Test-Specific and Context-Specific

The V3 framework proposed for biometric monitoring technologies separates verification, analytical validation, and clinical validation. Applied to functional assessment, verification asks whether the device sensors operate as specified. Analytical validation asks whether the algorithm-derived metric agrees with an appropriate reference. Clinical validation asks whether the metric meaningfully measures the intended functional state in the defined population and use case (Goldsack et al., 2020).

This framework is particularly useful because remote and self-administered testing can change the context of measurement. A 2025 study comparing remote home and in-person assessment in older adults found good agreement for 5xSTS and TUG, while 4-m walk and overall SPPB agreement were more moderate. This is a reminder that transferring a validated clinic protocol into the home is not automatically measurement-equivalent.

6.2. Self-Assessment and Clinician-Guided Assessment

Digital assessment can support several operating models: clinician-administered in the clinic, patient-performed in the clinic, synchronous remote assessment, or asynchronous self-assessment at home. Evidence indicates that older adults can perform selected smartphone-guided or remotely supervised functional tests, and studies of self-administered TUG, sit-to-stand, and gait tasks report promising reliability. However, feasibility depends on patient selection, cognition, balance, environment, device literacy, and the consequences of a measurement error.

The safest model is therefore not “self-assessment for everyone.” A clinically designed platform should support eligibility rules and escalation. Higher-risk patients may require direct clinician supervision, while lower-risk patients may be able to perform standardized reassessments remotely. The system should preserve the distinction between a measurement that is technically available and one that is appropriate for a particular patient.

7. Clinical Workflow, Security, and Data Governance

The largest operational advantage of digital assessment may occur after the measurement is complete. A structured system can associate each result with the test protocol, date, assessor or self-assessment status, device context, and prior results. It can also identify which patients are due for reassessment and present change over time without requiring clinicians to search through narrative notes.

This creates a different information architecture from a standalone measurement app. At the clinician level, the relevant view may be a patient timeline: baseline, most recent value, change, measurement quality, and next planned assessment. At the service-manager level, the relevant questions are different: what proportion of eligible patients were reassessed on schedule, where data are missing, and whether outcomes differ across clinics or pathways. These management-level metrics are only meaningful when the underlying test protocols are sufficiently standardized.

Functional-performance data in healthcare can constitute personal data and, where they reveal information about health, may fall within the special categories of data under Article 9 of the GDPR. Digital assessment therefore requires purpose limitation, data minimization, access control, secure authentication, encryption in transit and at rest where appropriate, logging, retention rules, and clear controller/processor responsibilities. A smartphone-based system does not become privacy-preserving merely because it uses consumer hardware.

For camera-based measurement, data-minimization choices are especially important. A product may not need to retain raw video once a test has been processed. In other cases, retaining selected recordings may be necessary for quality review or validation. The correct architecture depends on the intended use, risk analysis, and evidence needs. The design principle should be explicit: collect and retain only the data necessary for the clinical and technical purpose.

Interoperability also matters. A separate dashboard can be sufficient for a pilot, but routine adoption usually becomes easier when results can be exported or integrated with the organization's electronic health-record and reporting systems. At minimum, structured export should preserve the test name, protocol version, result, units, date/time, and relevant context so that the measurement is interpretable outside the application.

8. Regulatory and Clinical Validation Implications in the European Union

Whether software qualifies as a medical device under the EU Medical Device Regulation (MDR) depends on its intended purpose, not merely on the technology used. Software intended for general purposes or wellness is treated differently from software specifically intended for medical purposes. Where software provides information used to take diagnostic or therapeutic decisions, MDR Rule 11 can place it in Class IIa or higher depending on the potential impact of those decisions. The European Commission updated MDCG 2019-11 guidance on qualification and classification of software in June 2025.

For a functional-assessment product, this means product claims must be designed together with the evidence and regulatory strategy. A tool that records a stopwatch result, a tool that measures a validated performance endpoint, and a tool that recommends a treatment change may carry materially different regulatory implications. Clinical evidence, software lifecycle controls, usability engineering, risk management, cybersecurity, and post-market monitoring become more demanding as the intended medical function and associated risk increase.

8.1. Implications for Product Design and Adoption

The most defensible digital-assessment strategy is to begin with narrow, clinically recognizable endpoints and validate them rigorously. A small set of well-instrumented tests can be more useful than a broad catalogue of poorly characterized "AI measurements." For geriatric rehabilitation and older-adult services, a practical core may include chair-rise performance, mobility, gait or walking endurance, balance, and patient-reported symptoms or function, while strength domains such as grip require dedicated hardware.

The product also has to solve the workflow around these tests. If the only outcome is that a clinician replaces a stopwatch with a phone but still performs all preparation, documentation, scheduling, and comparison manually, the economic benefit may be modest. The more significant opportunity is to standardize the full reassessment loop: assign the protocol, guide execution, validate the attempt, capture the result, document it, compare it with baseline, and surface the next action.

Claims about time savings, increased clinical capacity, or cost reduction should nonetheless be treated as empirical hypotheses. They depend on how much clinician time is currently spent on the relevant workflow, what proportion can safely be shifted to the patient or automated, how often tests are repeated, and whether digital self-assessment introduces support work elsewhere. These outcomes should be measured in pilots rather than assumed.

A useful pilot therefore evaluates both measurement performance and operational performance. Measurement endpoints can include agreement with reference methods, repeatability, invalid-test rate, and missing-data rate. Operational endpoints can include completion rate, staff minutes per completed reassessment, time from due date to completion, documentation time, percentage of eligible patients with an up-to-date measure, and clinician acceptance. Economic value follows only if these operational metrics improve without compromising safety or measurement quality.

9. Conclusion

Digital physical-function assessment should not be understood as a replacement for clinical judgment or as a generic claim that a smartphone can measure all aspects of movement. Its strongest rationale is narrower and more practical: selected standardized tests can be instrumented, guided, and repeated digitally, creating the possibility of more consistent longitudinal measurement with less manual workflow.

The evidence base already supports smartphone or remote implementations of several common mobility and chair-rise tests, but performance is test-specific and context-dependent. TUG and sit-to-stand measures are comparatively promising candidates for digitalization; gait, balance, range-of-motion, and endurance measures require more careful control of environment and sensing method; and grip strength still requires a force-measuring device.

The decisive design principle is therefore protocol-aware measurement. The clinical test defines what should be measured; the sensor and algorithm define how it is captured; the validation program establishes whether the result is trustworthy; and the workflow determines whether repeated measurements actually become available when care teams need them. When these layers are combined, digital assessment can move from being a convenient measurement tool to becoming infrastructure for longitudinal physical-function management.

For Kinegrade, the relevant product thesis is not that every physical assessment should be automated. It is that standardized, repeatable functional measurements should be easier to perform, harder to lose, and more useful over time - in the clinic or, where clinically appropriate, remotely.

Scope note: This white paper is a technical and clinical overview, not medical or legal advice. Product classification and validation requirements depend on the final intended purpose, claims, population, and implementation.

References

- European Parliament and Council of the European Union. Regulation (EU) 2016/679 (General Data Protection Regulation), including Article 9 on special categories of personal data. <https://eur-lex.europa.eu/eli/reg/2016/679/>
- European Parliament and Council of the European Union. Regulation (EU) 2017/745 on medical devices (MDR), Annex VIII, Rule 11. <https://eur-lex.europa.eu/eli/reg/2017/745/>
- European Commission, Medical Device Coordination Group. MDCG 2019-11 rev.1: Qualification and classification of software - Regulation (EU) 2017/745 and Regulation (EU) 2017/746. June 2025. https://health.ec.europa.eu/medical-devices-sector/new-regulations/guidance-mdcg-endorsed-documents-and-other-guidance_en
- Podsiadlo, D., & Richardson, S. The Timed "Up & Go": A Test of Basic Functional Mobility for Frail Elderly Persons. *Journal of the American Geriatrics Society*, 1991;39(2):142-148. doi:10.1111/j.1532-5415.1991.tb01616.x
- Jones, C. J., Rikli, R. E., & Beam, W. C. A 30-s Chair-Stand Test as a Measure of Lower Body Strength in Community-Residing Older Adults. *Research Quarterly for Exercise and Sport*, 1999;70(2):113-119. doi:10.1080/02701367.1999.10608028
- Guralnik, J. M., et al. A Short Physical Performance Battery Assessing Lower Extremity Function: Association with Self-Reported Disability and Prediction of Mortality and Nursing Home Admission. *Journal of Gerontology*, 1994;49(2):M85-M94. doi:10.1093/geronj/49.2.m85
- Bohannon, R. W. Reference Values for the Five-Repetition Sit-to-Stand Test: A Descriptive Meta-Analysis of Data from Elders. *Perceptual and Motor Skills*, 2006;103(1):215-222. doi:10.2466/pms.103.1.215-222
- Connelly, D. M., Thomas, B. K., Cliffe, S. J., Perry, W. M., & Smith, R. E. Clinical Utility of the 2-Minute Walk Test for Older Adults Living in Long-Term Care. *Physiotherapy Canada*, 2009;61(2):78-87. doi:10.3138/physio.61.2.78
- Goldsack, J. C., Coravos, A., Bakker, J. P., et al. Verification, Analytical Validation, and Clinical Validation (V3): The Foundation of Determining Fit-for-Purpose for Biometric Monitoring Technologies. *npj Digital Medicine*, 2020;3:55. doi:10.1038/s41746-020-0260-4
- Coravos, A., Doerr, M., Goldsack, J. C., et al. Modernizing and Designing Evaluation Frameworks for Connected Sensor Technologies in Medicine. *npj Digital Medicine*, 2020;3:37. doi:10.1038/s41746-020-0237-3
- Chan, M. H. M., Keung, D. T. F., Lui, S. Y. T., & Cheung, R. T. H. A Validation Study of a Smartphone Application for Functional Mobility Assessment of the Elderly. *Hong Kong Physiotherapy Journal*, 2016;35:1-4. doi:10.1016/j.hkpj.2015.11.001
- Böttinger, M. J., Mellone, S., Klenk, J., et al. A Smartphone-Based Timed Up and Go Test Self-Assessment for Older Adults: Validity and Reliability Study. *JMIR Aging*, 2025;8:e67322. doi:10.2196/67322
- Bergquist, R., Vereijken, B., Mellone, S., Corzani, M., Helbostad, J. L., & Taraldsen, K. App-Based Self-Administrable Clinical Tests of Physical Function: Development and Usability Study. *JMIR mHealth and uHealth*, 2020;8(4):e16507. doi:10.2196/16507
- Villar, R., et al. Test-Retest Reliability and Agreement of Remote Home-Based Functional Capacity Self-Administered Assessments in Community-Dwelling, Socially Isolated Older Adults. 2024. <https://pmc.ncbi.nlm.nih.gov/articles/PMC11110522/>
- van Schooten, K. S., Steffens, D., Engeler, A., et al. Remote Assessment of Physical Function in Older People: Feasibility, Safety and Agreement with In-Person Administration. *Age and Ageing*, 2025;54(9):afaf266. doi:10.1093/ageing/afaf266
- Validation of a markerless motion capture app for automated scoring of sit-to-stand, Timed Up and Go, and Short Physical Performance Battery tests in adults with chronic disease. 2026. <https://pmc.ncbi.nlm.nih.gov/articles/PMC12773808/>