

Measuring physical capacity and performance in older people

Sarah E Lamb, DPhil^{1&2}

¹Centre for Rehabilitation Research, Nuffield Department of Orthopaedics Rheumatology & Musculoskeletal Sciences, Botnar Research Centre, University of Oxford, Windmill Road, Oxford, OX3 7LD UK,

²Centre for Statistics in Medicine, University of Oxford, Windmill Road, Oxford, OX3 7LD, UK.

Email address: Sarah.Lamb@ndorms.ox.ac.uk. Tel. +44 (0)1865 223462

David J Keene, DPhil³

³Centre for Rehabilitation Research, Nuffield Department of Orthopaedics Rheumatology & Musculoskeletal Sciences, University of Oxford, Critical Care, Trauma and Rehabilitation Trials Group Clinical Trials Office, Kadoorie Centre - Level 3, John Radcliffe Hospital, Headley Way, Oxford OX3 9DU

Email address: david.keene@ndorms.ox.ac.uk. Tel. +44 (0)1865 223121

Corresponding Author: Professor SE Lamb.

Professor SE Lamb, Centre for Rehabilitation Research, Nuffield Department of Orthopaedics Rheumatology & Musculoskeletal Sciences, Botnar Research Centre, University of Oxford, Windmill Road, Oxford, OX3 7LD UK. Email address: Sarah.Lamb@ndorms.ox.ac.uk. Tel. +44 (0)1865 223462

1.

Abstract

We highlight the important differences between the concepts of capacity and performance, and the development of measures and their application in common conditions encountered in health care practice with older people. A number of expert consensus projects have concluded that mobility, balance, muscle strength and dexterity are core domains for capacity measurement in older people. Instruments with evidence of adequate psychometric properties for the evaluation of capacity in response to intervention programmes include the Short Physical Performance Battery, hand grip strength, mini-BEST and 9-hole dexterity test. Measures that are able to track individual change and convey information that can be used to inform clinical decision making, individual prognosis or prediction of events require greater precision. Few such measures are available. Performance measurement usually focuses on basic or instrumental (advanced) Activities of Daily Living (ADLs) performed by people in their usual environments. Finally, we discuss the limitations of physical performance and capacity measures, and future developments that may enhance the use of these measures in health and clinical care.

Keywords

Physical capacity; performance; mobility; hand grip strength.

The difference between performance and capacity

There are a confusing array of terms used across the health and disablement literature. A particular problem is a lack of consensus on standardised definitions for the terms “performance” and “capacity” across major taxonomies and models of disability and healthy ageing (summarised in Table 1). The terms capacity and performance tend to be used interchangeably despite each having a distinct meaning.

The United Nations have adopted the International Classification of Functioning (ICF) as the method of classifying the consequence of disease, impairment, injury and other health conditions [1]. The ICF defines capacity as relating to what an individual can do in a ‘standardised’ environment, usually in a test situation, and hence reflects the environmentally adjusted ability of an individual. The standardised environment is designed to neutralise the varying impact of different environments on an individual and also to allow comparison across people, practitioners, organisations and international boundaries. ‘Performance’ relates to what the person actually does in his or her ‘current’ (usual) environment. In general, capacity measures are taken under conditions where there is no human or other assistance during the test situation. There are a variety of optional and additional qualifiers including the need for assistance. The qualifiers have particular use in institutionalised settings[1]. The gap between capacity and performance reflects the difference between the impacts of current and uniform environments, and thus provides a useful guide as to what can be done to the environment of the individual to improve performance [1].

In their recent report on healthy ageing, the World Health Organisation has proposed a new model of healthy ageing and defined the terms Intrinsic Capacity “as the composite of all the physical and mental capacities that an individual can draw on” and performance as “what individuals do in their current environment, including their involvement in life situations”[2] . Importantly for physical functioning, important aspects of intrinsic capacity affected by ageing include sensory, cognitive, and movement functions. Measures of intrinsic capacity within each of these functions are characterised using standardised tests such as gait speed and hand grip strength. At an overall level, intrinsic capacity is measured and conceptualised in the predominant form as Instrumental Activities of Daily Living (IADLs) and Activities of Daily Living (ADLs). The report authors recognise more work is needed on measurements within the proposed framework. Three levels of intrinsic capacity are recognised – high and stable, declining capacity and significant loss of capacity. The WHO taxonomy has some but not complete consistency with the ICF.

Older models of disablement have used a different taxonomy. The Nagi model underpins much of the conceptual work in the epidemiology of ageing and gerontology particularly the model of disability that is proposed by the Institute of Medicine, USA [3, 4]. In the Nagi model functional limitations are equivalent to capacity within the ICF, and disability is synonymous with measures of the ICF definition of performance.

Table 1

Definitions of performance and capacity in major models of disability and health ageing

	ICF model	Nagi/ Institute of Medicine Model	World Health Organisation
Performance	What a person actually does in his or her environment	Defined as functional limitations.	What individuals do in their current environment including their involvement in life situations
Capacity	What an individual can do in a standardised environment	Described as disability	Intrinsic capacity – the composite of all the physical and mental capacities that an individual can draw on. Generally described in terms of body system functions.

In summary, the taxonomy of disability has evolved in a dis-jointed manner. There is considerable scope for confusion in understanding the difference between performance and capacity measures. However, there is consistency in the concepts if not the terminology. The consistencies in concept can be summarised as:

1. The measurement of ability to complete a task under standardised test conditions is an important element of understanding and monitoring change in the status of older people over time. Standardised tests enable us to measure and isolate key components of the disablement process, to measure across settings and countries, and make direct comparisons between test participants.
2. That the ability to complete a task under standardised test conditions is not necessarily reflective of how an individual functions in their own or other non-standardised environments.
3. The environment is an important determinant of the ability to function in a given space, and the lack of fit or gap between environmental demands and capabilities is an important target for public health and clinical intervention.
4. Important factors to consider in addition to the physical environment are social support and availability of assistive devices, aids and appliances, and the support of other people. These factors should be considered within the measurement framework of performance and capacity.

For the purposes of this paper, measures that capture the capacity element of the ICF framework, and functional limitation in the IOM model, will be referred to as capacity tests. The defining characteristic is that capacity tests are performed under standardised conditions, often using timing or counting of repetitions of movements or tasks. We accept that in some parts of the literature, these measures are referred to as physical performance measures [5, 6].

Measures that capture the performance element of the ICF framework will be referred to as performance tests of daily living (to include Activities of Daily Living, Instrumental Activities of Daily Living and other self-report or self-test methods designed for use in or contextualised to the test recipients' usual environment).

Measurement properties of performance and capacity instruments

In this section we highlight the most important measurement properties of instruments designed to evaluate physical performance and capacity, which include reliability, validity, responsiveness, predictive and diagnostic accuracy, practicality, and feasibility. The relative importance of these different properties varies depending on the intended purpose of the measurement. Practicality considers refusal rates, missing data, and completion. Feasibility covers the demands or burdens placed on those who administer or complete an instrument, such as special training requirements, training time, mode of administration, required time for administration, costs, complexity of scoring, or time to interpret. Information on and evidence of feasibility is seldom readily available.

Reliability is usually reported as reproducibility and internal consistency. Reproducibility is an instrument's stability over time and between settings, and may be assessed using test-retest reliability or levels of inter- or intrarater agreement. The minimum standards for reliability coefficients are set at 0.70 for group comparisons [7] and between 0.90 and 0.95 for individual comparisons [8, 9]. Kappa coefficients may be used in ordinal data to indicate the level of observed agreement greater than that due to chance where a value of 1.0 represents perfect agreement [10]. The strength of agreement has been defined as fair (0.40), moderate (0.60),

substantial (0.80), and excellent (>0.80). Weighted Kappa should be considered, here the Kappa is weighted to account for the potential importance of misclassification [10, 11].

Validity is the extent to which an instrument measures what it claims to measure. Face validity relates to whether a measure depicts the information that it was designed to capture [12].

Content validity concerns the extent to which the component to be measured is comprehensively illustrated by the items in the instrument. Both are assessed through an evaluation of item content and judgment of the relationship with the proposed purpose or by evaluating the involvement of experts and patients in the development process. Quantitative evaluation of content validity can be obtained by comparing an instruments' relationship with other outcome measure instruments or variables. There are no agreed standards on coefficient values, because appropriate correlation coefficients would fluctuate depending on the hypothesized relationship. Content validity should be assessed against a gold standard if one exists. Construct validity is used when there is no gold standard available to assess the validity of an instrument against.

Responsiveness is an instrument's sensitivity in detecting clinical changes of importance to patients, significant others, or healthcare providers. There is no single agreed upon method to assess responsiveness, although there are several statistical methods available (e.g. change scores, effect size, standardised response mean, modified standardized response mean, relative efficiency, sensitivity and specificity, receiver operating characteristics). Effect size is calculated by dividing the raw score change of the measure by the standard deviation of the measure at baseline. An effect size of 0.2 is considered small, 0.5 medium, and greater than 0.8 large[13]. Standardized response mean is obtained by dividing the mean change in scores by the standard

deviation of the change in scores and should not be interpreted along the same thresholds as effect sizes because this may cause over or underestimation of an instrument's responsiveness.

Assessing whether an instrument to measure capacity or performance is fit for purpose.

The most common reasons for using an instrument to measure capacity or performance include predicting an event for an individual (e.g. a fall), to providing a prognosis for an individual (e.g. probability of recovery of mobility after hip fracture), identifying groups of people/populations at risk of events (e.g. a group of people at high risk of falling), measuring change over time within an individual (e.g. monitor an individual's response to intervention), measuring change over time within groups of individuals (e.g. an uncontrolled before and after experiment), and comparing two groups of people receiving different treatments (e.g. within a randomised controlled trial).

Validity, notably face validity, responsiveness, reliability and feasibility are the core of a good measure regardless of what the eventual intended use. A measure that is used to predict health events in a binary response (yes or no) for an individual demand much greater evidence on their predictive characteristics. A traditional measure of risk such as the odds ratio or relative risk provides evidence of an association between two variables. Evidence of an association across time, as is reported in many standard epidemiological studies, is reassuring but by no means confirmatory of adequate predictive and prognostic performance. To understand the performance of instruments that categorise individuals into binary or ordered categories (for example, low, medium or high risk) knowledge of the sensitivity and the specificity of an instrument is useful but sometimes difficult to interpret as estimates are sensitive to the

underlying prevalence of the target event in the sample. Calculation of the positive and negative likelihood ratios (PLR and NLR, respectively), and diagnostic odds ratio (DOR) for prediction provide more robust information. The PLR calculates how many more times a person who experiences a fall in the follow-up year is likely to have a positive test result, and the NLR calculates the number of times a person who has no fall in the follow-up year is likely to have a negative test result. The DOR is the ratio of PLR and NLR, and should ideally be 0.4 [14]. Guidelines suggest that a strategy with strong evidence of predictive ability at the level of the individual would have a PLR ≥ 5 and an NLR ≤ 0.2 [15]. In more recent guidance for the development and validation of prognostic models to predict health events, assessments of discrimination and calibration are recommended [16]. Discrimination refers to an instrument's ability to separate individuals who do and do not experience a health event, summarised by the concordance or C-index (for logistic regression, equivalent to the receiver operating characteristic curve). Calibration is the agreement between observed and predicted probabilities, assessed using calibration plots[17].

Instruments that are used to measure change in health status over time within an individual or within groups of individuals should have published evidence about the absolute level of change that can be considered clinically important. Statistics commonly used are the minimal clinically important difference (MCID) (the smallest change considered worthwhile to patients and clinicians) and the minimal detectable change (the smallest change that can be detected beyond measurement error). The between group difference is usually smaller than the MCID, given that we do not expect all people to respond uniformly to an intervention, and we expect at least some degree of non-compliance or treatment failure.

Instruments to measure physical capacity: rationale and domain choice

Tests of physical capacity are conducted under standardised environmental conditions, with careful control and consideration of aids, appliances and additional support. A person is asked to complete a specified task, and is evaluated in an objective uniform manner using pre-determined criteria which may include counting of repetitions or timing of the activity as appropriate [5].

Such tests have a growing currency in clinical research and increasingly clinical practice. The underpinning rationale that capacity test instruments can detect pre-clinical disability, predict future events, capture health status and track change over time in older people, is based on a well-founded and evidenced rationale [18]. Senescence begins between 20 and 40 years depending on the body cells and systems involved, and has a variable time course between individuals [2]. For the healthiest people, decline in cellular, organ and functional ability is slow and buffered by physiological reserve. Functional consequence might occur first in adaptation, for example in slowing the speed with which common activities are undertaken or reducing the complexity of tasks (for example avoiding dual task activities). Limitation in a functional activity can therefore be defined in terms of speed of completion, form of completion (quality, speed and accuracy), as well as the ability to complete an activity or not. Level of assistance, be that from an aid, appliance or person is also an additional indicator of limitation. Inability to complete a task occurs when the physiological demand of an activity is greater than the physiological capacity of an individual in a given environment. For example, the power required to rise from a chair has to be less than the leg extensor power that can be generated by a person to ensure they can raise their body weight from a chair. A number of factors are

known to impact on the amount of physiological buffer and speed of decline, and those which are modifiable include physical activity and exercise, and presence of disease or acute illness. Importantly, measures of physical capacity have greater predictive validity than diagnoses of chronic disease [19].

When selecting an instrument to measure physical capacity, there should be careful consideration as to the extent to which it relates the domain or underlying construct(s) of interest. The domain selection for physical performance needs to be sufficiently broad to provide a comprehensive picture of ageing and disability, and this knowledge has built over many years of sustained investment in epidemiological research [19-21]. There is broad but no absolute agreement on the domains and instruments that should be selected for evaluating physical capacity. There are numerous consensus statements about what types of measurement domains and/or instruments to include [22-27]. The level of methodological sophistication underpinning the consensus methods also varies. We have selected five recent exercises that have used a triangulation of systematic reviews, national or international consensus, and a recognised and quality controlled method of gaining consensus. The National Institute of Health (NIH) tool box project has undertaken extensive comparative testing of the reliability and validity of a variety of tests that are importantly, feasible across the life span [24]. The Biomarker set is a wider project that has agreed consensus on biomarkers across all body systems in later life, but has not considered feasibility for life span measurement [25]. The Osteoarthritis Research Society International (OARSI) has considered measures for hip and knee osteoarthritis and is driven predominantly by a paradigm of localised disease activity [27].

These consensus statements fill in important gaps in the ICF framework, which although it highlights neuromuscular and movement related functions as being important, provides no indication of specific measurement instruments, other than to state mobility and stability are important domains of physical performance.

All of the consensus statements are consistent in defining locomotion, strength, balance and dexterity as centrally important. With the exception of the Biomarker statements, endurance is an important element. Several of the statements go on to make recommendations for measurement instruments. The NIH tool box is by far the most rigorous in this respect, having undertaken independent testing of reliability and validity in a cohort spanning early to late adulthood. However, in its quest to cover the entire lifespan, the NIH tool box has not considered measures that have been developed and applied only in cohorts of older people, for predicting or measuring health outcome in later life (for example the SPPB). Several of the consensus methods also draw attention to a subtle but important element of interpretation of the test results.

There are a number of consensus statements that confine themselves to one of the identified domains that are important sources of reference for the selection of balance and instability [22, 23]. The Balance Consensus Statement considered a wide range of measures of standing balance, and the Prevention of Fall Injury Europe Consensus statement on measurements and definition of fall events and related injuries.

Table 2: Domains and linked recommendations for capacity test instruments from different consensus groups.

	NIH	Biomarker consensus	OARSI
Domains	[1] Dexterity [2] Strength [3] Locomotion [4] Balance [5] Endurance	[1] Dexterity [2] Strength [3] Locomotion [4] Balance	Not specified (Strength) Not specified (Locomotion) Not specified (Locomotion) [4] Ambulatory transitions [5] Endurance
Tests	[1] Rolyan 9-hole pegboard [2] Jamar hand dynamometer [3] 2 minute walk test [4] NIH standing balance test [5] 4m walk test usual/fast	[1] Pegboard test [2] Handgrip strength [3] Gait speed [4] Standing balance	30s Sit to stand 4*10m fast paced walk test Stair climbing [4] Timed up and go [5] Six minute walk test

Numbers are used to indicate corresponding tests and domain areas.

Measures of mobility and stability (locomotion and balance)

Mobility is the ability to move from place to place and includes walking and the transferring between different body positions. Locomotion is the act or power of moving from place to place. Mobility/locomotor limitation is the most prevalent type of disability in older people and a substantial concern to individuals, families and carers, health providers and society. Stability or balance is a key component of mobility. Measures of balance capacity can be either static tasks (holding balance in pre-defined postures) or dynamic balance (holding balance during movements). Additional levels of difficulty can be added to balance measures through the inclusion of cognitive or sensory challenges during the test procedures, such as blind-folding, counting backwards, and perturbation. More sophisticated methods of balance and mobility assessment are available but rely on equipment that is too costly for routine use or application.

There are many published instruments to measure mobility and stability in routine settings.

The key choices centre on whether to select:

1. A single task such as walking or multiple tasks such as chair rise, balance and walking speed all within one instrument (a multi-component test).
2. A multi-component test that allows the user to score each element separately as well as overall or not.
3. The constraints of the test situation. Long test distances are difficult to administer in some settings, particularly domestic settings.
4. Evidence of measurement validity and responsiveness.

The Short Physical Performance Battery (SPPB) is an excellent multi-component test instrument for assessing mobility and allows for separate and summation of ability in chair rise speed, walking speed over short distances and timed balance tests [5, 6, 28]. It is able to capture a broad range of function and has been used in institutional and community dwelling populations of older people. The SPPB has excellent validity with well characterised and strong associations with incident disability, hospitalisation and institutionalisation when applied to cohorts of older people [29-32]. It has evidence to support the selection of a minimal clinically important difference, which combined with good responsiveness to change means that it is a good measure to select for clinical trials [33]. The SPPB has used a novel methodology to circumvent a key challenge with timed measures, which is that an individual who cannot perform a task often cannot record a meaningful score. This means that data is lost from clinical trials which include frail and vulnerable people, and the pattern of data loss can lead to bias in many applications. By inverting the distribution of continuously scored data, identifying quartiles of the distribution, and then assigning an ordinal score to each quartile, all individuals can obtain a meaningful score. The cut-points used on the SPPB have been extensively validated in a range of data sets.

An advantage of the SPPB is that it can be undertaken in a range of environments with little impact on its reliability. The short walk distance can be accommodated in most homes, as evidenced by its use in large scale epidemiological research of ageing and has been used in care home settings. Standardised instructions about all elements are essential as even seemingly minor variants during the test elements, including slight assistance of arms during chair rise, chair height or selection of a walking aid, will impact the test results and the ability to compare between populations. The originators of the SPPB have produced standardised test

instructions, along with detailed manuals and examples of use, and these have been essential in promulgating the test in a reliable and comparable way across the world.

The main criticism of the SPPB would be in ceiling effects. For those who are already able, the measure will struggle to pick up further improvement and investigators should consider alternative measures like the fast walking tests, six minute walk test and shuttle walking tests [34-36]. These come in a variety of formats that combine incremental speed and/or distance.

Distance or the time spent walking are the two elements that can be standardised. Both measures have strong evidence of normative values, of sensitivity to change, a meaningful within and between group difference and a long history of use in interventional gerontology trials.

Walking aids have a significant impact on speed of walking, and tests require careful standardisation of procedures for using aids and assistance during walking tests. There are several options – selecting the test condition under which walking speed is maximal, or the test procedure under which the patients feels most confident, or to provide just one standard instruction. These issues need careful thought in situations where the trajectory of walking speed is likely to change substantially during short time periods – for example after hip fracture.

Both the six minute and shuttle walking tests provide a validated proxy for endurance and aerobic capacity, have established MCIDs and normative values to aid interpretation [34-36].

Although the SPPB is being used in an increasingly diverse range of clinical populations to good effect, there is no evidence that it can provide meaningful information across the entire life span and there are no large data sets to confirm normative values. Other multi-component measures are available, but generally have some degree of problem with managing people who

cannot complete some or all elements of the task and are not as well specified from an operational perspective.

The balance tests of the SPPB, NIH and Biomarker panels may also suffer from ceiling effects, and in situations where more healthy older populations are being measured, the recommendation of the Balance Consensus group to adopt the MINI-Best measurement should be considered [37].

Speed of walking (both usual speed and fast speed) are associated with frailty and general health in community, clinical and institutional settings [28, 33, 38-40]. A poorly recognised limitation of capacity tests is that they lose some precision as they are measurements grounded in the context of the day or hours that they are taken in. The American Geriatrics Society/British Geriatrics Society International Fall Guideline panel [41] systematically reviewed risk factors for falling and suggested that screening should comprise a question on fall history in the last year (any fall indicating risk), supplemented by a test of gait and balance in people reporting no fall history (with poor gait or balance indicating fall risk). However, at least one large cohort study has shown that asking test participants' to recall the frequency of feeling of loss of balance was substantially more accurate than a capacity test at a single time point [42].

Measures of muscle function.

There are a range of strength and muscle capacity tests available. Measures of maximal voluntary isometric (static) force are the most commonly reported. Lower limb power is more

closely related to capacity in many mobility tasks in older people than isometric force as it captures both the speed and amount of force that can be generated during muscle contraction [43]. However, power is less easily measured in everyday settings [43].

Maximal isometric strength can be measured effectively using either hand held or fixed dynamometers. Careful attention has to be paid to standardising the position of the participant and the tester as well as the environmental conditions for the test (for example the characteristics of the surface that the participant is positioned on). Verbal instructions should be standardised along with the number of repetitions and method of calculating the overall summary score of strength.

Both the NIH tool box and Biomarker consensus have recommended hand-grip strength as measured using a portable dynamometer as the method of choice, including guidance on the manufacturer and model. The ease of collecting hand grip strength has led to widespread adoption of this measure in large epidemiological studies [20, 21, 44], and in turn, much is known about the validity of hand grip strength. Normative data for hand grip is available across the life course [45]. Hand grip strength has demonstrated consistent strong associations with gait speed and balance and with activities of daily living [46]. It is recognised that hand grip strength one of the strongest predictors of mortality in middle-aged people [21]. The most likely explanation of these associations, is that the small muscles of the hand and the larger muscles of the forearm which are essential to generating grip force are good indicators of the reserve and functioning of the generalised skeletal muscle bulk of the body. Muscle is the largest organ in the body and has important actions not only in sustaining mobility and physical

activity essential to the health of other body systems, but acting as a metabolic buffer in various bioenergetics pathways such as insulin control and protein reserve.

Measures of dexterity

The NIH tool box defines dexterity as the ability to coordinate the fingers and manipulate objects in a timely manner [24]. The Rolyan 9-hole pegboard, single trial per hand, is recommended as the dexterity measure for inclusion in the NIH Toolbox. It demonstrates good reliability and validity and can be completed by people of all age ranges. In comparison to some other tests of dexterity, the Rolyan minimises visual perceptual demands. The Biomarker consensus panel concluded that the inclusion of measures of dexterity was one of the least evidence based areas of the recent recommendations for physical capacity measures. However, there is some evidence from older studies that dexterity and timed tests of hand performance can predict institutionalisation and other aspects of decline in health status with age [47].

Instruments to measure performance: Activities of Daily Living and Instrumental Activities of Daily living

ADL assessment can offer an insight into physical performance (i.e. what the person actually does in their usual environment). However, definitions and instruments measuring limitations in ADLs vary greatly. Over one hundred different instruments measure ADLs [48]. ADL measures are typically self-reported or require care providers and clinicians to make assessments, either from recall or from direct observation. The potential biases from such

assessments has been recognised and include lack of self-awareness particularly with cognitive impairment. More recent development of instruments has included more objectively assessed observed capacity to perform set tasks in peoples everyday settings [49]. For example, for the Performance ADL Test (PAT), people in their residential home are timed doing 16 progressively more challenging ADLs and have the quality of performance scored by an assessor [50]. Some of the most commonly used instruments measuring ADLs are the Katz Index of Activities of Daily Living (Katz ADL)[51], Functional Independence Measure (FIM)[52] and Barthel Index [53]. These instruments are generally better suited to measurement of people in institutionalised settings and in clinical populations with more severe functional limitations as the assessments focus on basic self-care and mobility tasks (e.g. bathing, transferring, toileting, dressing). Some instruments are more prone to ceiling effects and are less well suited for use on the community-dwelling older adult population. Instrumental activities of daily living (IADLs)[54] measure more complex daily tasks required to live independently (e.g. shopping, cleaning, managing finances and medication, preparing meals) and require greater cognitive and physical skills than basic ADLs. IADL instruments are better suited to assessing older adults in the community setting and in early or pre-clinical health states.

Physical performance is dependent on the task and the built and social environment. Accurate and individualised measurement of all parameters is most likely to help explain gaps between capacity and performance. Measurement and integration of data on environmental factors, particularly has been a challenge. The advent of wearable technology and a greater understanding of which environmental factors are most important to consider during measurement is substantially improving the assessment, characterisation and measurement of the fit between an individual's capacity and their unique environment [55].

Discussion

Measures of physical capacity have strong face validity and measure constructs important to older people and their families. Declining strength, mobility and stability are hallmarks of ageing, and variations in the onset of small decrements in mobility are associated with further decline and can be predictive of future health events [18]. In many situations a test of physical capacity provides a more accurate prediction of important future events such as premature mortality and major disability than tests commonly used in clinical settings. It is perhaps because of the ability to capture pre-clinical changes in function, and the reflection of multi-organ ability that this happens.

For the health care practitioner and/or researcher trying to select a measure, identifying the underlying construct and purpose is essential. For public health surveillance, a measure that has known associations with important end points such as quality of life, disability, hospitalisation, institutionalisation is an obvious choice. For a clinical trial, selecting a measure which is sensitive to within and between group change, and where there is knowledge of the meaning of the scale, is important. For use in clinical settings, the ability of a measure to convey individual meaning be that diagnosis, prognosis or monitoring response to treatment is essential.

Several texts hint at the possibility of using physical capacity tests in the way blood tests are used today [5, 6, 19]. This seems a strong possibility given the speed of development of the field, but will require much greater knowledge and evidence of predictive ability at the individual level. The key to harnessing this information will be in understanding the implication and refining clinical practice so that there are viable intervention strategies to follow.

There is remarkably little validation of measures to make predictions about individuals, although many instruments are able to identify and discriminate well between groups of people who are at increased risk. It is possible that current tests or test protocols are not sufficiently sensitive. Repeat testing or tests that seek recall of status over time may be needed to improve the precision of capacity tests of mobility.

A physical capacity test can tell you how well an individual can perform a set task, but it will not necessarily tell you why an individual is unable to perform a task. Physical capacity is nearly always attributable to multi-organ function. For example, the speed of walking under set conditions might be dependent on sight allowing assessment of direction and hazards, adequate balance functions (integration of vestibular, proprioceptive, visual and central nervous inputs), muscle power to enable the development of power around the joints, and joint movement allowing the joints to complete movements in the pre-requisite range. Walking over longer distances and at increased speeds requires cardiovascular reserve. The relative impairment of these functions will vary between individuals, as will the capacity of other systems to offset loss.

The fit between domain and test instrument are not absolute. The functions and tests selected express the final common pathway of phenotypic expression of many of the body systems. For

example, it would be difficult to complete a timed walking test in any of the following example situations – poor executive function, inability to see the test situation, poor muscle strength, lack of balance. There is evidence to suggest that different types of timed performance test are measuring the same underlying construct most likely non-chronological ageing. And also that a physical capacity test can predict cognitive decline at a range of 10 years [56] .

Physical tests of capacity have some cost associated with their collection, notably a test setting, basic equipment and personnel to conduct and run the test. However, the tests require substantially less investment than medical imaging for example. The main driver in determining how cost effective these tests may be will be in demonstrating that the use of the test enables more accurate targeting of health resources and that there are interventions that are effective in ameliorating the underlying conditions including frailty and cognitive impairment.

There are a range of sophisticated extensions to capacity and performance tests underway. Some of these are technology enabled, for example body worn sensors tests which allow assessment of the quality of walking by providing detailed information on all or selected parameters of the gait cycle [57]. Information from the gait cycle can be obtained from fixed or portable sensor mats, and increasingly from wearable sensor technology. Parameters include head sway, trunk sway (deviation), step length, step variation. All have been implicated as risk factors for falling, but whether they add significantly to either group stratification for falls (high versus low risk) or individual prediction has yet to be confirmed. A challenge is the association between increasing speed and the increase in a range of parameters such as step length and that these do not always increase in a linear way. Many investigators attempt to standardise test conditions using a metronome and multi-level modelling and standardising or accounting for

these sources of variation will need consideration particularly if physical capacity measurement evolves to free living assessment [58].

The other area of development is in understanding and measuring the built and social environment and integrating these elements into measures of performance of ADL and IADL. Although many cohorts have demonstrated the strong associations between capacity for example in the form of gait speed or muscle strength, accurate and meaningful ways of capturing and accounting for people's usual environment have been elusive. Better understanding of the environment/capacity gap should accelerate development to enable older people in broader society.

Summary

Measurement of physical capacity and performance is emerging as an important element of clinical intervention and research evaluation through the lifespan and in older people. Although there has been a disjointed development of concepts and models our paper underpins the importance of considering physical capacity and performance as separate but linked entities.

Practice points

- Selection of an instrument should be based on an assessment of the domains of interest and measurement properties (including reliability, validity, responsiveness, predictive and diagnostic accuracy, practicality, and feasibility).

- Tests of physical capacity are conducted under standardised environmental conditions, with careful control and consideration of aids, appliances and additional support.
- Physical capacity tests with evidence of adequate psychometric properties for the evaluation of capacity in response to intervention programmes include the Short Physical Performance Battery, hand grip strength, mini-BEST and 9-hole dexterity test.
- Performance measurement usually focuses on basic or instrumental (advanced) Activities of Daily Living (ADLs) completed by people in their usual environments, the latter better suited to assessing older adults in the community setting and in early or pre-clinical health states.

Research agenda

- There is remarkably little validation of measures to make predictions about health events and outcomes of individuals, although many instruments are able to identify and discriminate well between groups of people who are at increased risk. It is possible that current tests or test protocols are not sufficiently sensitive.
- An important area for development is the understanding and measurement of the built and social environment and integrating these elements into measures of performance of ADL and IADL.
- Technology enabled measurement promises to enhance research and clinical practice, for example body worn sensors tests which allow assessment of the quality of walking by providing detailed information on all or selected parameters of the gait cycle.

Conflict of Interest Statement

None of the authors report a conflict of interest.

Acknowledgement

We thanks Mrs Sue Davolls for editing and assisting with manuscript preparation

Funding

The research was supported by the National Institute for Health Research (NIHR)

Collaboration for Leadership in Applied Health Research and Care Oxford at Oxford Health

NHS Foundation Trust, and supported by the NIHR Biomedical Research Unit, Oxford. Dr

D. Keene is supported by the NIHR Biomedical Research Centre, Oxford and the NIHR

Post-Doctoral Fellowship programme (PDF- 2016-09-056). The views expressed are those of

the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

References (* 10 most important)

- [1] World Health Organization. International Classification of Functioning, Disability and Health: ICF: World Health Organization; 2001.
- [2] World Health Organization. World report on ageing and health: World Health Organization; 2015.
- [3] Nagi SZ. An epidemiology of disability among adults in the United States. The Milbank Mem Fund Q Health Soc. 1976;54:439-67.
- [4] Pope A, Tarlov A. Disability in America: Toward a national agenda for prevention. Washington, DC: National Academy Press; 1991.
- [5] Guralnik JM, Branch LG, Cummings SR, Curb JD. Physical performance measures in aging research. J Gerontol. 1989;44(5) :M141-M6.
- [6] Guralnik JM, Ferrucci L. Assessing the building blocks of function: utilizing measures of functional limitation. Am J Prev Med. 2003;25:112-21.
- [7] Scientific Advisory Committee of the Medical outcomes Trust. Instrument Review Criteria. Med Outcomes Trust Bull. 1995;I-IV.
- [8] Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating patient-based outcome measures for use in clinical trials. 1998.
- [9] Nunnally J, Bernstein JC. Psychometric Theory, 3rd Ed. New York: McGraw-Hill, 1994.
- [10] Fleiss JL. Statistical Methods for Rates and Proportions, 2nd Ed. Chichester: Wiley, 1981.
- [11] Stiell IG, McKnight RD, Greenberg GH *et al*. Interobserver agreement in the examination of acute ankle injury patients. Am J Emerg Med 1992;10: 14-17.
- [12] Streiner DL, Norman GR. Health Measurement Scales: A Practical Guide to Their Development and Use, 2nd Ed. Oxford: Oxford University Press, 1995.

- [13] Cohen J. Statistical Power Analysis for the Behavioural Sciences. New York: Academic Press, 1977.
- [14] Deeks JJ. Systematic reviews of evaluations of diagnostic and screening tests. *BMJ*. 2001;323(7305):157-62.
- [15] Deeks JJ. Using evaluations of diagnostic tests: understanding their limitations and making the most of available evidence. *Ann Oncol*. 1999;10:761-8.
- [16] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*: 2015;350.
- [17] Thangaratnam S, Allotey J, Marlin N, Mol BW, Von Dadelszen P, Ganzevoort W, et al. Development and validation of Prediction models for Risks of complications in Early-onset Pre-eclampsia (PREP): a prospective cohort study. *Health Technol Assess*. 2017;21:1-100.
- *[18] Kuh D, Karunananthan S, Bergman H, Cooper R. A life-course approach to healthy ageing: maintaining physical capability. *Proc Nutr Soc*. 2014;73:237-48.
- [19] Chaudhry SI, McAvay G, Ning Y, Allore HG, Newman AB, Gill TM. Geriatric impairments and disability: the cardiovascular health study. *J AM Geriatr Soc*. 2010;58:1686-92.
- *[20] Cooper R, Kuh D, Cooper C, Gale CR, Lawlor DA, Matthews F, et al. Objective measures of physical capability and subsequent health: a systematic review. *Age Ageing*. 2010;40:14-23.
- *[21] Cooper R, Kuh D, Hardy R. Objectively measured physical capability levels and mortality: systematic review and meta-analysis. *BMJ*. 2010;341:c4467.
- *[22] Sibley KM, Howe T, Lamb SE, Lord SR, Maki BE, Rose DJ, et al. Recommendations for a core outcome set for measuring standing balance in adult populations: a consensus-based approach. *PloS One*. 2015;10:e0120568.

- *[23] Lamb SE, Jorstad-Stein EC, Hauer K, Becker C. Development of a common outcome data set for fall injury prevention trials: the Prevention of Falls Network Europe consensus. *J Am Geriatr Soc*. 2005;53:1618-22.
- *[24] Reuben DB, Magasi S, McCreath HE, Bohannon RW, Wang Y-C, Bubela DJ, et al. Motor assessment using the NIH Toolbox. *Neurology*. 2013;80:S65-S75.
- *[25] Lara J, Cooper R, Nissan J, Ginty AT, Khaw K-T, Deary IJ, et al. A proposed panel of biomarkers of healthy ageing. *BMC Med*. 2015;13:222.
- *[26] Morley JE, Abbatecola AM, Argiles JM, Baracos V, Bauer J, Bhasin S, et al. Sarcopenia with limited mobility: an international consensus. *J Am Med Dir Assoc*. 2011;12:403-9.
- [27] Dobson F, Hinman R, Roos EM, Abbott J, Stratford P, Davis A, et al. OARSI recommended performance-based tests to assess physical function in people diagnosed with hip or knee osteoarthritis. *Osteoarthritis Cartilage*. 2013;21:1042-52.
- *[28] Chung J, Demiris G, Thompson HJ. Instruments to assess mobility limitation in community-dwelling older adults: a systematic review. *J Aging Phys Act*. 2015;23:298-313.
- [29] Freire AN, Guerra RO, Alvarado B, Guralnik JM, Zunzunegui MV. Validity and reliability of the short physical performance battery in two diverse older adult populations in Quebec and Brazil. *J Aging Health*. 2012;24:863-78.
- [30] Volpato S, Cavalieri M, Sioulis F, Guerra G, Maraldi C, Zuliani G, Fellin R, Guralnik JM. Predictive value of the Short Physical Performance Battery following hospitalization in older patients. *J Gerontol A Biol Sci Med Sci*. 2011;66:89-96.
- [31] Guralnik JM, Ferrucci L, Pieper CF, Leveille SG, Markides KS, Ostir GV, et al. Lower extremity function and subsequent disability: consistency across studies, predictive models, and value of gait speed alone compared with the short physical performance battery. *J Gerontol A Biol Sci Med Sci* 2000;Apr 55:M221-31.
- [32] Guralnik JM, Simonsick FM, Ferrucci L, Glynn RJ, Berkman LF, Blazer DG, et al.

A short physical performance battery assessing lower extremity function: association with self-reported disability and prediction of mortality and nursing home admission. *J Gerontology*. 1994;Mar;49(2):M85-94.

[33] Perera S, Studenski S, Newman A, Simonsick E, Harris T, Schwartz A, et al. Are estimates of meaningful decline in mobility performance consistent among clinically important subgroups? (Health ABC Study). *J Gerontol A Biol Sci Med Sci*. 2014;69:1260-8.

[34] Bohannon RW, Bubela D, Magasi S, McCreath H, Wang Y-C, Reuben D, et al. Comparison of walking performance over the first 2 minutes and the full 6 minutes of the Six-Minute Walk Test. *BMC Res Notes*. 2014;7:269.

*[35] Bohannon RW, Glenney SS. Minimal clinically important difference for change in comfortable gait speed of adults with pathology: a systematic review. *J Eval Clin Pract*. 2014;20:295-300.

[36] Bohannon RW, Wolfson LI, White WB. Timed mobility: description of measurement, performance, and dimensionality among older adults. *Disabil Rehabil*. 2017:1-4.

[37] Franchignoni F, Horak F, Godi M, Nardone A, Giordano A. Using psychometric techniques to improve the Balance Evaluation Systems Test: the mini-BESTest. *J Rehabil Med*. 2010;42:323-31.

[38] Atkinson HH, Rosano C, Simonsick EM, Williamson JD, Davis C, Ambrosius WT, et al. Cognitive function, gait speed decline, and comorbidities: the health, aging and body composition study. *J Gerontol A Biol Sci Med Sci*. 2007;62:844-50.

[39] Kuys SS, Peel NM, Klein K, Slater A, Hubbard RE. Gait speed in ambulant older people in long term care: a systematic review and meta-analysis. *J Am Med Dir Assoc*. 2014;15(3):194-200.

[40] Peel NM, Kuys SS, Klein K. Gait speed as a measure in geriatric assessment in clinical settings: a systematic review. *J Gerontol A Biol Sci Med* 2013;68(1):39-46.

- [41] Kenny RA, Rubenstein LZ, Tinetti ME, Brewer K, Cameron KA, Capezuti EA et al. Summary of the updated American Geriatrics Society/British Geriatrics Society clinical practice guideline for prevention of falls in older persons. *J Am Geriatr Soc.* 2011;59:148-57.
- [42] Lamb SE, McCabe C, Becker C, Fried LP, Guralnik JM. The optimal sequence and selection of screening test items to predict fall risk in older disabled women: the Women's Health and Aging Study. *J Gerontol A Biol Sci Med Sci.* 2008;63(10):1082-8.
- [43] Byrne C, Faure C, Keene DJ, Lamb SE. Ageing, Muscle Power and Physical Function: A Systematic Review and Implications for Pragmatic Training Interventions. *Sports medicine (Auckland, NZ).* 2016;46:1311-32.
- [44] Xue QL, Beamer BA, Chaves PH, Guralnik JM, Fried LP. Heterogeneity in rate of decline in grip, hip, and knee strength and the risk of all-cause mortality: the women's health and aging study II. *J Am Geriatr Soc.* 2010;58(11):2076-84.
- [45] Dodds RM, Syddall HE, Cooper R, Kuh D, Cooper C, Sayer AA. Global variation in grip strength: a systematic review and meta-analysis of normative data. *Age Ageing.* 2016;45(2):209-16.
- [46] Vermeulen J, Neyens JC, van Rossum E, Spreeuwenberg MD, de Witte LP. Predicting ADL disability in community-dwelling elderly people using physical frailty indicators: a systematic review. *BMC Geriatr.* 2011;11:33.
- [47] Falconer J, Hughes SL, Naughton BJ, Singer R, Chang RW, Sinacore JM. Self report and performance-based hand function tests as correlates of dependency in the elderly. *J Am Geriatr Soc.* 1991;39:695-9.
- [48] McDowell I. *Measuring health : a guide to rating scales and questionnaires.* 3rd ed. New York ; Oxford: Oxford University Press; 2006.
- [49] Mlinac ME, Feng MC. Assessment of Activities of Daily Living, Self-Care, and Independence. *Arch Clin Neuropsychol.* 2016;31(6):506-16.

- [50] Weening-Dijksterhuis E, Kamsma YP, van Heuvelen MJ. Psychometric properties of the PAT: an assessment tool for ADL performance of older people living in residential homes. *Gerontology*. 2011;57(5):405-13.
- [51] Katz S, Ford AB, Moskowitz RW, Jackson BA, Jaffe MW. Studies of illness in the aged. The index of ADL: a standardized measure of biological and psychosocial function. *JAMA*. 1963;185:914-9.
- [52] Keith RA, Granger CV, Hamilton BB, Sherwin FS. The functional independence measure: a new tool for rehabilitation. *Adv Clin Rehabil*. 1987;1:6-18.
- [53] Mahoney FI, Barthel DW. Functional evaluation: The Barthel Index. *Md State Med J*. 1965;14:61-5.
- [54] Lawton MP, Brody EM. Assessment of older people: self-maintaining and instrumental activities of daily living. *The Gerontologist*. 1969;9(3):179-86.
- [55] Van Holle V, Van Cauwenberg J, Gheysen F, Van Dyck D, Deforche B, Van de Weghe N, et al. The Association between Belgian Older Adults' Physical Functioning and Physical Activity: What Is the Moderating Role of the Physical Environment? *PloS One*. 2016;11:e0148398.
- [56] Steves CJ, Mehta MM, Jackson SH, Spector TD. Kicking back cognitive ageing: leg power predicts cognitive ageing after ten years in older female twins. *Gerontology*. 2016;62:138-49.
- [57] van Schooten KS, Pijnappels M, Rispens SM, Elders PJ, Lips P, Daffertshofer A, et al. Daily-life gait quality as predictor of falls in older people: a 1-year prospective cohort study. *PloS One*. 2016;Jul 7;11(7):e0158623.
- [58] Keene DJ, Moe-Nilssen R, Lamb SE. The application of multilevel modelling to account for the influence of walking speed in gait analysis. *Gait & posture*. 2016;43:216-9.

