Table 1: Reliability and reliability of outcome measure variables to be implemented in the EXVITD study

Outcome variable	Measurement variable	Reliability/validity/appropriateness in an older population
Muscle power	LLEP (Nottingham power rig)	The Nottingham power rig is used to measure explosive power in the dominant leg of participants. The test is completed in a seated position, with the participant extending their dominant leg in order to depress the foot press as quickly and forcefully as possible. The depression of the foot press rotates the flywheel and an optoswitch is used to record angular velocity and calculate LEP[1]. This action incorporates hip extension, knee extension, and ankle plantar flexion and is considered the "gold standard" of power measurement in older adults[2]. Since the test is performed seated, it is considered a safe alternative measure of muscle power to jump tests in frail older adults[3]. The measurement of muscle power using the Nottingham leg rig has been found to be significantly associated with measurement of power using an isokinetic dynamometer (Spearman's rho = 0.73, p<0.001) and two-legged jumps on a force plate (Spearman's rho = 0.86, p<0.001)[1]. The test-retest coefficient of variation has been reported as 8% in a population of 419 women aged 63-75 years[4]. Muscle power was selected as the primary outcome measure since muscle power is a superior predictor of functional status than muscle strength[5,6] and power of the lower limbs has been found to be an independent predictor of self-reported functional status in community-dwelling older women[7]. Muscle power declines more rapidly with ageing than muscle strength [8] and low leg muscle power is associated with a 2-3 fold greater risk of mobility limitations compared with low muscle strength [9]; mobility limitation is an independent risk factor for disability, hospitalization and mortality and reduces quality of life and independence in older adults[10].
Body composition	DXA	Total body composition and BMD was assessed using a GE Lunar iDXA encore 2011 running software version 13.60.033. Quality assessment checks are completed 3 times per week with the machine switched on and water and encapsulated phantoms provided by the manufacturer are used to calibrate the machine 3 times per week. DXA differentiates and quantifies different materials in the body using 2 beam energies which are attenuated or absorbed in relation to tissue type and amount. DXA discriminates between lean and bone materials, and provides estimates of 3 body compartments; lean, bone and fat. Since lean and fat masses comprise more water than bone, these tissues will attenuate the beam energy to a lesser degree than bone[11]. DXA is widely available, produces lower radiation doses and measurements of muscle mass and quality are highly correlated with both CT (multi-slice thigh fat-free mass, r^2 = 0.96)[12] and MRI (whole body lean mass r = 0.94)[13]. Although DXA does have a number of limitations, such as the inability to measure intramuscular fat and differences in results between devices[11], it has been cited as a reliable method of indirectly estimating muscle mass in older adults[13]. Indeed, DXA is the current "reference technique of choice for estimating muscle mass and body composition in research and clinical practice", and is widely used in RCTs to estimate skeletal muscle mass[11].

BMD	DXA	BMD is calculated by converting the radiation energy per pixel into areal density, which is the number of pixels in the area and the amount of bone in each pixel[14]. The assessment of BMD to predict fracture risk has limitations, namely, BMD does not fully describe bone strength nor quality, does not account for bone size nor architecture, is not a true three dimensional measure and cross-comparison between different DXA machines is difficult[15]. However, from BMD, T and Z scores can be calculated, which are used to diagnose osteoporosis[16]; BMD estimation via DXA has become "universally adopted as a standard to define osteoporosis"[17]. In the EXVITD study, additional images of the hip and spine were chosen since these site-specific scans are preferred to diagnose/confirm osteoporosis and predict fracture risk[18]. DXA used to estimate BMD has been described as highly precise (with a maximum acceptable precision error of 2-2.5%[17]) and BMD of the hip is a strong predictor of hip fracture risk in men and women[19].
Muscle function	SPPB	The Short Physical Performance Battery test (SPPB) was designed to assess lower extremity function in older adults and consists of 3 domains; balance, gait speed and chair rise tests[20]. The test has a maximum score of 12 points, with a score ≤8 points indicative of poor physical performance[21]. SPP score is associated with mobility disability, hospitalization and mortality[22,23]. The SPPB test has been found to be a highly valid and reliable measure of function in a variety of older populations[24-26] and also sensitive to change in the event of medical events including myocardial infarction, stroke and hip fracture; participants experiencing such medical events were significantly more likely to record poorer summary performance change scores[27]. Therefore, the SPPB test is a standard measure of physical performance in older adults in a research and clinical setting[21] and is also a validated test for sarcopenia severity diagnosis[28].
	TUG	The Timed Up and Go test (TUG) is a measure of muscle function developed specifically for older adults and is the time (in seconds) taken to stand up from a chair, walk 3 meters, turn around, return and sit down[29]. Aspects of the TUG test are characteristic of activities of daily living important for the maintenance of physical independence, i.e. the muscle strength to rise from a chair, gait speed, turning 180°, and the co-ordination to sit safely from a standing position[30] and poor performance in the TUG test has been associated with mortality (HR = 1.79; 95% CI, 1.33, 2.42, p <0.001)[31]. The TUG test has previously been found to be a sensitive and specific measure for identifying older people with sarcopenia (sensitivity = 67%, specificity = 88.7%)[32] and those who are prone to falls (sensitivity = 87%, specificity = 87%)[33]. Additionally, a high inter-rater reliability has been reported amongst community-dwelling older adults (ICC = 0.98)[33] and the TUG test is recommended by both the American and British Geriatrics societies to assess gait speed, balance and fall risk in older adults[34]. The TUG test is recommended by The European Working Group on Sarcopenia in Older People to assess low physical performance[28] and a cut-off point of ≥20 seconds for low performance is suggested[35].

Physical activity	Accelerometery (ActivPAL^{TM})	Accelerometry was used to estimate average physical activity over 7 days. The activPAL^{TM} monitor (PAL Technologies Ltd, Glasgow, Scotland) is a small accelerometer device affixed to the anterior midline of the thigh where it is able to determine static and dynamic acceleration, classify activities into 3 categories (sitting/lying, standing, stepping) and provides an estimation of energy expenditure (EE) expressed in metabolic equivalents (METs). Although subjective measures of activity estimation such as self-report questionnaires and activity diaries are more practical for large-scale cohorts[36], they may overestimate sedentary time in older adults[37], they reply on participant recall and they can place a high level of burden on the participant[38]. Conversely, activPAL monitoring has reported excellent inter-device reliability (ICC = 0.99) and an overall 95.9% agreement between a second-by-second observed analysis and the monitor[39]. Additionally, activPAL has been reported to record step count and cadence under 3 treadmill conditions (fast, normal and slow speeds) and during a 500 metre outdoor walk to a high degree of accuracy (<1% absolute percentage error) in a group of 20 older adults of mean age 71.9 years[40].
Musculoskeletal pain	VAS pain	The visual analogue scale (VAS) for pain is a generic questionnaire for evaluating pain. It consists of a coloured scale from 0 to 10, with 0 being “no pain” and 10 being “worst imaginable pain”, with a score <4 indicative of good pain management[41] and a score of ≤2 a cut-off point for “acceptable” pain[42]. The minimum change of clinical relevance in a cohort of patients with rheumatoid arthritis was shown to be 1.1 points on the 10 point scale, which was the calculated standard error of the mean[42]. The questionnaire is completed by the participant, with a line drawn perpendicular to their experience of pain both in a seated and standing position for the previous 24 hours. The participant burden of completing the questionnaire is low, since it can be completed in under 1 minute and the VAS is reported to be sensitive to change in clinical trials[41]. Test-retest reliability has been reported as r = 0.937 for literate and r = 0.712 for illiterate patients with rheumatoid arthritis[43]. Chronic pain assessed by the VAS has been shown to be significantly associated with poor self-reported health status, sleep disorders, depression and malnutrition in 105 nursing home residents of mean age 82.2 years[44].
Quality of life	SF-36	The general population health questionnaire the Short Form (SF)-36 is a quick and comprehensive version of the 149-item health questionnaire developed as part of the Medical Outcomes study[45]. The SF-36 questions encompass 3 domains of health; functional status, wellbeing and an overall evaluation of health. A 2005 review of quality of life assessment in older adults found that the SF-36 was the most extensively used assessment, with good evidence supporting the questionnaire reliability, validity and responsiveness to change, particularly in community-dwelling older adults with lower levels of morbidity[46]. The SF-36 was completed quickly (median time of 8 minutes) by 195 patients aged 65 years and over and the authors remarked questions regarding functional ability made the SF-36 particularly relevant for use in older adults[47].

Gound reaction force & peak mechanical power of the lower limbs	Leonardo Mechanograph® Ground Reaction Force Plate	The Leonardo Mechanograph® Ground Reaction Force Plate (Novotec Medical, Pforzheim, Germany) running the Leonardo Mechanograph® GRFP Research Edition Software version 4.2 was used to assess ground reaction force and peak mechanical power of the lower limbs. The Leonardo jump plates are calibrated at set up with no ongoing checks. The Leonardo Mechanograph® Ground Reaction Force Plate comprises a bench (45cm high) and a force platform separated into 2 equal halves (left and right), each containing 4 strain gauge force sensors used to measure vertical ground reaction force[48]. The equipment and associated software records the maximal total relative power per body weight ($CRTP_{rel}$) and the maximal velocity (CRTv) during standing throughout the 5 time chair rise test (also performed as part of the SPPB test[20]). The chair rise test was chosen, rather than the single two-legged jump as this test may be contraindicated in frail older adults, the chair rise test evaluates an action which is relevant in everyday life[49] by assessing the muscle power required to coordinate the multiple-joint movement of standing from a seated position[50]. Additionally, one study found that lower limb power measured using jump mechanography was more highly correlated with age than the SPPB, grip strength and lean mass assessed by DXA[51], suggesting that lower limb power may be an important
		measure to detect early deficits associated with sarcopenia[52]. The Leonardo Mechanograph® Ground Reaction Force Plate has been found to have excellent test-retest reliability (ICC of up to 0.99) in a range of different age groups, including older adults[49,53,54], good inter and intra-rater reliability[48] and measurements of lower limb power correlate well with the Nottingham power rig (r = 0.6) and isokintetic dynamometry (r = 0.68)[3].
Musculoskeletal pain	VAS Pain	The visual analogue scale (VAS) for pain is a generic questionnaire for evaluating pain. It consists of a coloured scale from 0 to 10, with 0 being “no pain” and 10 being “worst imaginable pain”, with a score <4 indicative of good pain management (Burckhardt and Jones, 2003) and a score of ≤2 a cut-off point for “acceptable” pain (Wolfe and Michaud, 2007). The minimum change of clinical relevance in a cohort of patients with rheumatoid arthritis was shown to be 1.1 points on the 10 point scale, which was the calculated standard error of the mean (Wolfe and Michaud, 2007). The questionnaire is completed by the participant, with a line drawn perpendicular to their experience of pain both in a seated and standing position for the previous 24 hours. The participant burden of completing the questionnaire is low, since it can be completed in under 1 minute and the VAS is reported to be sensitive to change in clinical trials (Burckhardt and Jones, 2003). Test-retest reliability has been reported as r = 0.937 for literate and r = 0.712 for illiterate patients with rheumatoid arthritis (Ferraz et al., 1990). Chronic pain assessed by the VAS has been shown to be significantly associated with poor self-reported health status, sleep disorders, depression and malnutrition in 105 nursing home residents of mean age 82.2 years (Zanocchi et al., 2008).

LLEP: Lower limb extensor power; LEP: Leg extensor power; DXA: Dual-energy X-ray absorptiometry; BMD: Bone mineral density; CT: Computerised tomography; MRI: Magnetic resonance imaging; RCT: Randomized controlled trial; HR: Hazard ration; CI: Confidence interval; ICC: Intraclass correlation

References

1. Bassey E, Short A. A new method for measuring power output in a single leg extension: feasibility, reliability and validity. *European journal of applied physiology and occupational physiology* 1990;60(5):385.
2. Gray M, Paulson S. Developing a measure of muscular power during a functional task for older adults. *BMC geriatrics* 2014;14(1):145.
3. Lindemann U, Claus H, Stuber M, et al. Measuring power during the sit-to-stand transfer. *European journal of applied physiology* 2003;89(5):466-70.
4. Portegijs E, Sipilä S, Alen M, et al. Leg extension power asymmetry and mobility limitation in healthy older women. *Archives of physical medicine and rehabilitation* 2005;86(9):1838-42.
5. Suzuki T, Bean JF, Fielding RA. Muscle power of the ankle flexors predicts functional performance in community-dwelling older women. *Journal of the American Geriatrics Society* 2001;49(9):1161-67.
6. Byrne C, Faure C, Keene DJ, et al. Ageing, muscle power and physical function: a systematic review and implications for pragmatic training interventions. *Sports Medicine* 2016;46(9):1311-32.
7. Foldvari M, Clark M, Laviolette LC, et al. Association of Muscle Power With Functional Status in Community-Dwelling Elderly Women. *The Journals of Gerontology: Series A* 2000;55(4):M192-M99. doi: 10.1093/gerona/55.4.M192
8. Skelton DA, GREIG CA, Davies JM, et al. Strength, power and related functional ability of healthy people aged 65–89 years. *Age and ageing* 1994;23(5):371-77.
9. Bean JF, Leveille SG, Kiely DK, et al. A Comparison of Leg Power and Leg Strength Within the InCHIANTI Study: Which Influences Mobility More? *The Journals of Gerontology: Series A* 2003;58(8):M728-M33. doi: 10.1093/gerona/58.8.M728
10. Pahor M, Guralnik JM, Ambrosius WT, et al. Effect of structured physical activity on prevention of major mobility disability in older adults: the LIFE study randomized clinical trial. *Jama* 2014;311(23):2387-96.
11. Buckinx F, Landi F, Cesari M, et al. Pitfalls in the measurement of muscle mass: a need for a reference standard. *Journal of Cachexia, Sarcopenia and Muscle* 2018;9(2):269-78. doi: 10.1002/jcsm.12268
12. Levine JA, Abboud L, Barry M, et al. Measuring leg muscle and fat mass in humans: comparison of CT and dual-energy X-ray absorptiometry. *Journal of Applied Physiology* 2000;88(2):452-56. doi: 10.1152/jappl.2000.88.2.452
13. Chen Z, Wang Z, Lohman T, et al. Dual-Energy X-Ray Absorptiometry Is a Valid Tool for Assessing Skeletal Muscle Mass in Older Women. *The Journal of Nutrition* 2007;137(12):2775-80. doi: 10.1093/jn/137.12.2775
14. Berger A. 0*BMJ: British Medical Journal* 2002;325(7362):484.
15. Pors Nielsen S. The Fallacy of BMD: A Critical Review of the Diagnostic Use of Dual X-ray Absorptiometry. *Clinical Rheumatology* 2000;19(3):174-83. doi: 10.1007/s100670050151
16. World Health Organization. Assessment of fracture risk and its application to screening for postmenopausal osteoporosis: report of a WHO study group [meeting held in Rome from 22 to 25 June 1992] 1994 [Available from:

https://apps.who.int/iris/bitstream/handle/10665/39142/WHO_TRS_843_eng.pdf?sequence=1&isAllowed=y accessed 13/03/19 2019.
17. Link TM. Osteoporosis imaging: state of the art and advanced imaging. *Radiology* 2012;263(1):3-17.
18. Sözen T, Özışık L, Başaran NÇ. An overview and management of osteoporosis. *European journal of rheumatology* 2017;4(1):46.
19. Johnell O, Kanis JA, Oden A, et al. Predictive value of BMD for hip and other fractures. *Journal of bone and mineral research* 2005;20(7):1185-94.
20. Guralnik JM, Simonsick EM, Ferrucci L, 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.
21. Cruz-Jentoft AJ, Baeyens JP, Bauer JM, et al. Sarcopenia: European consensus on definition and diagnosis: Report of the European Working Group on Sarcopenia in Older PeopleA. J. Cruz-Gentoft et al. *Age and Ageing* 2010;39(4):412-23. doi: 10.1093/ageing/afq034
22. Guralnik JM, Ferrucci L, Pieper CF, 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. *The Journals of Gerontology: Series A* 2000;55(4):M221-M31. doi: 10.1093/gerona/55.4.M221
23. Beaudart C, McCloskey E, Bruyère O, et al. Sarcopenia in daily practice: assessment and management. *BMC geriatrics* 2016;16(1):170.
24. Freire AN, Guerra RO, Alvarado B, et al. Validity and reliability of the short physical performance battery in two diverse older adult populations in Quebec and Brazil. *Journal of aging and health* 2012;24(5):863-78.
25. Gómez JF, Curcio C-L, Alvarado B, et al. Validity and reliability of the Short Physical Performance Battery (SPPB): a pilot study on mobility in the Colombian Andes. *Colombia medica* 2013;44(3):165-71.
26. Olsen CF, Bergland A. Reliability of the Norwegian version of the short physical performance battery in older people with and without dementia. *BMC geriatrics* 2017;17(1):124.
27. Ostir GV, Volpato S, Fried LP, et al. Reliability and sensitivity to change assessed for a summary measure of lower body function: results from the Women's Health and Aging Study. *Journal of clinical epidemiology* 2002;55(9):916-21.
28. Writing Group for the European Working Group on Sarcopenia in Older People 2 , EWGSOP2 tEGf, Cruz-Jentoft AJ, et al. Sarcopenia: revised European consensus on definition and diagnosis. *Age and Ageing* 2018;48(1):16-31. doi: 10.1093/ageing/afy169
29. 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):142-48.
30. Vervoort D, Vuillerme N, Kosse N, et al. Multivariate analyses and classification of inertial sensor data to identify aging effects on the Timed-Up-and-Go test. *PloS one* 2016;11(6):e0155984.
31. Bergland A, Jørgensen L, Emaus N, et al. Mobility as a predictor of all-cause mortality in older men and women: 11.8 year follow-up in the Tromsø study. *BMC health services research* 2017;17(1):22.
32. Martinez BP, Gomes IB, Oliveira CSd, et al. Accuracy of the Timed Up and Go test for predicting sarcopenia in elderly hospitalized patients. *Clinics* 2015;70(5):369-72.
33. Shumway-Cook A, Brauer S, Woollacott M. Predicting the Probability for Falls in Community-Dwelling Older Adults Using the Timed Up & Go Test. *Physical Therapy* 2000;80(9):896-903. doi: 10.1093/ptj/80.9.896
34. Panel on Prevention of Falls in Older Persons AGS, Society BG. Summary of the updated American Geriatrics Society/British Geriatrics Society clinical practice guideline for prevention of falls in older persons. *Journal of the American Geriatrics Society* 2011;59(1):148-57.

35. Bischoff HA, Stähelin HB, Monsch AU, et al. Identifying a cut-off point for normal mobility: a comparison of the timed 'up and go'test in communitydwelling and institutionalised elderly women. *Age and ageing* 2003;32(3):315-20.
36. Aguilar-Farías N, Brown WJ, Olds TS, et al. Validity of self-report methods for measuring sedentary behaviour in older adults. *Journal of science and medicine in sport* 2015;18(6):662-66.
37. Van Cauwenberg J, Van Holle V, De Bourdeaudhuij I, et al. Older adults' reporting of specific sedentary behaviors: validity and reliability. *BMC Public Health* 2014;14(1):734.
38. Atkin AJ, Gorely T, Clemes SA, et al. Methods of Measurement in epidemiology: Sedentary Behaviour. *International Journal of Epidemiology* 2012;41(5):1460-71. doi: 10.1093/ije/dys118
39. Grant PM, Ryan CG, Tigbe WW, et al. The validation of a novel activity monitor in the measurement of posture and motion during everyday activities. *British journal of sports medicine* 2006;40(12):992-97.
40. Grant PM, Dall PM, Mitchell SL, et al. Activity-monitor accuracy in measuring step number and cadence in community-dwelling older adults. *Journal of aging and physical activity* 2008;16(2):201-14.
41. Burckhardt CS, Jones KD. Adult measures of pain: the McGill pain questionnaire (MPQ), rheumatoid arthritis pain scale (RAPS), short-form mcgill pain questionnaire (SF-MPQ), verbal descriptive scale (VDS), visual analog scale (VAS), and west haven-yale multidisciplinary pain inventory (WHYMPI). *Arthritis Care & Research: Official Journal of the American College of Rheumatology* 2003;49(S5):S96-S104.
42. Wolfe F, Michaud K. Assessment of pain in rheumatoid arthritis: minimal clinically significant difference, predictors, and the effect of anti-tumor necrosis factor therapy. *The Journal of rheumatology* 2007;34(8):1674-83.
43. Ferraz MB, Quaresma M, Aquino L, et al. Reliability of pain scales in the assessment of literate and illiterate patients with rheumatoid arthritis. *The Journal of rheumatology* 1990;17(8):1022-24.
44. Zanocchi M, Maero B, Nicola E, et al. Chronic pain in a sample of nursing home residents: prevalence, characteristics, influence on quality of life (QoL). *Archives of gerontology and geriatrics* 2008;47(1):121-28.
45. Ware Jr JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36): I. Conceptual framework and item selection. *Medical care* 1992:473-83.
46. Haywood K, Garratt A, Fitzpatrick R. Quality of life in older people: a structured review of generic self-assessed health instruments. *Quality of life Research* 2005;14(7):1651-68.
47. Hayes V, Morris J, Wolfe C, et al. The SF-36 health survey questionnaire: is it suitable for use with older adults? *Age and ageing* 1995;24(2):120-25.
48. Matheson L, Duffy S, Maroof A, et al. Intra-and inter-rater reliability of jumping mechanography muscle function assessments. *J Musculoskelet Neuronal Interact* 2013;13(4):480-6.
49. Buehring B, Krueger D, Fidler E, et al. Reproducibility of jumping mechanography and traditional measures of physical and muscle function in older adults. *Osteoporosis International* 2015;26(2):819-25.
50. Dietzel R, Felsenberg D, Armbrecht G. Mechanography performance tests and their association with sarcopenia, falls and impairment in the activities of daily living–a pilot cross-sectional study in 293 older adults. *Journal of musculoskeletal & neuronal interactions* 2015;15(3):249.
51. Siglinsky E, Krueger D, Ward RE, et al. Effect of age and sex on jumping mechanography and other measures of muscle mass and function. *Journal of musculoskeletal & neuronal interactions* 2015;15(4):301.

52. Hannam K, Hartley A, Clark E, et al. Feasibility and acceptability of using jumping mechanography to detect early components of sarcopenia in community-dwelling older women. *Journal of musculoskeletal & neuronal interactions* 2017;17(3):246.
53. Rittweger J, Schiessl H, Felsenberg D, et al. Reproducibility of the jumping mechanography as a test of mechanical power output in physically competent adult and elderly subjects. *Journal of the American Geriatrics Society* 2004;52(1):128-31.
54. Veilleux L-N, Rauch F. Reproducibility of jumping mechanography in healthy children and adults. *Journal of Musculoskeletal and Neuronal Interactions* 2010;10(4):256-66.