UiT / NORGES ARKTISKE UNIVERSITET

Inter-rater reliability of the Norwegian translation of the Wolf Motor Function Test

Martin Vatshaug

Mastergradsoppgave i helsefag, flerfaglig studieretning

Institutt for Helse og Omsorgsfag

Det Helsevitenskapelige fakultet

Universitetet i Tromsø

Oktober 2014

Foreword:

This thesis was written for my master degree in health at the University in Tromsø, The Arctic University of Norway, and is part of the NORCIMT study.

I would like to thank the following people, whose help, support and guidance made this thesis possible. First I would like to express my gratitude to the participants in the NORCIMT study, as well as the management group of the NORCIMT study. I would also like to thank my supervisor Gyrd Thrane at the faculty of health science at the University in Tromsø, The Arctic University of Norway for inspiring and thorough counseling.

I would direct a special thanks to all participants in the NORCIMT study. I would also like to express my gratitude to the NORCIMT study for granting me access to participant's data.

My employers at the therapeutic department at the University hospital in Northern Norway (UNN) have been supportive by granting me study leaves, which I am grateful for.

I would also like to thank my fellow students, and especially Øyvind Bernhardsen for good discussions.

Last I would like to thank my family, Ann Kristin, Oliver Liam and Oscar Andreas for being supportive and giving me diversions from the writing.

Martin Vatshaug
Tromsø 2014

Abstract:

Background: The Wolf Motor Function Test (WMFT) is a test for motor function in the most affected upper extremity after stroke, measuring both performance time and functional ability of movement. The original version of WMFT has shown good psychometric properties, but these have not yet been investigated in the Norwegian translation. Inter-rater reliability is one of the properties that should be investigated before the test is implemented in assessment, rehabilitation and research. **Objective:** To assess the inter-rater reliability of the Norwegian translation of WMFT on hemiparetic stroke patients in the early phase after stroke. **Design:** Methodological Inter-rater reliability study. **Method:** 41 hemiparetic stroke patients (31 male, 10 female) with a mean age of 62.63 (11.56) and a mean time of 17.1 (7.1) days since the onset of stroke participated in this reliability study that is part of the NORCIMT study. Patients were assessed at baseline and videos were scored by two raters. For performance time, two- way mixed (3.1) intraclass correlation coefficients (ICC`s) was calculated to estimate inter rater reliability and Standard Error of Measurement (SEM) was computed to calculate measurement error. Performance time was also log10-transformed and analyzed. For the Functional Ability Scale (FAS) two- way mixed ICC`s and weighted kappa was computed to assess inter rater reliability. Analyses were made for three different models, using both the complete sample (n=41) and two subsamples (n=29 and n=12). **Results:** Total scores performance time had high agreement ($ICC_{agreement}$ =0.90) for all three models, while a minimum of 12 of 15 individual items had adequate agreement ($ICC_{agreement}$ >0.75) in all three models. Total score Functional Ability Scale had adequate agreement (ICCagreement = 0.76, Weighted kappa = 0.75). 6 of 15 individual FAS items had adequate $ICC_{agreement}$ (>0.75), while for weighted kappa 10 of 15 items had adequate levels (>0.61). **Limitations:** Sample consisted of a lower percentage of females (24.4%). Patients with cognitive impairments was not included, also subjects had a moderate to high level of functioning, limiting the generalizability. **Conclusion:** Inter-rater reliability of WMFT total scores was excellent for performance time, and adequate for Functional Ability Scale. For the individual items, reliability was adequate for nearly all performance time items For the FAS individual items reliability was adequate for under half the items using $ICC_{agreement}$, but 2/3 of the items were adequate when applying weighted kappa. Both performance time and FAS showed a very high internal consistency.

Key words: Inter-rater reliability, Wolf Motor Function Test, stroke, motor function, arm.

Sammendrag

Bakgrunn: Wolf Motor Function Test (WMFT) er en test for motorisk funksjon i den mest affiserte overekstremiteten etter hjerneslag og måler både utførelsestid og funksjonell kvalitet på bevegelse (Functional Ability Scale- FAS). Den originale versjonen av WMFT har vist gode psykometriske egenskaper, men disse har ikke blitt utforsket i den norske oversettelsen. Inter-rater reliabilitet er en av egenskapene som bør undersøkes før testen blir implementert i undersøkelse, rehabilitering og forskning. **Hensikt:** Å undersøke inter-rater reliabiliteten til den Norske oversettelsen av WMFT på hemiparetiske slagpasienter i tidlig fase etter hjerneslag. **Design:** Metodestudie som undersøker inter-rater reliabilitet. **Metode:** 41 hemiparetiske slagpasienter (31 menn, 10 kvinner) med gjennomsnittsalder på 62.63 (11.56) år med gjennomsnittlig 17.1 (7.1) dager etter slaget deltok i denne reliabilitetstudien som er del av NORCIMT studien. Pasienter ble undersøkt ved baseline og filmet. Videoer ble analysert av to ratere. For utførelsestid ble two- way mixed (3.1) intraklasse korrelasjonskoeffisienter (ICC`s) kalkulert for å undersøke inter rater reliabilitet, mens Standard Error of Measurement (SEM) ble kalkulert for å undersøke målefeil. Utførelsestid ble også log10-transformert og analysert.

For FAS ble two- way mixed ICC`s og vektet kappa kalkulert for å undersøke inter-rater reliabilitet. Analyser ble utført på 3 forskjellige modeller, både for den totale gruppen (n=41) og to subgrupper (n=29 og n=12). **Resultater:** Total score for utførelsestid hadde høy enighet (ICCagreement>.90) for alle tre modeller, mens ett minimum av 12 av 15 øvelser hadde adekvat enighet (ICCagreement>.75) for alle tre modeller. Total score for Functional Ability Scale hadde adekvat enighet (ICCagreement = 0.76, vektet kappa = 0.75). Minst 6 av 15 FAS øvelser hadde adekvat enighet (ICCagreement>.75), mens vektet kappa viste at 10 av 15 øvelser hadde adekvate verdier (>0.61). **Begrensninger:** Utvalget hadde en lavere prosent kvinner (24.4%). Pasienter med kognitive utfall ble ikke inkludert. I tillegg hadde pasienter moderat til høy grad av funksjon, noe som begrenser generaliserbarheten. **Konklusjon:** Inter-rater reliabilitet for WMFT total score var meget god for utførelsestid og adekvat for FAS. For de individuelle øvelsene var reliabilitet adekvat for nesten alle øvelser på utførelsestid. For FAS individuelle øvelser var reliabilitet adekvat for under 50 % av øvelsene når $ICC_{agreement}$ ble brukt, mens cirka 2/3 av øvelsene hadde adekvat verdi når vektet kappa ble brukt. Både utførelsestid og FAS viste meget høy intern konsistens. **Nøkkelord:** Inter-rater reliabilitet, Wolf Motor Function Test, hjerneslag, motorisk funksjon, arm.

Innhold

1 Introduction

A common goal in stroke rehabilitation is to promote personal independence and the ability to return to activities outside of the hospital setting (1). Stroke is one of the most frequent causes of death and disability in Norway, and stroke prevalence is expected to increase with an growing elderly population the next 50 years (2).

Persons suffering from stroke often experience disability's that limit their independence, and it is therefore of importance that the rehabilitation is well documented and efficient, securing the patients the best possible care (2). Standardized assessments used in clinical practice can help to identify and measure areas of problem, as well as being used as outcome measures for rehabilitation (3). Experiences from the clinic indicate that evaluation of the effect of interventions rarely takes place. Evaluation with standardized measurements should be performed regularly during the rehabilitation process to see if interventions have effect (2). The measurements or instruments used should have adequate psychometric properties, measuring what they intend to measure (4). Measurements are a central part of clinical practice as they form the basis for diagnosing, making prognosis and evaluating the results of interventions (5), so before choosing a particular measurement to use for this purpose we should ensure that its qualities have been tested through thorough research. In scientific research, data collection should be accurate, truthful and sensitive (6). With this in mind it should be clear that a minimum requirement is that the instruments and measures we use for assessment should be trustworthy (5, 7).

Reliability is an essential requirement for the measurements used in clinical practice and research, and a prerequisite for validity (5). Instruments and measurements should give the same results and scores, independent of the person administering the instrument, or if the same person administers the instrument twice (5).

There are several measures and instruments available for assessment of function in the upper extremity after stroke, one being the Wolf Motor Function Test (WMFT). It is developed in America by Taub et.al, and is commonly used as an outcome measure in many stroke studies (8), particularly in studies examining the effect of Constraint Induced Movement Therapy (9). Being one of the most frequently cited outcome measures in stroke rehabilitation (10), it has now been translated to Norwegian (11). It is also recommended for standardized assessment of hand- and arm function in stroke rehabilitation by the Norwegian Directorate of Health (2).

It is most commonly used by occupational and physical therapists, but demands no training (8, 12-14). The American version of WMFT has shown good inter-rater reliability (9, 14) and validity (9), but has not yet been assessed for validity and reliability in the Norwegian translation (11). Reliability is a central part of and prerequisite for validity (6). It is therefore needed to investigate the reliability of the Norwegian translation of WMFT before its validity can be established.

1.1 Purpose of the study

The aim of this study was to investigate the inter-rater reliability for the Norwegian translation of Wolf Motor Function Test. This study is part of the NORwegian-CI-Multicite Trial (NORCIMT), and includes 41 subjects with sub-acute hemiparetic stroke.
The purpose is to answer the following research question:

Is the Norwegian translation of the WMFT a reliable instrument when applied to hemiparetic stroke patients in the early phase after stroke?

1.2 Professional background

This master thesis is written by an occupational therapist working in rehabilitation. Although occupational therapists have their role in stroke rehabilitation, I consider highlighting occupational therapy and occupational therapy theory and philosophy beyond the scope of this study. Measurement properties of instruments and tests are relevant to all health workers, and this thesis therefore has a focus on measurement theory with an interdisciplinary angle.

2 Theoretical background

In this chapter literature and theory is presented to define and clarify key concepts, as well as presenting the theoretical background for this study. Literature related to the topic of the study was obtained from articles and textbooks. Literature searches were performed in different databases using the search term *Wolf motor function test* in combination with terms like; *reliability, stroke, rehabilitation, ICF, intraclass correlation coefficient, kappa, standard error of measurement.*

2.1 Stroke

In Norway, approximately 15 000 persons are affected by stroke every year, making stroke a leading cause of serious disability. It is also the third most frequent cause of death in Norway, hereby leading to significantly economic consequences for the society (2), in addition to the health consequences for the affected and their relatives (15). One stroke is estimated to cost around 600 000 Norwegian kroner, while the annual cost for strokes in Norway is above 7 billion Norwegian kroner (15). These costs are closely connected to the degree of disability, so treatment and rehabilitation that reduces degree of disability could also reduce the economic burden of society. Treatment in stroke units have shown decreased death and disability, assuming the treatment offered is of adequate quality (2). It is therefore important to document the outcome of rehabilitation programs with accurate assessments (14), which is further emphasized by

the increasing amount of elders, meaning that frequency of strokes could increase by 50 % the next 20 years (2). Paresis is one of the most frequent results of stroke, affecting over 80 % of the patients. Prominent impairments following stroke is reduced force, impaired motor control, reduced tempo of movement, disturbances with automated movements, increased tiredness and loss of coordination (2). Between 55-75% has lasting reduced motor function in one arm, causing problems in the ability to perform daily activities, participating in the work field and in recreational activities (2). Motor impairments can therefore be devastating for the patient and the relatives (3). When we consider the consequences of stroke, both economic and social, we would want the rehabilitation programs we implement to be effective. Edmans states that assessing motor impairments in a thorough way is essential for understanding the impact of stroke on functional tasks, and to form basis for developing an intervention plan (3). It is therefore of importance that the measures we use have adequate psychometric properties.

2.2 Wolf Motor Function Test

Wolf Motor Function Test is a test for motor function in the most affected upper extremity after stroke (9), and was developed by Wolf et.al to assess the effect of Constraint Induced Movement Therapy (CIMT) on survivors after stroke and traumatic head injury (14). Originally developed as the Emory Motor Test, for determining the time used by stroke patients on everyday tasks with the upper extremities, it was modified by Taub and colleagues. It consists of 17 tasks, 15 of these where time of performance is measured, and functional quality of movement, called Functional Ability Scale (FAS), is evaluated. The 2 remaining tasks are strength measuring tasks (8). Tasks 1-6 involves joint-segment movements, while the rest of the timed tasks are integrative functional movements (9). Detailed descriptions of the WMFT tasks are in the procedure of assessment (appendix 2). FAS is scored on a 6-point ordinal scale, ranging from 0 to 5, where 0 indicates that no attempt to move the upper extremity is observed, and 5 indicates that movement is considered normal (13).

There are several instruments and assessments that can be useful in assessing upper extremity function, but none has been accepted as the standard (14). One challenge is that many upper extremity motor function outcomes do not produce obvious links for planning treatment (9). WMFT assesses motor function in the upper extremity with focus on both quality and speed of movement (14), by quantifying upper extremity movement ability through timed and functional tasks (9). It also includes measures of impairment and disability, and therefore differs from other motor function assessments (8). Psychometric properties of the American version of the WMFT has been established (10), where Morris showed an intraclass correlation coefficient (ICC) of .97 for Time and .88 for the FAS-scale, and concluded that WMFT had high inter-rater reliability, internal consistency and test retest reliability (14). Wolf showed an ICC ranging from .97 to .99 and also concluded that inter-rater reliability was good for the American version (9). Filming is commonly used for assessing WMFT tasks in research (8, 9, 14, 16), but validity and reliability has also been found adequate without videotaping (17). In addition, WMFT has shown a high correlation with duration of movement in the more affected arm (1), as well as good construct and criterion validity (9). It also differentiates well between higher and lower functioning patients (8).

2.3 Measurement theory

This study is based on a quantitative research paradigm. This paradigm has become known as the traditional method of science, also including the rehabilitation field (4). Characterized by the emphasis on measurement, it has its roots from the development of physics and mathematics. It is closely related to the philosophical positions labeled positivism and logical positivism, which were proposed by the French philosopher Auguste Comte. The central idea of positivism is that only knowledge that can be verified through measurement and observation can be considered certain (4). The goal was for all sciences to follow the ideals of physics by rendering observations in an exact and objective form (18).

The quantitative paradigm has according, to Carter, Lubinsky and Domholdt, been based on some general assumptions (4). The first is that there is a single objective reality, where one goal of quantitative research is to determine the nature of this reality through measurement and observation. The second assumption is that investigator and subject can be independent of one another, where the investigator is considered a neutral and objective observer of a reality that is not altered through this study. The third assumption is that results of quantitative research should be generalizable characterizations of reality, making research lacking generalizability flawed. Fourth, it is assumed that causes and effect can be determined and differentiated from one another. The fifth assumption is that research should be value free, where the researcher is an impartial and objective discoverer, hereby avoiding influence of investigator opinions and societal norms (4). This gives the researcher an important responsibility controlling for systematic errors, random errors and confounding factors.

In quantitative studies we measure variables to derive data for understanding phenomena (6). Whether the data is nominal, ordinal, interval or ratio, different statistical methods are chosen analyses (6). In nominal scales, categories are classified without having a designated order, for example gender or nationality. Ordinal or categorical scales classify categories in a determined order, but the distance between the categories are not equal. Interval scale possesses the same properties as real numbers with constant distances between the values, like temperature. Ratio scales possesses the same properties as interval scale, but in addition has an absolute zero point, like weight or height (6). Methodological research is centered around documenting and improving measurements used in the clinic and research (4). When choosing an instrument or measure one must first know that it is suited for the purpose, and that its

measurement properties are adequate (5). Measurements are developed for different reasons, thereby having different measurement properties, like discrimination, prediction and evaluation (4).
As measurements form the basis for diagnosis, prognosis and evaluation of the results by the applied medical interventions, they are central to clinical practice and medical and health research (5). When we apply interventions to patients we do so hoping or expecting for some change in the patients functioning. If outcome measures are used to measure this change, they should at least address the aims the interventions are applied for (6, 19). All measurement used in medicine should therefore be tested for its properties (5), securing that they have adequate quality.

Systematic errors or bias are errors in the study design or in the implementation of the study that are recurring throughout the study, and could interfere with the basis for comparison with other studies (20). Random errors could be caused by the range of variation in the sample . The sample we are using to test WMFT should reflect the population which we intend to use WMFT in. A lower sample size increases the odds that the sample characteristics does not reflect the characteristics of the original population (20), decreasing the potential for generalizing the results. The sample size also affects to what degree we can obtain an acceptable confidence interval (CI) around the estimated reliability parameter. This is not a matter of statistical significance, since the issue is whether the reliability parameter approaches 1, and not its statistical difference from 0 (5). The requirements for sample size in reliability analysis is much debated, but De Vet et.al claims that 50 commonly is regarded as acceptable (5).

For outcomes research to be effective, it must use a systematic approach for describing and meaningfully classify outcomes (21). There are several challenges if one tries to make outcome research more standardized. It aims to understand the end results of health services, but results and discussions are often presented outside a common framework (6). Concepts like health status, functional status, well- being, quality of life and health related quality of life are often applied undifferentiated, making it difficult to understand- and to compare study results (19).
The tests we use should have a clear distinction between scoring alternatives, and be non-overlapping (5). In the case of WMFT the FAS scale has 6 alternatives for scoring, and it is

important that raters are able to distinguish between the different parts of the scale. If the categories are overlapping the chance that raters will disagree in scoring of the same patient could increase.

2.4 Methods for evaluating the quality of an assessment tool:

When performing studies to investigate the psychometric properties of a measurement or instrument, the primary aim usually is to improve its quality. In methodological research the goal is to document and improve qualities of clinical and research measurements (4).
Data of interest in methodological research is often referred to as "psychometric properties", which reliability and validity are a part of (5).
It is customary to start with investigating the reliability, as an instrument must be reliable to be valid (6).

2.4.1 Reliability:

Reliability can be defined as the proportion of total variance in the measurements (5), or how consistent an instrument or measure is when measuring a certain attribute (6). Reliability is not a constant characteristic of an instrument or measurement, meaning that a high level of reliability in one population, not necessarily is transferrable to another population (6).
In addition to reliability and measurement error, internal consistency is considered an aspect of reliability (5).

Reliability has several aspects and can be defined as:

> *"the extent to which scores for patients who have not changed are the same for repeated measurement under several conditions: e.g. using different sets of items from the same multi-item measurement instrument (internal consistency); over time (test–retest); by different persons on the same occasion (inter-rater); or by the same persons (i.e. raters or responders) on different occasions (intra-rater)"* (5).

The reliability of an instrument or measure is decided by its stability, consistency and equivalence. These terms are used by Polit and Beck (6).

When investigating relationships it is important to have a reliable instrument, because unreliable measures reduces statistical power and will therefore affect validity. The reliability coefficients are therefore important indicators of an instruments quality, where high reliability

is critical for interpreting research results. If the dataset doesn't support our hypothesis, one might instinctively assume that the expected relationship doesn't exist, but in fact it could be due to use of an unreliable instrument (6).

Reliability can be quantified as relative or absolute reliability. Relative reliability is based on the idea that if a measurement is reliable, then individual measurements within a group will stay in the same position within the group after repeated measurement. This can be measured using a correlation coefficient (4), which reflects the relationship of variability caused by measurement error to total variability in the data material (5). Choosing the appropriate statistical method depends on the nature of the data. For categorical data Cohen's Kappa is commonly used, while ICC is applied when the data is on the interval scale (5).

Absolute reliability concerns the extent a score varies after repeated measurement. This is measured using the standard error of measurement (SEM). If clinicians or researchers should be able to judge if patients conditions have changed, they must know how much variability could be expected due to measurement error (4). This can enable rehabilitation professionals to evaluate the clinical changes compared to changes that might be expected from measurement error (4).

The extent that an instrument gives similar scores on separate occasions is defined as its stability. Stability can be assessed using test-retest reliability procedures, where researchers use the same measure on a sample twice, and then compare the scores (6). Test-retest reliability is not investigated in this study, but WMFT has shown high test-retest reliability in American version (14).

Items need to have internal consistency when they are used to form a scale (22). If this involves summing of item scores, we usually evaluate their internal consistency (6). Internal consistency measures the extent items assess the same construct, and if any items measures something else they will have a lower inter-item correlation than other items (5).
It is, like other reliability measures, sample dependent. Therefore we cannot say that an instrument has high internal consistency, but rather high internal consistency in that specific population and situation (6). We desire a high degree of internal consistency, and expect scales that are designed to measure an attribute to be composed of items that measure only

that particular attribute (6). We can say that an instrument is internally consistent when its items measure the same trait (6), and therefore is correlated with one another (22). Examination of internal consistency can also be aimed at reducing the number of items, as items with low inter-item correlation possibly could be deleted (5).

The degree that two or more independent raters agree on scoring of an instrument is called equivalence. A high level of agreement indicates that there is a low level of measurement error (6). This is called the interrater or interobserver reliability, and can be assessed by using an intraclass correlation (ICC) (6). It has been common to use Pearson`s *r* to investigate whether raters scores are correlated, but it is not a very stringent parameter for assessing reliability since it does not take systematic errors into account (6). We aim for the raters to have a common interpretation of the construct, hereby reaching exact agreement. (6). This can be assessed by using a intraclass correlation coefficient for continuous variables, or weighted kappa for ordinal type variables (5). For this study the equivalence is of high importance, since it is an inter-rater reliability study. Equivalence is measured for both absolute agreement and consistency, and these terms are therefore used throughout the paper, instead of equivalence.

A reliability study can be divided into generalizability or decision studies, where we in generalizability studies aim to generalize results to other clinicians. If reliability is high then we can expect to be able to generalize the scorings from one clinician to another clinician. Decision studies aim to find the best strategy for achieving a high reliability, for example by doing more measurements and taking the average of these (5). This study is a generalizability study, since the aim is to see if the scorings of one rater can be generalized to the other raters in the study. If this is the case then we could expect the WMFT to be a reliable measure when used by other clinicians in a similar context, something that will be further discussed in the discussion.

2.4.2 Validity

The degree to which an instrument measures what it is supposed to measure is called validity (6). High reliability is a requirement for, but does not guarantee high validity. An inconsistent and inaccurate instrument cannot validly measure an attribute, because it contains too much error to be a valid indicator of the target variable (5). It is important to measure the construct

one intends to measure, and a reliable measurement is therefore made valid in combination of being reliable and bringing forth meaningful information (4, 5).

Validity is generally divided into external and internal validity (5), where internal validity concerns whether it is the independent variable that caused the outcome, and not something else. The challenge for the researcher is to rule out the plausibility that something else than the independent variable caused the observed relationship (6). In experimental research, the main question is whether the independent variable (treatment or intervention) caused the effect in the dependent variable (4), which usually is a patient or subject.

External validity is concerned with the generalizability of the results to other situations (5). One can generalize to other groups, settings or times with similar characteristics as the one studied (4). The results from this study should therefore be generalizable to populations with similar characteristics if external validity is good.

Three different types of internal validity can be identified: content validity, criterion validity and construct validity, in addition to face validity, which is an aspect of content validity (5). Face validity is how well the measurement instrument seems to measure the construct on first impression by doing an overall view. As this is a subjective judgment, it cannot be quantified, and there is no standard to how it should be done. Face validity is often underestimated because of this, but can be useful when choosing an instrument, as an overview of a questionnaire can give a good impression of whether it is suitable for the purpose (5). Even though not considered to give strong evidence of validity, face validity can be helpful when other types of validity has not been shown (6). Lack of face-validity is also considered a strong argument for selecting to not use an instrument (5), so clinicians planning to use a new questionnaire or instrument for assessing a patient would probably make an overview of the different elements to make a judgment of whether it is suited for the purpose. When viewing the different test items of the WMFT one should get a first perspective of whether it measures the construct it purports to measure. All items should be measuring motor function in the most affected upper extremity after stroke, so if any items seems to be out of place, this should be looked at closer by using statistical methods (5).

Content validity shows to what extent the content of the instrument is representative for the construct one intends to measure. When using multi-item questionnaires or assessments, all of the items should be relevant and including, covering all aspects of the construct to be measured (5). As there are no completely objective methods of evaluating an instruments content validity, it must be based on judgment. One way is to use an expert panel to evaluate whether the different items are relevant to the construct they are designed to measure, and whether they cover all the dimensions of the construct (6). Although based on qualitative assessment, it is possible to quantify content validity more by for example using the ICF as a framework. By classifying the items in accordance with the ICF- domains, one can compare several questionnaires content (5). This was done by Thrane et.al, by linking the WMFT aspects motor function, arm use and dependency together with the ICF domains bodily function/impairment, mobility and activity (23). It is not immediately obvious how WMFT as a measurement is linked to activity, but a study by Lang et.al showed a relationship between motor impairment and use of the more affected arm (1), suggesting that WMFT could be an indicator of real-world use of the upper extremities.

Criterion validity is the degree the scores of a measurement instrument gives a good reflection of a gold standard. Therefore it can only be investigated when a gold standard measuring the same construct is available (5). The key issue is to what degree the instrument is a useful predictor of other behaviors, experiences or conditions (6). This demands the availability of a reliable and valid instrument that the instrument can be compared to. The sample used to assess criterion validity should also be appropriate for the target population which it will ultimately be used in (5). What level of agreement is considered acceptable should be decided before comparing the two instruments, to prevent one from drawing positive conclusions on data that are showing a less than convincing correlation. Scores for the instrument and the gold standard must also be independent of one another (5). The WMFT criterion validity has been supported when compared to the Fugl- Meyer Motor Assessment, which is considered a reliable and valid measurement of upper extremity motor function after stroke (9).

Construct validity is, unlike criterion validity, applicable when there is no gold standard. It reflects whether the instrument gives the scores one would expect based on the existing knowledge about the construct (5). It is an important part of validity because constructs are used for linking the methods in a study to conceptualizations and mechanisms, thereby

making it possible to transfer results into practice (6). Although considered to be less powerful than criterion validity, it is possible to find evidence to support that the measurement instrument is measuring what it claims to. This demands strong theories and specific, challenging hypotheses (5). To maximize construct validity one must first clearly define the constructs one wants to measure, before making the construct measurable by operationalizing it (4). The construct motor function in the upper extremity is operationalized and measured as time of performance and quality of movement in the WMFT. Whether these constructs are good indicators of motor function will influence the internal validity (4).

2.4.3 Factors affecting reliability and validity for assessment tools

Reliability can be affected by a number of factors one should be aware of. Scales can partly be affected by their length, so by adding more items to address the same concept, reliability could be increased. The scales should however have discriminating power (6). Regarding the WMFT, this implies that it should be able to separate between patients with different levels of functioning. The American version of WMFT has been shown to have good discriminative power in patients with sub-acute stroke (8).

In the NORCIMT study patients are assessed with the WMFT before and after treatment. Raters should be able to register if changes in level of patient functioning have taken place, as well as being sensitive to differences between patients with lower level of functioning and patients with higher level of functioning. Training of observers is one effective way of enhancing reliability (6), ensuring that raters feel comfortable applying the WMFT and using the scoring manual.

The heterogeneity of a sample also affects reliability. A sample that elicits homogeneous answers will give a lower reliability coefficient, since instruments measure differences between those being measured (6). This means that when the sample has low diversity, reliability will decrease (5).

Furthermore reliability is not a fixed entity that can be transferred to all situations and populations, and it must be considered as a property of the instrument when applied to a certain population under certain conditions (5). Therefore, it is important to know the characteristics of the group an instrument was developed for, when choosing what instruments to apply (6).

It should also be noted that reliability varies according to what type of reliability is tested. Test-retest reliability will rarely give the same value as internal consistency or inter-rater reliability, making it a priority deciding which aspect of reliability is relevant when selecting instruments or measures (6). For this study interrater and internal consistency is investigated.

Cook and Campbell proposed several threats to internal validity. Of these 11 are highlighted as particularly important to the rehabilitation researcher (4). They are history, maturation, testing, instrumentation, statistical regression to the mean, assignment, subject attrition, interactions between assignment and maturation, history or instrumentation, diffusion or imitation of treatments, compensatory equalization of treatments and compensatory rivalry or resentful demoralization (4). History could threaten internal validity if events unrelated to the treatment occur during the study, possibly changing the dependent variable. Maturation or changes within the participant during the study could influence the dependent variable. Repeated testing is likely to change results in the dependent variable, because participants could become more familiar with the test, giving improved measurements (4). These particular threats are apparent when several measurements are performed over time, allowing the possibility for change in the participants.

Changes in the measurement tools could also be responsible for changes in the dependent variable. This is called instrumentation and is especially apparent when using technical equipment that needs calibration between each testing, but could also occur when researchers themselves are measuring tools (4). Humans could evaluate the same situation in different ways, for example when rating a WMFT video of the same patient. Training is one way of calibrating humans as measuring tool (4), and an effective way to improve reliability (5).

Another threat is that extreme values tend to regress to the mean. This could be present when participants are included based on a single measure as criteria for inclusion. Single scores may not reflect true ability, something that could be avoided by using the mean of measurements over time as criteria for inclusion.
Assignment of participants to group poses a threat to internal validity if assignment is not performed randomly (4). In interrater reliability analyses where all participants are analyzed at the same time, assignment poses a minimal threat if all participants are rated by all raters. Still

it could pose a major threat if repeated measurements are analyzed to investigate effect of an intervention (4).
Losing participants during the study could cause groups to have different characteristics at the end of the study, compared to in the beginning, thereby affecting internal validity (4). Interactions between assignment and maturation, history or instrumentation could also occur and could affect treatment and control groups in different ways (4). When all participants is assessed at a single time point in an inter rater study this threat would be less present, although one could imagine that the functional level of participants assigned could influence agreement of scoring between raters. If participants in a treatment and control group share their experiences about their treatments with each other, this could cause diffusion or imitation of treatments. By minimizing contact between participants this can be controlled, as well as informing them about the importance of adhering to their specific regimens (4).
If researchers give extra attention to one group this could lead to increased effort and adherence in this group, thereby affecting the dependent variable. Also rivalry between groups with different treatments could give increased or decreased effort in participants (4).

External validity is also threatened by several factors, most notably selection, setting and time (4). If participants selected for inclusion is different from the original population which one wants to generalize the results to, this would decrease external validity. To avoid this, strict criteria for inclusion is required, controlling the participants selected to be as much alike the target population as possible. Controlling the setting where research is conducted is also necessary to ensure external validity (4). The videos with the WMFT assessments should be standardized and conducted in the same manner to increase external validity. Time could be a threat to external validity because knowledge and practice changes as time pass by, possibly making results less applicable when they are published. Researchers should describe possible changes that may influence the application of results in the future, compared to when data were collected (4).

3 Method

In this chapter key concepts and themes related to the sample, instrument and the different analyses will be further explained.

3.1 Design and sample

This study is a quantitative cross sectional reliability study of the 15 timed items of WMFT, where subjects where assessed using the WMFT at Pre 1-baseline, before they had received any treatment.

The study is based on a sample of 41 hemiparetic stroke patients who were recruited to the NORwegian- CI- Multicite- Trial (NORCIMT) from September 2008 (24). The NORCIMT study is approved by the Regional Ethics Committee (REK) and has the number 2009/1903. The patients were recruited to investigate if modified Constraint Induced Movement Therapy (mCIMT) in the early phase after stroke gave better results than traditional treatment, and how it compares to mCIMT 6 months after stroke (24). As seen in figure 1, patients were recruited from 5 hospitals, while the assessment of the patients was done on 4 locations. All patients were assessed at baseline after inclusion and before randomization to control or mCIMT-treatment group. Control group received mCIMT treatment approximately 6 months after onset of stroke (24).

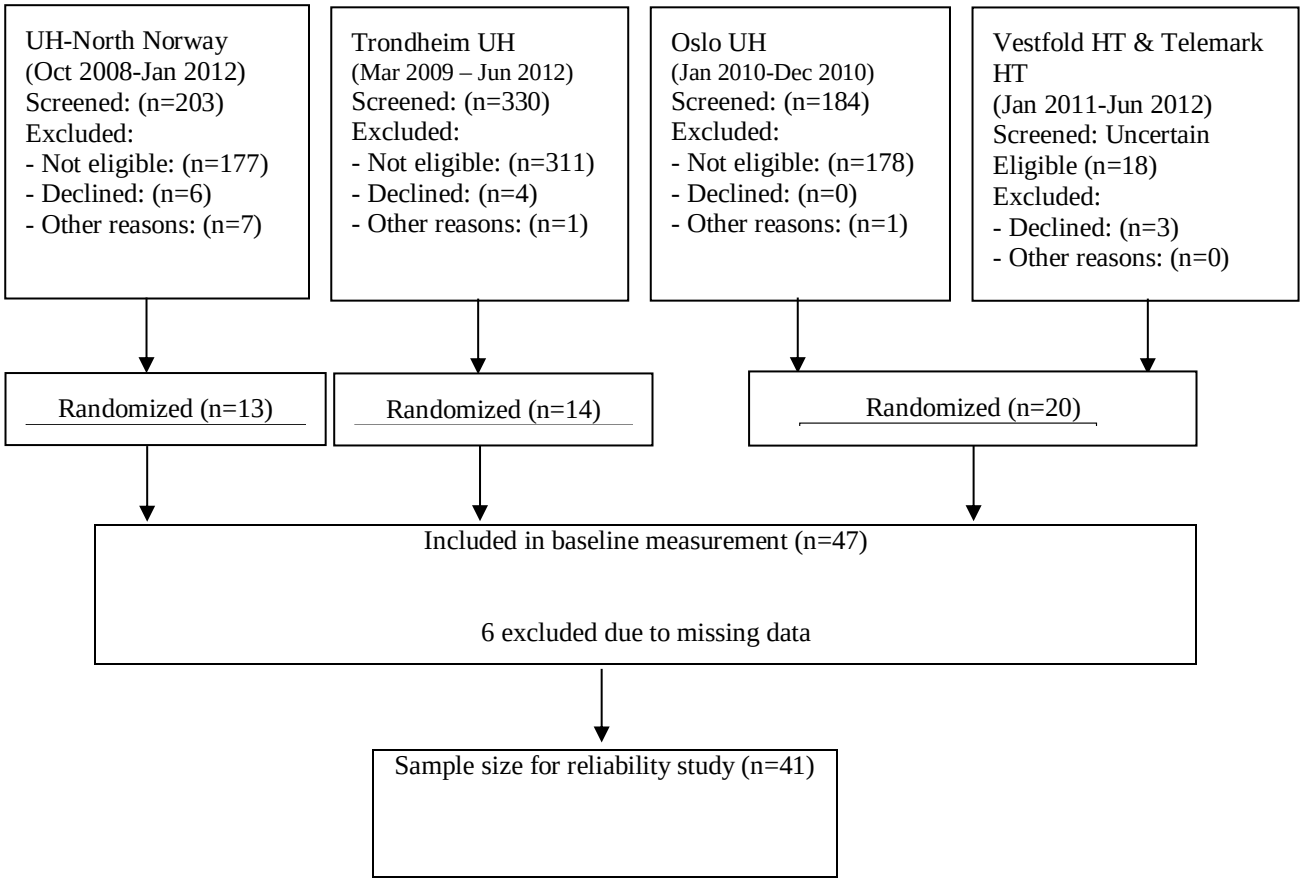
3.1.1 Inclusion criterias

- Stroke at more than 5 and less than 26 days ago (Either first stroke or second stroke without detectable arm weakness after the first stroke).
- Modified rankin scale 0-2 before admission
- Persistent unilateral arm or hand paresis (Scandinavian Strokes scale (SSS) arm motor function 2-5 or SSS hand motor function 2-4)
- Able to lift two fingers with the forearm pronated on the table or able to extend the wrist at least 10 degrees from fully flexed position.
- Able to follow a two-step command.
- Mini Mental State examination score of more than 20 (or more than 16 in combination with expressive aphasia)
-

3.1.2 Exclusion criterias

- Modified Rankin Scale > 4
- Unable to give informed consent
- Large hemispatial neglect (more than two cm on the Line Bisection Test)
- Not expected to survive one year due to other illnesses (eg cardiac, malignancy)
- Injury or condition in the affected upper extremity that limited use prior to the stroke.
- Other neurological condition affecting motor function

Figure 1: Flow diagram of the inclusion in the Norwegian CI-Therapy Multicite Trial interrater reliability study



3.1.3 Raters

11 raters rated the videos, with 1 rater rating all 41 videos, and the other 10 raters rating a varying amount of videos. Of the 10, one rater rated 12 videos, one 7 videos, one 6 videos, one 4 videos, one 3 videos, four 2 videos and one 1 video. The rater who rated all 41 videos is referred to as *rater 1*, the rater that rated 12 videos is *rater 2*, while the remaining raters are referred to as *other raters*. Rater 2 and other raters combined are referred to as *all raters*. Raters were trained in the administration of the WMFT, and masked to the treatment group designation of the participants. 10 raters were coursed for four days, and were given two days training in the treatment protocol, undergoing a standardization procedure. Rater 1 was included in the study later and was therefore trained independently by scoring 4 WMFT videos unrelated to the reliability study, and discussing these scorings with a representative from the NORCIMT study. Afterwards rater 1 and rater 2 were calibrated by separately rating 3 videos, and having a meeting over telephone where ratings were discussed to form a common agreement on scoring. These 3 videos were not used in the reliability analysis.

3.1.4 Procedures for assessment

The Norwegian translation of WMFT by Dahl, Stock, Langøren and Askim was used. This has some adjustments made from the original version.

Final timescore is the median of all timed tasks. Both the median and mean of 15 items was calculated. 120 seconds is the maximal time allowed for performing a task. If the subject was unable to perform the task, the score 121 seconds was given. To standardize the placing of the test equipment a template with marked guidelines was placed on the table. These were printed on paper, or transparent paper. A detailed description for positioning of the chair and camera was used. Procedures for how to carry out and score each task was described in detail, in addition to the general guidelines for how to conduct the WMFT. Videos had a quality of 25 frames per second, making it possible to time tasks down to every 0.04 seconds. Windows Movie Maker was used for analyzing the videos.

Timing when using video is not mentioned in the original protocol, and therefore a procedure was made to adress this. Performance time is decided by using frame by frame videoanalysis. Each test is videotaped, but the instructions between the tests are not taped to decrease the amount of work during analysis. The examiner had to watch through the video immediately after the test to check if all test items are included on the tape. This will give the patient a

short break and ensure that all content is included. Tests not included must be repeated. After the assessment, performance time of each test is timed in the following manner:

- To ensure that start time is included, camera is first turned on. Then the patient is asked if he/she is ready before saying "ready, set, go!"
- When going through the video the person doing the analysis fast-forwards the tape to the point where the tester says "…go!"
- By going back and forth, frame by frame, one finds the first image frame where the patient moves shoulder, arm or hand.
- The first frame after "…go!" where the patient has moved shoulder, arm or hand is set as starting time.
- Then the picture is moved forward until the patient has placed the arm in the end position as described for each test. The first picture in end position is set as end time.
- Patients time score for each test is calculated using time codes for start and end time. Score is rounded to the nearest tenth of a second and written down in the form.
- When the film is analyzed it is sent for reliability analysis and storing at the University Hospital of Northern Norway (UNN).

Scoring of the FAS is done by using the instructions shown in table 1 when watching the WMFT assessment videos. In addition, each task had instructions for scoring that particular task, defining how movement should be performed (appendix 2).

Table 1: Functional Ability Scale

Scoring of the Functional Ability Scale

0- Does not attempt to use the upper extremity (UE) being tested.

1- UE being tested does not participate functionally, but attempt is made to use it. In unilateral tasks the UE not being tested may be used to move the UE being tested.

2- Uses UE, but assistance of UE not being tested is required for minor adjustments or change of position. Or: More than two attempts is needed to complete task. Or: Task is done very slowly. In bilateral tasks UE being tested is used only for support.

3- Uses UE, but movement is to some degree influenced by synergy, done slowly or with effort.

4- Uses UE, movement is close to normal*, but slightly slower, may lack precision, fine coordination or fluidity.

5- Uses UE, movement appears to be normal*.

* To decide what is considered normal, the unaffected UE may be used for comparison. Hand dominance before stroke is taken into consideration.

3.5 Statistical analysis

Data in the study was analyzed with IBM SPSS Statistics version 21, and MedCalc version 13. The nominal, ordinal and interval nature of the variables (25) have been taken into consideration when selecting the appropriate analyses to be done. Demographic data is described using descriptive statistics. Means and standard deviations were used to describe normally distributed variables, while median and interquartile ranges were used for non-normally distributed variables. It is often preferable to use the median instead of the mean, even on interval variables, because it is not affected by outliers. The interquartile range gives a more robust description of the spread in the material, because outliers are removed (20).

The distribution of all variables was explored using histograms and normal probability plots. Visual inspection of scatterplots was performed when looking for outliers regarding disagreement between raters. Since the WMFT-time has non-normality with positive skew, a Log10-transformation was performed to investigate if this improved the distribution or changed the ICC levels. If data are non-normal or variances are unequal, a transformation may be appropriate to perform, as it changes the scores to correct for distributional problems, outliers or unequal variances (25).

Mean of timed items and median of timed items for rater 1, rater 2, all raters and other raters were calculated on the 15 items. Paired T-test was computed to assess the relationship of the means from the raters on WMFT-Time. When the measures obtained are from the same people, they are not independent anymore, and should therefore be computed by using a paired/dependent t-test (6). A t-test is used on interval or ratio scales and is based on the normality distribution assumption, but is not dependent on the assumption of equal variances (20). A p-value below 0.05 means that there is a significant difference between means (25).

Median of scored items was calculated for rater 1, rater 2, all raters and other raters on the FAS. A Wilcoxon signed-rank test was used to assess the relationship of the median of scored items from the raters. As ordinal data cannot be characterized using the normal distribution, a non-parametric test is needed (20). The Wilcoxon signed rank-test is considered the non-parametric counterpart of the dependent t-test (25), and is well suited for ordinal data (20).

Agreement on performance time was calculated using intraclass correlation (ICC), with a two-way mixed absolute agreement model. A two way mixed model is also referred to as 3.1, and takes the systematic error between raters into account (26). This was calculated using single score values. ICC ranges from 0 to 1, where 1 means that the error variance is negligible compared to the patient variance, while 0 means that it is extremely large (5). The closer the ICC value for the WMFT is to 1 therefore indicates a high reliability. Generally, an ICC between 0.50-0.69 is considered as a moderate correlation, 0.70-0.89 high and 0.90-1.00 as very high (4). For this study an ICC above 0.75 is considered adequate, as this is the level used in the reliability study on the American WMFT performed by Morris et.al (14).

Rater agreement on WMFT-FAS was calculated using Weighted Kappa with quadratic weights. Weighted Kappa is a reliability parameter for ordinal variables and ranges from -1 to 1. By applying weights we ignore that we are using an ordinal scale, and pretend that the distances between classes are equal (5). Linear or quadratic weights can be applied (5), but the determination of which weights to apply is a subjective issue even experts can disagree on (27). Quadratic weighting is considered the most common type of weighting (5, 28), and is often recommended because its coefficients are equivalent to the intraclass correlation coefficients (29). Unweighted kappa tends do decrease when numbers of categories increase, while quadratic weights increase with increasing number of categories (29). Kappa values was interpreted using the classification by Landis & Koch, where a kappa value of 0.41-0.60 generally is considered moderate, 0.61-0.80 substantial and 0.81-1 almost perfect (5, 27). A kappa of above 0.61 will be considered adequate in this study. $ICC_{agreement}$ was also used on WMFT-FAS for comparison with the studies of Morris et.al and Wolf et.al (9, 14).

Two- way mixed $ICC_{consistency}$ (3.1) was used when calculating consistency of scores between raters on performance time. ICC can be calculated for both consistency (ranking), and absolute agreement. De Vet e.al (2011) points out that we in medicine rarely are interested in the ranking of patients, and absolute agreement therefore is the most relevant analyses for reliability (5). One example of when ranking is appropriate is when we have to prioritize which patients should get at certain treatment, based on their condition. This is what De Vet et.al calls $ICC_{consistency}$ (5). Morris et.al reports both the consistency and absolute agreement ICC values, when reliability testing the original WMFT (14), so for comparison with their study, I chose to perform analysis with both versions of the ICC.

Internal consistency of the FAS scale and performance time was assessed using Cronbach's alpha. This was calculated on the scores of rater 1. When tests involving summing of item scores, it is common to evaluate their internal consistency, and Cronbach's alpha is the most commonly used measure (6). Cronbach's alpha depends on the number of items, and like all other reliability parameters depends on the variation in the population. Heterogeneous populations will get higher values of Cronbach's alpha, than homogeneous populations (5). Nunnally recommended a level of Cronbach's alpha above 0.80 for basic research tools, and 0.90 as the lowest tolerable level for clinical purposes. However, Streiner points out that a

Cronbach's alpha above 0.90 could be an indication of redundancy in the scale (30). A level of 0.80 was therefore considered adequate for this study.
An inter-item correlation above 0.50 is considered adequate, as this was the level used by Morris et.al for the American version of WMFT (14).

Inter-rater reliability was explored both for the mean and median of timed items, median of the Functional Ability Scale, and for the individual WMFT items.

Measurement error for performance time was calculated using Standard Error of Measurement (SEM). This was done by taking the square root of the within subject variance from the ANOVA table, as explained by Bland & Altman (31). The size of measurement error can be measured by taking the standard deviation of repeated measurements on the same subject, which is commonly known as the *within-subject standard deviation* or S_w *(31)*. This is a parameter of measurement error, that measures how far apart the outcomes of repeated measurements are (5). It is expected that the difference between a subjects measurement and the true value is less than 1.96 S_w for 95 % of observations (31). The SEM quantifies the precision of individual scores on a test, and has the same units as the measurement used (32). SEM therefore, unlike the ICC, provides an absolute index of reliability (5). Cronbach states that he considers the SEM as the single most important information to report regarding an instrument. Unlike a coefficient, it is easy to understand, and therefore is more applicable (33). For WMFT-time, SEM shows measurement error in seconds, which gives a clear view of the size of measurement error between the raters. It is not calculated on WMFT-FAS, as there are no parameters of measurement error for categorical variables (5).
It is largely independent from the population we determined it from, and could therefore be considered a fixed characteristic of the measure, regardless of the sample (32). It should be noted that the SEM is partly a function of the sample size, and therefore will decrease when increasing the sample size (6). Unlike the ICC, it is not affected by between-subjects variability and may be more informative as an absolute reliability measure (32).

The reliability study is part of the NORCIMT study and had approval for a reliability study to be performed. .All participants had given informed and written consent for two raters to score each video. All raters were blinded, and had no knowledge of the other rater's scores.

I have worked as a research assistant on the NORCIMT study from January 2011 to December 2013. My main role in the study has been administration, storing and analysis of research material. I have analyzed and rated all WMFT videos. At the time I analyzed the WMFT videos, this study had not yet been planned and my ratings should therefore not be influenced by my position in the NORCIMT study. Throughout the process I have also done my best to remain objective, and give ratings based on the procedure, but it is still important to clarify my role in the study

4 Results

4.1 Descriptives

4.1.1 Participants

Over 700 patients that were screened for participation in the NORCIMT study from October 2008 to June 2012. Of these, 47 patients were recruited for the NORCIMT study, but 6 patients were not analyzed by 2 raters due to limited resources and logistical reason. Only 41 analyzes were therefore used in this reliability study. From table 1 we see that, on average, participants was assessed 17.1 (7.1) days after stroke, with a range of 7-32 days.
Mean Fugl Meyer score was 48,93 (sd 10,90), ranging from 21-66. Mean Modified Rankin Scale was 2.6 (0.8) and 63.4 % of the participants had a NIHSS score of 0 for the affected arm.

Table 1: Clinical characteristics of participants for full sample and subsamples

	Ra 1 vs all n=41	Ra 1 vs Ra 2 n=12	Ra 1 vs other n=29
Age, mean (SD)	62,63 (11,56)	63,17 (12,79)	62,41 (11,24)
Range	34-85	35-85	34-82
Females, n (%)	10 (24, 4%)	2 (16, 7%)	8 (27, 6%)
Days post stroke, mean (SD)	17,10 (7,1)	18,42 (6,8)	16,55 (7,9)
Range	7-32	10-32	7-29
meanFMA (SD)	48,93 (10,90)	52,17 (10,35)	47,59 (11,02)
Range	21-66	29-66	21-61
Modified Rankin Scale (sd)	2,6 (0,8)	2,5 (0.8)	2,6 (0,8)
Range	1-4	2-4	1-4
NIHSS Mean (sd)	1,8 (1,9)	0,5 (0,7)	2,4 (2)
NIHSS Affected arm			
0	26 63,4 %	12 100 %	14 48,3 %
1	11 26,8 %	0 0 %	11 37,9 %
2	3 7,3 %	0 0 %	3 10,3 %
3	1 2,4 %	0 0 %	1 3,4 %
4	0 0 %	0 0 %	0 0 %

4.2 Performance time:

Performance times for all models had a positive skew, indicating a non-normal distribution. For rater 1 vs all raters, rater 1 performance times had a skewness of 2.06, while for all raters performance times this was 1.99. Rater 1 vs rater 2 showed a skewness of 1.98 for rater 1 and 1.70 for rater 2, while for rater 1 vs other raters skewness was 2.04 for rater 1 and 1.97 for other raters.

Log transformed performance times skewness was lower than for the untransformed performance times. Rater 1 vs all raters showed a skewness of 1.27 for rater 1 and 1.06 for all raters. In the rater 1 vs rater 2 model, skewness was 1.07 and 0.87 for rater 1 and rater 2. For rater 1 vs other raters skewness was 1.35 and 1.24 for rater 1 and other raters respectively. The distribution before and after log10-transformation is shown with histograms in appendix 3.

Table 2: Performance time descriptive statistics, inter-rater reliability with upper and lower limit of the 95% confidence interval and measurement error shown with the Standard Error of Measurement.

	Mean times (sd)	$ICC_{agreement}$ (95% CI)	SEM
Mean time			
Ra1 vs all	17.27±22.08 vs 19.51±25.47	0.93 (0.88-0.96)	6.03
Ra1 vs Ra2	16.84±18.34 vs 18.42±19.15	0.98 (0.95-0.99)	1.96
Ra1 vs other	17.38±23.63 vs 19.99±27.96	0.92 (0.84-0.96)	7.04
Log mean time			
Ra1 vs all	0.655±0.344 vs 0.700±0.347	0.93 (0.87-0.96)	
Ra1 vs Ra2	0.688±0.260 vs 0.775±0.244	0.91 (0.15-0.98)	
Ra1 vs other	0.641±0.376 vs 0.669±0.381	0.94 (0.87-0.97)	
	Median times (IQR)		
Median time			
Ra1 vs all	3.52 (2.76) vs 4.00 (2.50)	0.95 (0.90-0.97)	7.31
Ra1 vs Ra2	3.36 (1.29) vs 4.15 (1.92	1.00 (0.99-1.00)	0.44
Ra1 vs other	3.71 (3.12) vs 3.90 (2.95)	0.93 (0.85-0.96)	8.70
Log Median Time			
Ra1 vs all	0.684 (0.519) vs 0.708 (0.506)	0.96 (0.93-0.98)	
Ra1 vs Ra2	0.575 (0.406) vs 0.677 (0.375)	0.98 (0.94-0.99)	
Ra1 vs other	0.554 (0.447) vs 0.625 (0.461	0.95 (0.91-0.97)	

Rater 1 mean of timed items was not different from All raters mean of timed items (p=0.087). This was also the case for the median of timed items of rater 1 and all raters (p=0.821). For the Log10- transformed data the mean was different (p=0.016), but the median of timed items was not different (p=0.271). Rater 1 and rater 2 mean of timed scores were not significantly different from each other (p=0.055), but median of timed scores showed a difference (p= 0.025). The log10-transformed data for rater 1 vs rater 2 showed a difference for the mean of timed items (p=0.00), and for the median of timed items (p=0.046). Rater 1 vs other raters were not different from each other for the mean (p=0.164), or median of timed items (p=0.770). Log10-transformed performance time showed no difference for the mean (p=0.247) or the median of timed items (p=0.544).

From table 2 we can see that agreement was very high for total scores of all rater models. This was the case for both the mean and median of timed items, with the median giving slightly higher levels of agreement and narrower confidence intervals. Using the mean or median of timed items did not affect the SEM very much, but SEM was considerably lower for the rater 1 vs. rater 2 comparisons. For the Log10-transformed performance time we see that $ICC_{agreement}$ is excellent for all models, with narrow confidence intervals, with the exception of the confidence interval for mean of timed items for rater 1 vs. rater 2, which became very wide when the Log10-transformation was applied.

Consistency for the median of timed items was near perfect on all rater models, with narrow confidence intervals. $ICC_{consistency}$ was 0.94 (0.90-0.97) for rater 1 vs. all raters, 1.00 (1.00-1.00) for rater 1 vs. rater 2 and 0.92 (0.85-0.96) for rater 1 vs. other raters. The mean of timed items gave an $ICC_{consistency}$ slightly lower, but still near perfect with the values 0.93 (0.88-0.96), 0.99 (0.97-0.98) and 0.92 (0.85-0.96) for the three different models.
$ICC_{consistency}$ for Log10- transformed items did not differ much from non-transformed data and gave the values 0.96 (0.93-0.98), 0.99 (0.97-0.99) and 0.95 (0.90-0.97) for the median of timed items for the different rater models. For the mean of timed items these values were 0.94 (0.89-0.96), 0.97 (0.90-0.99) and 0.94 (0.87-0.97). As we can see from the numbers, the confidence interval of rater 1 vs. rater 2 stayed narrow and high, unlike the confidence interval for $ICC_{agreement}$ values.

4.2.2 Individual items performance time

For rater 1 vs all raters ICC ranged from 0 to 0.99, with 12 of 15 test items above 0.75 for both $ICC_{agreement}$ and $ICC_{consistency}$. SEM had a mean of 10.03 seconds, ranging from 0.51 to 21.53 seconds. Test items “forearm to table”, “forearm to box” and “reach and retrieve” were below adequate ICC level. These were heavily affected by outliers, where one rater had considered a task “completed”, and the other considered it “not completed” and scored it with 121 seconds.
In addition to the abovementioned items, the items “extend elbow”, “extend elbow with weight”, “lift can”, “lift pencil”, “lift paperclip”, “Stack checkers”, “flip cards”, “turn key in lock” and “fold towel” were affected by outliers and large measurement error, despite having adequate ICC levels. When cases of outliers were removed from the analysis of rater 1 vs all

raters, these items gained a notably higher $ICC_{agreement}$, as well as narrower confidence intervals and lower SEM. By removing the outliers all the items except for "forearm to table" achieved a well above adequate level of $ICC_{agreement}$. The model with rater 1 vs. other raters behaved in the same manner as rater 1 vs. all raters, with all items except "forearm to table" gaining adequate level of $ICC_{agreement}$ when outliers were removed.

Rater 1 vs. rater 2 had 11 of 15 items above 0.75, with "Forearm to table", "forearm to box", "hand to table" and "hand to box" having below adequate ICC. Mean SEM was low with a value of 3.1 seconds, which was considerably lower than for the other models. Outliers was not as apparent in this analysis as for rater 1 vs. all raters and rater 1 vs. other raters, but items "lift pencil" and "flip cards" had outliers where raters had major disagreements. With these removed from the analysis, the items gained adequate $ICC_{agreement}$ and mean SEM for 15 items decreases from 3.1 seconds to 0.96 seconds.

For the Log10- transformed performance time individual items there were 13 of 15 items with $ICC_{agreement}$ above 0.75, one more than for the non-transformed data. "Reach and retrieve", which had very low level of ICC on the non-transformed data, gained an adequate level when transformed. Only "forearm to table" and "forearm to box" had inadequate levels of ICC. Rater 1 vs. rater 2 had 9 of 15 items above 0.75, two less than for the non-transformed data, while rater 1 vs. other raters had 13 of 15 items above adequate ICC level. In general the models with rater 1 vs. all raters and rater 1 vs. other raters fared better with transformed data, while the model with rater 1 vs. rater 2 actually fared worse using transformed data.

Table 3: Descriptive statistics, ICC and 95% confidence interval for the individual items of the WMFT on rater 1 vs. all raters.

TIME	Rater 1 Median (min-max)	All raters Median (min-max)	$ICC_{agreement}$ (95%CI)	SEM
1 Forearm to table	1 (0.4-3)	1.3 (0.5-121)	0 (0-0.29)	13.29
2 Forearm to box	2 (0.5-121)	2.1 (0.4-121)	0.65 (0.43-0.79)	13.31
3 Extend Elbow	1.2 (0.4-121)	1.5 (0.5-121)	0.78 (0.63-0.88)	14.95
4 Extend elbow w/ weights	1.1 (0.3-121)	1.5 (0.4-121)	0.93 (0.88-0.96)	6.03
5 Hand to table	1.2 (0.6-5.4)	1.3 (0.5-5.8)	0.83 (0.70-0.82)	0.51
6 Hand to box	1.1 (0.4-121)	1.2 (0.5-121)	0.99 (0.99-0.99)	0.87
8 Reach and retrieve	1.2 (0.4-8.2)	1.7 (0.5-84.8)	0.16 (0-0.44)	8.55
9 Lift can	3.7 (1.9-121)	3.9 (1.4-121)	0.80 (0.66-0.89)	19.04
10 Lift pencil	5.1 (1.2-121)	6.1 (1.4-121)	0.77 (0.61-0.87)	21.53
11 Lift paper clip	5.1 (1.4-121)	5.5 (2-121)	0.89 (0.81-0.94)	13.45
12 Stack checkers	18.3 (4.4-121)	28.5 (0.7-121)	0.98 (0.96-0.98)	6.64
13 Flip cards	15.1 (5.5-121)	15.2 (5.8-121)	0.97 (0.94-0.98)	6.47
15 Turn key in lock	8.8 (3.2-121)	8.8 (3.4-121)	0.93 (0.88-0.96)	9.09
16 Fold towel	14 (5.8-121)	16.3 (3.5-121)	0.88 (0.78-0.93)	12.96
17 Lift basket	5.1 (1.9-121)	5.8 (1.8-121)	0.99 (0.98-0.99)	3.78

4.3 Functional Ability Scale

4.3.1 Total scores

In general FAS ICC`s were lower than for performance time. From table 4 we see that total FAS scores had adequate $ICC_{agreement}$, except for rater 1 vs. other raters, which was slightly below 0.75. $ICC_{consistency}$ was adequate with the values of 0.77 (0.61-0.87) for rater 1 vs. all raters, 0.83 (0.52-0.94) for rater 1 vs. rater 2 and 0.78 (0.58-0.89) for rater 1 vs. other raters. Weighted Kappa values indicated that the comparisons of rater 1 vs. all raters and rater 1 vs. other raters had substantial amount of agreement, while rater 1 vs. rater 2 was slightly above the "near perfect agreement" limit proposed by Landis & Koch (34). Both agreement and consistency values were lower for the FAS than for performance time, with rater 1 vs. rater 2

having the best ICC`s. Confidence intervals was however wide for all models, going well below adequate ICC level.

Table 4: Functional Ability Scale total scores medians, inter-rater reliability with lower and upper limits of 95% confidence interval

Functional ability scale	Medians (range)	$ICC_{agreement}$ (95%CI)	Weighted kappa
Ra1 vs all	3 (1-5) vs 4 (1-5)	0.76 (0.58-0.86)	0.75
Ra1 vs Ra2	4 (1-4) vs 3.5 (1-4)	0.84 (0.54-0.95)	0.83
Ra1 vs other	3 (1-5) vs 4 (2-5)	0.74 (0.46-0.87)	0.70

4.3.2 Individual items

For the FAS individual items, 6 of 15 items had adequate $ICC_{agreement}$ for the rater 1 vs. all raters comparison. For rater 1 vs. rater 2, 8 of 15 items had adequate $ICC_{agreement}$, while 5 of 15 items were adequate for rater 1 vs. other raters.

When looking at individual items $ICC_{agreement}$, rater 1 vs. all raters had 6 of 15 adequate items, rater 1 vs. rater 2 had 9 of 15 adequate items and rater 1 vs. other raters 7 of 15 adequate items. Weighted Kappa showed that rater 1 vs. all raters had 10 of 15 adequate items. Rater 1 vs. rater 2 had 11 of 15 adequate items, while rater 1 vs. other raters had 10 of 15 adequate items.

Items “forearm to table”, “forearm to box”, “extend elbow with weight”, “hand to table”, “hand to box” and “reach and retrieve” had inadequate levels of both $ICC_{agreement}$ and $ICC_{consistency}$ for all rater models.

For weighted kappa The items “Forearm to table”, “forearm to box” and “hand to box” had below adequate kappa levels for all the rater models, but for rater 1 vs. all raters, which is the full sample, “extend elbow” and “reach and retrieve” also had below adequate levels. Rater 1 vs. rater 2 had 11 of 15 items with adequate weighted kappa, but confidence intervals were quite wide for this comparison.

Table 5: Descriptive statistics, inter-rater reliability with lower an upper limit of the 95% confidence interval for the individual items of WMFT-FAS

FAS	Rater 1 Median (min-max)	All raters Median (min-max)	Weighted kappa (95% CI)	$ICC_{agreement}$ (95%CI)
1 Forearm to table	4 (3-5)	4 (2-5) 1	0.31 (0.12-0.51)	0.32 (0-58)
2 Forearm to box	4 (1-5)	4 (2-5) 2	0.42 (0.28-0.56)	0.43 (0.08-0.67
3 Extend Elbow	3 (1-5) 1	4 (2-5) 4	0.54 (0.33-0.75)	0.61 (0.30-0.79)
4 Extend elbow	3 (1-5)	4 (1-5) 1	0.68 (0.53-0.83)	0.69 (0.19-0.86)
5 Hand to table	3 (3-5)	4 (3-5) 2	0.61 (0.42-0.80)	0.62 (0.26-0.8)
6 Hand to box	3 (2-5)	4 (2-5) 2	0.50 (0.31-0.70)	0.51 (0.25-0.71)
8 Reach and retrieve	4 (2-5)	4 (2-5) 1	0.58 (0.34-0.81)	0.59 (0.34-0.76)
9 Lift can	3 (1-5)	4 (1-5) 3	0.84 (0.76-0.93)	0.86 (0.76-0.92)
10 Lift pencil	3 (1-5)	3 (1-5) 3	0.86 (0.78-0.94)	0.87 (0.77-0.93)
11 Lift paper clip	3 (1-5)	3 (1-5) 3	0.86 (0.77-0.95)	0.87 (0.77-0.93)
12 Stack checkers	3 (1-5)	3 (1-5) 3	0.84 (0.77-0.91)	0.85 (0.74-0.92)
13 Flip cards	3 (1-5)	3 (1-5) 4	0.74 (0.63-0.86)	0.66 (0.45-0.81)
15 Turn key in lock	3 (1-5)	3 (1-5) 2	0.73 (0.56-0.91)	0.74 (0.56-0.85)
16 Fold towel	3 (1-5)	3 (1-5) 4	0.80 (0.71-0.88)	0.82 (0.69-0.90)
17 Lift basket	3 (0-5)	3.5 (1-5) 3	0.76 (0.58-0.93)	0.79 (0.64-0.88)

4.4 Internal consistency

Internal consistency of WMFT was high, with a Cronbach's alpha of 0.91 for performance time. Corrected inter- item correlations was above 0.50 for all test items except items "hand to table" (0.46) and "reach and retrieve" (0.30), indicating that most items measure the same construct. Cronbach's alpha would not increase if the items "forearm to table", "hand to table" and "reach and retrieve" were deleted.

For the Log10-transformed performance time, Cronbach's alpha was 0.93, with inter item correlations all above 0.56, except for "reach and retrieve" that had an inter item correlation of 0.43. Cronbach's alpha would be reduced if any of the items were deleted, with the exception of "reach and retrieve", which would remain the same.

Cronbach's alpha for the FAS was 0.94, with inter item correlations above 0.50 for all items. The exception being "reach and retrieve" (0.43), but deletion of this items did not increase Cronbach's alpha.

5 Discussion

5.1 Summing of results

- Very good agreement and consistency for total score performance time (ICC>0.90 for all three rater models).
- Adequate agreement and consistency for most of the individual items for performance time.
- Adequate agreement and consistency for total score FAS, except for rater 1 vs other raters.
- Adequate agreement and consistency for approximately half of the individual items for the FAS.
- Very high internal consistency for both performance time and the FAS, with high inter-item correlations.

5.2 Results Total scores:

This study showed a high agreement and consistency with ICC> .90 on total scores performance time for all rater models. This was also the case when data was log10-transformed. Confidence intervals were also good for all models, except for the comparison between rater 1 and rater 2 with log10 transformed data. One reason for the very high reliability on performance time could be that the scoring was varied, which is illustrated by the large standard deviation (table 2). Polit and Beck states that samples with homogenous scores will give low reliability coefficients as instruments measures differences between the individuals being measured (6). Performance time was scored from below 1 second to 121 seconds, giving a very heterogeneous sample.

There was little difference between the models when using ICC, but SEM was considerably lower for the comparison of rater 1 vs rater 2 (1.96) than when rater 1 was compared to all raters (6.03) and other raters (7.06). The SEM is displayed in the original unit, which is seconds, so it becomes obvious that rater 1 vs rater 2 had less measurement error than the other comparisons. It can be debated what is an acceptable level of SEM, but as mentioned the difference between a subjects measurement and the true value is expected to be less than 1.96 S_w for 95 % of observations (31). Having this in mind we see that the confidence interval around the mean will become very wide when SEM is as large as 6.03 and 7.06 seconds.

Agreement, consistency and measurement error did not vary much when using either mean or median of performance time total scores, even though median often is recommended to avoid outliers (5). Median is also chosen as the preferred measurement of central tendency in the NORCIMT-study (appendix 1), but this did not seem to influence the reliability of this study.

Agreement and consistency for the FAS total scores were adequate, except for agreement of rater 1 vs other raters comparison. Both $ICC_{agreement}$ and $ICC_{consistency}$ were lower for the FAS than for performance time, with weighted kappa values being similar to the $ICC_{agreement}$. The FAS has 6 categories, ranging from 0-5, and scoring of the FAS is thoroughly specified with both general guidelines, and specific instructions for each item. Weighted kappa counts scoring in adjacent categories with partial agreement and it should therefore be expected that raters either agree on scoring or score in adjacent categories the most of the time. The lower values of ICC and weighted kappa could indicate that scoring of functional ability was particularly challenging for the raters. Rater 1 vs rater 2 had the highest amount of agreement on the FAS, so it would seem like calibration eases the scoring.

When analyzed separately we see that the different test items in general gave an adequate level of ICC on performance time. Still, some items had below adequate levels. When investigating the ratings of the independent test items it is clear that several were affected by outliers, where raters had a large disagreement. However, this was less apparent for the calibrated rater 1 and rater 2.

For the individual items of the FAS, between 5 to 9 of 15 items had an adequate $ICC_{agreement}$, with rater 1 vs rater 2 having the highest amount of adequate items. Weighted kappa for the individual items gave better results than ICC, if one considers a substantial amount (>0.61) of agreement as adequate. Still rater 1 vs rater 2 gave the best results, but had wide confidence intervals. For example a weighted kappa of 0.77 gave a confidence interval ranging from 0.51 to 1.00, which must be considered as wide.
Disagreements could occur because raters interpreted the procedure for analysis differently or some irregularities in the videos that affected the ratings. This is further adressed in the chapter “improvement in training, equipment and quality procedures”.

Internal consistency, measured with Cronbachs alpha, was high for both performance time and Functional Ability Scale, with high inter item correlations, indicating that all items of the WMFT measures the same construct. The only items having lower than 0.50 of Cronbach's alpha on performance time was "forearm to box (0.49)", "hand to table (0.46)" and "reach and retrieve (0.30)". For the FAS, Cronbachs alpha was 0.94 with inter item correlations above 0.50, except for "reach and retrieve" (0.43). Even though some items had a low inter item correlation, deletion of any of the items only influenced the internal consistency minimally, both for performance time and FAS. It is therefore difficult to recommend deleting any items based on the inter-item correlations.

The Log10-transformed performance time had slightly higher Cronbach's alpha than the untransformed performance time. This is probably due to that the transformation minimizes outliers and unequal variances, making data more homogeneous (25).
The high Cronbach`s alpha gives a good indication that all the items of the WMFT measures the same construct. However, a high Cronbachs alpha is desirable, but when it becomes higher than 0.90 it could be a sign of asking the same question several times (30). With a Cronbachs alpha of 0.91 for performance time and 0.94 for the FAS one could question if the number of items could be reduced. If so, workloads on testers, raters and patients tested could be reduced. The large number of items also contributes to a higher reliability, and could therefore give better results than it would have with fewer items.

5.3 Improvement in training, equipment and quality procedures

The effect of outliers and suggestions for minimizing outliers are presented in this section. Outliers are presented for each particular item with suggestions for improvement in the procedure for assessment (appendix 2). This mainly revolves around scoring of performance time, as outliers appeared most frequent here. Training of raters is also discussed and evaluated.

5.3.1 Procedures for assessment

For rater 1 vs all raters and rater 1 vs other raters the items "forearm to table", "forearm to box" and "reach and retrieve" had less than adequate levels of $ICC_{agreement}$, but also "lift can",

“lift pencil”, “lift paperclip”, “stack checkers”, “flip cards”, “turn key in lock” and “fold towel” were affected by outliers, though they still gained an acceptable level of $ICC_{agreement}$. When looking at Rater 1 vs all raters we see that “forearm to table” was scored by one rater with 0.7 seconds and 4 on the FAS for one subject, while the other rater scored it with 121 seconds and 4 on the FAS. It appears very unlikely to score the task “not completed”, and still score it with 4 on the FAS. Further inspection of the video gave no answer to the score of 121 seconds, so one explanation may be that it is a case of mistyping of some kind. When removing the outlier $ICC_{agreement}$ increases, but is still well below adequate level, so this particular case did not have a significant impact on the reliability of the item.

In “Forearm to box” one subject had been rated with 0.58 seconds and 5 on the FAS by one rater, while the other gave it 121 seconds and 4 on the FAS. This was the same subject as mentioned on “forearm to table”, and it seems like a mistyping of some kind, as the performance time and FAS score are contradictory. When removing the outlier, $ICC_{agreement}$ increased to an adequate level of 0.88 for this item.

The item “extend elbow” was rated with 121 seconds by one rater, while the other rater scored it with 61, 9 seconds on one subject. Another subject was scored with 121 seconds by one rater, with the other giving it 1,20 seconds. The procedure for this item says that it is considered completed when the thumb crosses the specified line on the table. When investigating the videos for the two subjects one can observe that the table template is printed on transparent paper, making the marking lines for completion difficult to see on the video. This is a probable cause of disagreement, as this also was the case for “Extend elbow with weight”. One rater gave the score of 121 seconds, while the other rater timed it to 66, 7 seconds. Raters have probably disagreed whether the hand and weight crosses the line as this is the criteria for completion of the task as specified in the procedure. Marking lines printed on transparent paper was less visible than on white paper, which could be a source of disagreement between raters. I therefore suggest that effort should be taken to ensure the quality of the table template. One simple way of solving this is using white paper instead of transparent to print the marking lines on.

“Reach and retrieve” had a very low $ICC_{agreement}$ due to one subject was given the time score of 7.52 seconds by one rater and 84.80 by the other rater. When looking at the video one can

see that the task is done with 3 attempts where the tester also explains the task for the subject after the first two attempts. One rater has scored the third attempt, while the other rater has considered all 3 attempts as one attempt and therefore gave it a higher time score. The procedure for assessment states that timing should be started when first movement is observed after the command "go!" but gives no clear guidelines of how to score this case. This could be further specified in the procedure by clarifying that only one attempt should be scored when the subject stops the task and the tester gives new instructions, even though the commands "ready, set, go!" is not used. Timing should then be started at the first observed movement in the shoulder, elbow or hand in the affected extremity.
One reason for subjects using several attempts could be that they didn't understand the task. The procedure for assessment (appendix 2) states that testing both arms could be advantageous to ensure that the subject has understood the test and test procedure, especially if subjects have cognitive deficits after stroke. The less affected arm should be tested first by going through the complete test, and then testing the affected arm. The possibility of inter-manual transfer of motor training is present but is outweighed by reducing the chance of misunderstanding and cognitive tentativeness from affecting the testing of the affected arm (appendix 2). It is not clear whether this has been done since the videos shows no tasks performed with the less affected arm, but by making this a mandatory part of the procedure for assessment rather than a recommendation, one could decrease level of confusion in the subjects.

"Lift can" has three major cases of disagreement on timing. One rater gave the scores of 42.32, 121 and 7.87 seconds, while the other rater gave the scores 121, 17.40 and 121 seconds. The procedure states that timing should be stopped when the can is approximately 2,5 cm from the patient's mouth. As this is subjectively measured with the raters visually inspecting the video, it is likely that raters have disagreed whether the can was close enough to the mouth, and therefore considered it "not completed". One way of minimizing disagreement between raters could be to add a picture in the procedure of a subject holding the can in the position when the task should be considered completed.

"Lift pencil" had three cases where one rater considered the task "not completed" with 121 seconds and the other rater gave the scores 16.88, 7.87 and 1.23 seconds. The procedure for this task states that timing is stopped when the whole pencil is lifted from the table using a 3-

jaw chuck grasp. The task should be considered “not completed” if the task is done incorrect, and this could be the cause of disagreement. It is unclear whether using a different grip than 3-jaw chuck is considered incorrect to the degree that task should be considered “not completed”. In the scoring procedure for this task it is written that use of other grip than 3-jaw chuck should be scored with maximum 2 on the FAS, but at the same time the task description states that 3-jaw chuck is the grasp to be used. This somewhat contradictory information could have caused one rater to consider the task “not completed”. The procedure should therefore specify more clearly when to score the task “not completed”, for example by stating that only 3-jaw chuck grip is accepted.

“Flip cards” was given 65.45 seconds and 121 seconds by the two raters. The procedure states that cards should be drawn towards the end of the table and then flipped through the long side with pincer grasp. In this particular video it is unclear whether a pincer grasp is used, and not all cards are flipped through the longside. The procedure is unspecific to whether this qualifies for the task to be rated “not completed”, but with rater 1 scoring over 65 seconds this indicates that performance was very slow. A clearer instruction for scoring “not completed” is needed to avoid this problem. One suggestion is to state that if pincer grasp is not used then task should be considered “not completed” and scored with 121 seconds. If other grasps than pincer is to be accepted it should be defined how to score this.

“Turn key in lock” was scored with 40.32 seconds by one rater, while the other rater scored 121 seconds. The procedure states that the patient should use a lateral pincer grasp for turning the key from a vertical position through a 180 degree-arc, with 90 degrees to each side. This particular task is filmed from the opposite side of the hand tested, but it is difficult to see if the patient moves the key through the whole 180 degree-arc. The procedure for assessment states that if the patient doesn’t turn the key in the correct sequence, maximal score for functional ability is 3, but lacks information about scoring when 180 degree- arc is not used. One way of clarifying this is by stating that if 180 degree-arc is not used then maximal score for functional ability is 2 or 3. Another way is to state that task is to be considered “not completed” if full 180 degree-arc is not used.

“Fold towel” was considered “not completed” and given 121 seconds by one rater, and 7 seconds by the other rater. The procedure for this item specifies that edges of the towel should

lie approximately upon each other, but within 4 cm apart would be tolerated. If the edges are within this limit of tolerance is hard to assess using video, and is a subjective decision. One rater has probably considered the towel edges to be more than 4 cm apart, while the other rater considered it less than 4 cm. Adding example pictures of accepted and non-accepted folding to the procedure of assessment would give clearer instructions of whether to score "completed" or "not completed", which could reduce disagreement.

5.3.2 Training of raters

One significant question is if training of raters is worthwhile. Wolf Motor Function Test does not require training, so should we invest time and resources on this? With all raters in this study receiving 4 days of coursing, as well as 2 days of training in the treatment, a good amount of resources was spent on this. In addition rater 1 and rater 2 was calibrated against each other by analyzing 3 videos and discussing scoring over telephone. This can seem costly and time consuming, but the level of reliability gained in this study was very good, and could be an indication that training was beneficial. Agreement on performance time was very good for the whole sample, but FAS agreement was best for the calibrated raters, suggesting that calibration is beneficial for scoring quality of movement.

Calibration of rater 1 and rater 2 was done by scoring 3 videos and discussing them afterwards. The level of agreement between the raters gives an indication that calibration of raters was beneficial without demanding much more effort than for the other raters. The question is whether training for all raters could be reduced to a lower amount and still gets the same level of agreement, especially if some of the coursing is replaced by calibration like rater 1 and rater 2 received. This question is hard to answer based on this study, but should perhaps be investigated in future studies.

5.4 Statistical analysis

To what degree different statistical approaches are appropriate for different situations is a much debated subject, and therefore needs to be discussed. ICC is much used in reliability studies, but there are several models to choose from, all suited for different situations (26, 35). When deciding to use ICC a central question was which model was best suited for this study. These can all give different results when applied to the same data, and is suited for different situations and study designs. Many researchers are not aware of the difference between the

forms, and even those aware of the difference rarely reports which form is used (35). In this study a two-way mixed ICC model was used, using single measures and an absolute agreement definition of ICC, while consistency was investigated for comparison with previous studies. A two-way model ICC was therefore considered appropriate because a systematic source of variance can be expected from the judges, and this model controls for systematic error (35). The consistency model deems column variance as an irrelevant source of variance, and it is therefore excluded, while in the absolute agreement model it is not (26). De Vet recommends using an absolute agreement model, claiming that a consistency model rarely is desirable in the medicine field (5). Paired scores between raters would result in perfect agreement in the consistency version (26), and we could encounter a scenario where rater 1 scored performance time consistently 1 second higher than rater 2. Obviously there is a systematic difference that would not be taken into account when applying the consistency definition. I therefore argue that the absolute agreement definition is the appropriate for this kind of study.

Weighted kappa with quadratic weights was applied for the FAS. There was some uncertainty of what type of weights were the appropriate to apply, which was not made clearer by the literature. For example, Viera states that the choice of weights is a subjective matter which experts may disagree on in different settings (27). It is also claimed that the quadratic weights are the most common in use (5, 28), not stating specifically why this is the case. That something is common to use is also hard to consider a solid argument for choosing a particular statistical method. The coefficients of quadratic weights often give equivalent values to the intraclass coefficient is one argument often used in the literature (29), and Sim even claims that to apply kappa without weights for ordinal data is inappropriate (28). Unweighted kappa also tends to decrease the more categories we have, while weighted kappa increases when we add more categories (29).

There is also the question of whether one should use weighting in the first place. Obviously weighting will give higher kappa values than unweighted kappa, because it considers scorings in adjacent categories as a partial agreement. For a scale like the FAS which have 6 categories, ranging from 0 to 5, we could expect raters to score in adjacent categories more often than not, for example if one rater scores 3 and another assigns the score of 2 or 4. One could wonder if weighting in this instance is an artificial way of inflating the agreement to an

acceptable level. When using unweighted kappa, it becomes clear that it can be challenging to come to exact agreement on the Functional Ability Scale. For example test item 10 had a weighted kappa of 0.86 when comparing rater 1 to all raters, but only 0.50 with unweighted kappa. It also appears unclear what level of kappa is acceptable when applying weights. I have referred to the suggestions of Landis & Koch (5, 27) for interpretation of kappa values. These are presented for unweighted kappa, but it is not specified what levels are acceptable for weighted kappa. If these guidelines are applicable to weighted kappa is therefore unclear, but they were the only guidelines available that I have been able to find.

SEM was assessed and presented for total scores and individual test items performance time as a parameter of measurement error. The SEM is shown in seconds, and therefore gives an easy understandable picture of the size of measurement error. Cronbach also claimed that the measurement error is the most important information about a measurement (33), and it is therefore calculated in my study. The SEM clearly showed its use when comparing the three rater models performance time total scores. ICC was almost similar in all three models, but SEM was significantly lower for the rater 1 vs rater 2 model, showing that there was less measurement error.

The literature does not state whether ICC requires a normal distribution, at least I have not been able to find such guidelines. Nor have I seen transformations in use for reliability studies on the WMFT. Because the performance time of WMFT does not fulfill the assumption of normality, I performed a transformation on the data to see if this improved the distribution and if it affected the reliability. The use of transformations is much debated, but they can correct data that violates the assumptions of normality, especially in heavy-tailed distributions (25). Field describes a log-transformation as well suited for correcting a positive skew in data (25), but I have not found the literature to give clear specifications for when and where to use the different transformations.

The log-10 transformation did not yield a perfect normal distribution, but improved the positive skew (appendix 3). It affected the reliability in the three models very little, making it hard to recommend or discourage the use of this transformation. It could be that other transformations would have improved the skew more and thereby affected the reliability in another way, but this has not been investigated in this study.

The sample size can have a large impact on the results from statistical tests, and therefore is of significance for evaluating the results for statistical analysis. One interesting question when designing a reliability study is how many subjects we need to obtain an acceptable confidence interval. If the lower bound of the CI is far below the cut off for adequate ICC, in this study 0.75, then a 95 % CI gives few assurances that other samples will give adequate ICC (5).
In general it seems that the total scores of WMFT were less affected by sample size, while individual items clearly needed a larger sample size, at least when ICC goes below 0.90. Although the rater 1 vs rater 2 model gained a high level of agreement, the confidence interval on the individual items was very wide. It therefore appears that a sample size of 12 is too small to give solid conclusions about measurements properties, at least when analysis is only based on 2 raters.
This is of importance when planning reliability studies, because the amount of patients and raters affects the cost and amount of work required for the study. If one can get acceptable confidence intervals with less patients and raters, then it would be easier to conduct these studies. One way to give better confidence intervals would have been to let more raters score the videos. If three raters had scored all videos a lower sample size would be required. For two raters the sample size required to obtain a 95% Confidence interval of ±0.1 around an ICC of 0.8 is 50 subjects. With three raters the sample size required to obtain the same confidence interval would be 35 (5).
One improvement of the study therefore would have been to increase the amount of raters.

The timing of the performance time also presents some challenges for both statistical analyses, as well as observing change in more impaired subjects. This study had many scores of "not completed", where patients were given the score of 121, making the distribution skewed. Hodics et.al writes that patients that are unable to complete at least half the tasks will suffer from a floor effect and receive a median score of 120 seconds, regardless of how well they performed in the other tasks. Use of the mean will also suffer through a large amount of skewing (36). They therefore propose that WMFT measurements could be calculated as rate of performance, where how many times they perform a task in 60 seconds is used. If a patient is unable to perform a task, then he/she will be given the score of 0, which will make the distribution more normal (36). This could be one way to improve the distribution in this study, both for minimizing the floor effect and gaining a more normal distribution. It should therefore be considered for future studies.

5.5 Internal validity

Internal validity will be evaluated and discussed using the threats outlined by Cook and Campbell (4), which was presented in the theoretical background .

As previously mentioned the threats *history, maturation* and *testing* could influence internal validity in a negative way because participants can change due to other factors than the intervention we apply. However, these propose less of a threat in this interrater reliability study design because all participants were assessed only one time. This was also performed at baseline, before training started.

Instrumentation is important to consider as inter- rater reliability focus on the degree two or more raters agree on scoring, in this case WMFT assessments on video. All raters had training in using the Norwegian WMFT manual for scoring videos. Two raters also were calibrated against each other through discussing scoring of videos. This would contribute to the internal validity of the study, as training of raters is considered an effective way of enhancing internal validity (5). The calibration of rater 1 and rater 2 also gives an opportunity to see if calibration improves reliability in this study.

Assignment of participants to the study was done through strict inclusion and exclusion criteria, which is strength for internal validity. However, when looking at the demographics (table 1), we see that there were differences between the three rater models. The rater 1 vs rater 2 model had 12 patients where all scored 0 on the NIHSS, indicating a high level of functioning. When keeping in mind that rater 1 and rater 2 had the most training, we could get an interaction effect between *assignment* and *instrumentation*. One could hypothesize that high functioning patients are easier to score, which in combination with the extra calibration would give a higher level of agreement. This could be a weakness to internal validity.

5.6 External validity

External validity is concerned with what degree the results can be generalized (5).

Selection of patients was done using the inclusion and exclusion criteria described in the procedures for assessment (appendix 2). Strict inclusion and exclusion criteria strengthen internal validity, making sure that participants have similar characteristics, but could reduce the possibility for generalizing the results to a wider population (4). The mean score of 2.6 on the MRS and the fact that of the total sample, 90.2 % had a score of 0 (63.4%) or 1 (26.8%)

on the NIH stroke scale (NIHSS) of the affected arm, indicates that participants in average had light to moderate body function impairments (appendix 2). Stroke survivors constitute a very heterogeneous group, and often have a high level of disability with lower level of functioning (2). The high level of functioning in patients in this study makes it difficult to generalize to a broader population of stroke survivors. Still, generalizability to a similar population is possible.

This reliability study had a relatively large sample size with 41 participants. This is quite close to 50, which is the recommended number proposed by De Vet (5). A large sample size is a requirement for obtaining an acceptable confidence interval around the estimation, especially when ICC values are low (5). The sample size contributes to the generalizability, at least for the full sample. For the subsample of rater 1 vs rater 2 (n=12) confidence intervals became very wide for the individual items, and it is therefore difficult to generalize these results. It can also be mentioned that the sample size for this study was quite large compared to reliability studies on the American WMFT (9, 14).

Only two raters scored all videos, with all of them receiving extensive training in scoring. Generalizing results to other settings could therefore be difficult, since few health professionals in the clinical field have been trained to use WMFT. Scoring was done by viewing videos of assessments, and procedures for videotaping and assessments were described thoroughly, which makes it possible for other researchers to replicate the setting used.

Results from this study can be generalized to similar groups of hemiparetic stroke survivors in the early phase after stroke. To what degree these results apply to clinical practice is therefore still uncertain.

Traditionally internal validity has been given the highest priority, using the argument that if we cannot show enough evidence that an intervention really works, there is no point generalizing the results (6). This can be at odds with the emphasis on evidence- based practice, because if study results cannot be generalized to real-world clinical settings, then what does it matter if the study has strong internal validity (6)? It is therefore important to evaluate both internal and external validity as they both are important for the quality of the study. They also affect each other, as sometimes a high level of internal validity decreases the generalizability and vice versa.

5.7 Clinical implications

Transferring a measurement used in research to clinical practice can obviously have some implications, both practical and philosophical. In this study the WMFT has shown adequate to high level of reliability, and in American version shown good psychometric properties. It therefore appears to be a measurement to be relied on, and is also recommended by the directorate of health for assessing upper extremity function after stroke (2). If it is to be implemented as a standard approach for assessment, then the clinical practice field has to accept it. One challenge with the WMFT is the equipment required, which among other objects, includes a table, a box and a basket (appendix 2). This has to be stored somewhere or should preferably stand ready for testing at all times for practicality. Otherwise it would be very time consuming to set up all equipment in the correct positions each time testing is to be performed. Videotaping, which was used in this study, also complicates testing, due to ethical issues regarding filming of patients. However, WMFT has been shown to be both reliable and valid without videotaping (17). WMFT is also correlated with amount of use with the affected upper extremity (1), which has been known to provide an objective, real world index of more-impaired arm activity (37). The WMFT, as a reliable and valid measure in American version, and now reliable in Norwegian translation, could be a useful measure of upper extremity motor function in the clinical practice field, if properly integrated.

6 Conclusion

The purpose of this study was to investigate the inter-rater reliability of the Norwegian translation of the Wolf Motor Function Test used in hemiparetic stroke survivors in the early phase after stroke. The study suggests that the Norwegian translation of the WMFT is a reliable measure when applied to hemiparetic stroke patients in the early phase after stroke.

The study also shows that calibrating raters against each other can be beneficial, even though it is difficult to conclude based on the low sample size.

Inter-rater reliability was close to the levels of the American version of the WMFT, and is considered sufficient for clinical use.

7 References

1. Lang CE, Wagner JM, Edwards DF, Dromerick AW. Upper extremity use in people with hemiparesis in the first few weeks after stroke. J Neurol Phys Ther. 2007 Jun;31(2):56-63. PubMed PMID: 17558358.
2. Helsedirektoratet. Nasjonal retningslinje for behandling og rehabilitering ved hjerneslag. 2010.
3. Edmans J. Occupational therapy and stroke. 2nd ed. Chichester: Wiley-Blackwell; 2010.
4. Carter R, Lubinsky J, Domholdt E. Rehabilitation research: principles and applications. 4th ed. St.Louis, Missouri: Elsevier Saunders; 2011.
5. De Vet H, Terwee C, Mokkink L, Knol D. Measurement in Medicine. Cambridge: Cambridge University Press; 2011.
6. Polit DFB, Cheryl T. . Nursing research: generating and assessing evidence for nursing practice 9th ed. Philadelphia, Pa: Wolters Kluwer/Lippincott Williams & Wilkins; 2012.
7. Helsedirektoratet. Internasjonal klassifikasjon av funksjon, funksjonshemming og helse. In: Sosial- og Helsedirektoratet, editor. 2 ed. Trondheim: Aktietrykkeriet2006.
8. Wolf SL, Thompson PA, Morris DM, Rose DK, Winstein CJ, Taub E, et al. The EXCITE trial: attributes of the Wolf Motor Function Test in patients with subacute stroke. Neurorehabil Neural Repair. 2005 Sep;19(3):194-205. PubMed PMID: 16093410.
9. Wolf S, Catlin P, Ellis M, Archer A, Morgan B, Piacentino A. Assessing Wolf Motor Function Test as Outcome Measure for Research in Patients After Stroke. Stroke. 2001;32(7):1635-9.
10. Fritz SL, Blanton S, Uswatte G, Taub E, Wolf SL. Minimal detectable change scores for the Wolf Motor Function Test. Neurorehabil Neural Repair. 2009 Sep;23(7):662-7. PubMed PMID: 19498013.
11. Askim T, Dahl A, Stock R, Langøren E. Wolf Motor Function Test på norsk. Fysioterapeuten. 2009.
12. Wolf SL, Winstein CJ, Miller JP, Taub E, Uswatte G, Morris D, et al. Effect of Constraint-Induced Movement Therapy on Upper Extremity Function 3 to 9 Months After Stroke The EXCITE Randomized Clinical Trial. JAMA. 2006;296(17):2095-104.
13. Wolf S, Newton H, Maddya D, Blanton S, Zhang Q, Winstein C, et al. The Excite Trial: Relationship of intensity of constraint induced movement therapy to improvement in the wolf motor function test. Restorative Neurology and Neuroscience. 2007;25:13.
14. Morris DM, Uswatte G, Crago JE, Cook EW, 3rd, Taub E. The reliability of the wolf motor function test for assessing upper extremity function after stroke. Arch Phys Med Rehabil. 2001 Jun;82(6):750-5. PubMed PMID: 11387578.
15. Fjærtoft H, Indredavik B. Kostnadsvurderinger ved hjerneslag. 2007.
16. Pereira N, Michaelsen S, Menezes I, Ovando A, Lima R, S. T. Reliability of the brazilian version of the Wolf Motor Function Test in adults with hemiparesis. Revista Brasileira de Fisioterapia. 2011;15(3):257-65.
17. Whittall J, Savin D, Harris-Love M, Waller S. Psychometric Properties of a Modified Wolf Motor Function Test for People With Mild and Moderate Upper-Extremity Hemiparesis. Archives of Physical Medicine and Rehabilitation. 2006;87(5):656–60.
18. Thornquist E. Vitenskapsfilosofi og vitenskapsteori for helsefag. Bergen: Fagbokforlaget; 2003.
19. Cieza A, Geyh S, Chatterji S, Kostanjsek N, Ustun B, Stucki G. ICF linking rules: an update based on lessons learned. Journal of rehabilitation medicine : official journal of the

UEMS European Board of Physical and Rehabilitation Medicine. 2005 Jul;37(4):212-8. PubMed PMID: 16024476.
20. Bjørndal A, Hofoss D. Statistikk for helse- og sosialfagene. 2.utgave, 3.opplag ed. Oslo: Gyldendal Akademisk; 2004. 269 p.
21. Salter K, Jutai JW, Teasell R, Foley NC, Bitensky J, Bayley M. Issues for selection of outcome measures in stroke rehabilitation: ICF Participation. Disability and rehabilitation. 2005 May 6;27(9):507-28. PubMed PMID: 16040555.
22. Bland MJ, Altman DG. Statistics notes Chronbachs alpha. BMJ. 1997;314.
23. Thrane G, Emaus N, Askim T, Anke A. Arm use in patients with subacute stroke monitored by accelerometry: association with motor impairment and influence on self-dependence. Journal of rehabilitation medicine : official journal of the UEMS European Board of Physical and Rehabilitation Medicine. 2011 Mar;43(4):299-304. PubMed PMID: 21347506.
24. Thrane G. Prosjektprotokoll: Intensiv trening av arm- og håndfunksjon hos personer med hjerneslag – en randomisert kontrollert multisenterstudie. 2009.
25. Field A. Discovering statistics using SPSS Los Angeles: SAGE; 2009.
26. McGraw K, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychological Methods. 1996;1(1):16.
27. Viera A, Garrett J. Understanding Interobserver Agreement: The Kappa Statistic. Family Medicine. 2005;37(5):3.
28. Sim J, Wright C. The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. PHYS THER. 2005;85:11.
29. Brenner H, Kliebsch U. Dependence of weighted kappa coefficients. Epidemiology. 1996;7(2):3.
30. Streiner DL. Starting at the beginning: an introduction to coefficient alpha and internal consistency. J Pers Assess. 2003 Feb;80(1):99-103. PubMed PMID: 12584072.
31. Bland MJ, Altman DG. Statistical notes measurement error. BMJ. 1996;312.
32. Weir JP. Quantifying Test-Retest reliability using the Intraclass Correlation Coefficient and the SEM. Journal of Strength and Conditioning Research, 2005, 19(1), 231–240. 2005;19(1):231–40.
33. Cronbach LJ, Shavelson RJ. My Current Thoughts on Coefficient Alpha and Successor Procedures. Educational and Psychological Measurement. 2004;64(3):391-418.
34. Landis R, Koch G. The Measurement of Observer Agreement for Categorical Data. Biometrics. 1977;33(1):17.
35. Shrout P, Fleiss J. Intraclass Correlations: Uses in Assessing Rater Reliability. Psychological Bulletin. 1979;86(2):8.
36. Hodics T, Nakatsuka K, Upreti B, Alex A, Smith P, Pezzullo J. Wolf Motor Function Test for Characterizing Moderate to Severe Hemiparesis in Stroke Patients. Archives of Physical Medicine and Rehabilitation. 2012;93(1):1963–7.
37. Uswatte G, Giuliani C, Winstein C, Zeringue A, Hobbs L, Wolf SL. Validity of accelerometry for monitoring real-world arm activity in patients with subacute stroke: evidence from the extremity constraint-induced therapy evaluation trial. Arch Phys Med Rehabil. 2006 Oct;87(10):1340-5. PubMed PMID: 17023243.

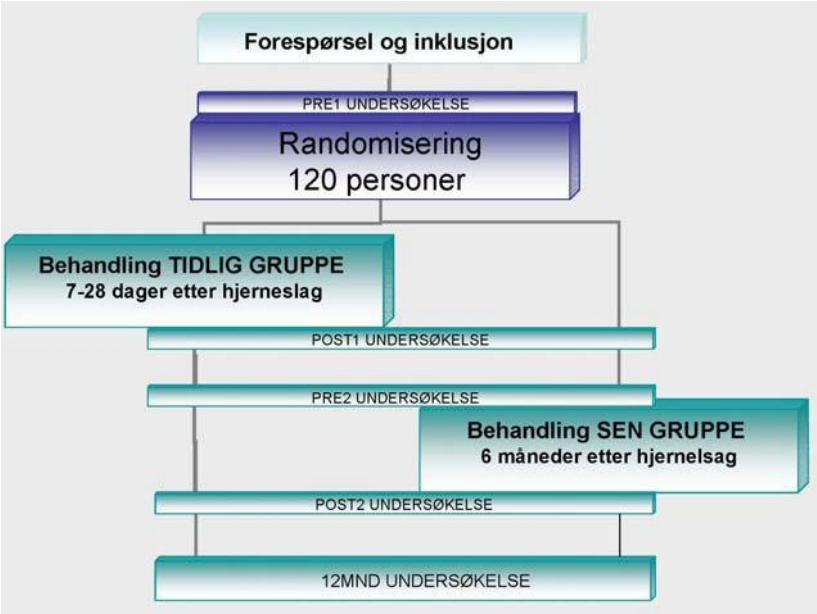
List of appendices

Appendix 1: Protocol of the NORCIMT study

Prosjektprotokoll:

Intensiv trening av arm- og håndfunksjon hos personer med hjerneslag

– en randomisert kontrollert multisenterstudie.



UNIVERSITETSSYKEHUSET NORD-NORGE
DAVVI-NORGGA UNIVERSITEHTABUOHCCEVIESSU

ST. OLAVS HOSPITAL
UNIVERSITETSSYKEHUSET I TRONDHEIM

UNIVERSITETET I TROMSØ

NTNU
Det skapende universitet

Bakgrunn

Hjerneslag er en svært vanlig sykdom i industrialiserte land med 250 tilfeller pr. 100 000 innbyggere pr år[1]. Hver 5. person vil få et hjerneslag i løpet av livet, og dette er den

hyppigste årsak til alvorlig funksjonshemning. Etter rehabilitering vil en høy andel av pasientene fortsatt ha motoriske funksjonsutfall (50-60%)[2]. I en dansk studie forble 21 % uten funksjonell aktivitet i begge armer, og armen forble uten nytte hos 56% av pasientene med alvorlige pareser i den initiale fasen[3]. Motorisk funksjon har vist seg å ha en klar sammenheng med senere evne til å utføre daglige aktiviteter, grad av selvhjulpenhet og senere livstilfredshet [4].

Constraint Induced Movement Therapy (CI-terapi)

Constraint Induced Movement Therapy er en metode for å trene opp arm og håndfunksjon etter et hjerneslag. Metoden kjennetegnes ved 1) intensiv trening av den mest affiserte overekstremiteten opp til 6 timer pr. dag; 2) et adferdsterapeutisk opplegg for å fremme bruken av den mest affiserte armen i daglige aktiviteter; og 3) bruk av en vott på den friske hånden for å øke bruken av den mest affiserte siden[5, 6]. Behandlingen gjøres over 10 arbeidsdager. Metoden ble utviklet forkroniske pasienter mer en 12 måneder etter hjerneslag, og i denne gruppen fant Taub et al [5] effekt av behandlingen på kort og lang sikt. Taub [7] har senere sammenlignet CI-terapi med tilsvarende intensivt treningsprogram på 41 pasienter med signifikante forskjeller i favør CI gruppen. I tillegg er behandlingen testet ut i subakutt fase 3-9 måneder etter hjerneslag. En amerikansk multisenterstudie på 222 pasienter viste statistisk signifikante og klinisk relevante forbedringer i armfunksjon som vedvarte over 1 år [8].

Modifiserte former for CI-terapi

Også modifiserte versjoner av CI-terapi er prøvd ut i kronisk fase. I den klassiske CI-terapien gis den intensive treningen i en til en situasjoner, og legges opp etter en systematisk protokoll. Van der Lee et al [9] ga den intensive treningen som gruppebehandling, og fant en liten men klinisk og statistisk signifikant effekt i forhold til bilateral armtrening. En norsk studie med gruppebasert trening ga bedret motorisk funksjon på kort sikt, men ingen langsiktig effekt [10]. Det ser også ut til at 3 timers intensiv trening kan gi bedret motorisk funksjon [11, 12], men dette har ikke vært testet ut i større studier i kronisk fase.

CI-terapi i tidlig fase etter hjerneslag

Forskjellige former for CI-terapi er prøvd ut i tidlig fase. Seks timers intensiv trening i denne fasen kan være til ulempe for pasientene, mens to timers CI terapi ikke gir bedre resultater

enn annen behandling av tilsvarende intensitet (n=52) [13]. Tre timers trening gir en positiv trend i favør CI-gruppen (n=23) men ingen signifikante resultater er vist [14]. CI-behandling er belyst i flere systematiske oversiktsartikler [15, 16] og kommentarartikler [6, 17-20]. Det er utført få studier i tidlig fase, og de fleste av dem har lavt deltakerantall[13-15]. Effekten av å starte CI-behandling i tidlig fase er derfor usikker, og det trenges større studier med før man kan anbefale dette som standard behandling. Det etterlyses studier av CI-terapi som er mer forenlig med klinisk praksis [20].

Spesielt bør man undersøke mindre intensive metoder som er mer anvendelig i tidlig rehabilitering etter hjerneslag [19] og man bør også undersøke om pasienter med større motoriske utfall kan profitere på behandlingen [6]. Pasienter som er inkludert i klassiske CI-studier har minimum 10 graders ekstensjon i håndledd og fingre, og må ha balanse som gjør at de kan bevege seg sikkert innendørs [5, 7, 8]. Det er antydet at dette ikke vil gjelde mer enn 10% av pasienter med hjerneslag [17]. Krav til høy MMS-Score gjør at også pasienter med ekspressiv afasi utelukkes. Gjennomsnittlig alder i studier av CI terapi har variert fra 50,7 år 63,9 år med standardavvik fra 12 til 19 år[7-9, 13, 14]. Den lave alderen kan trolig tilskrives inklusjonskriteriene og fremtidige studier bør inkludere eldre pasienter og undersøke korrelasjonen mellom alder og treningseffekt. Det er ingen studier som har undersøkt om pasientene føler seg bedre eller blir mer tilfreds med livet etter CI behandling. I tillegg bør det undersøkes om pasienten oppnår sine egne mål gjennom deltakelse i et slikt intensivt opplegg [21]. I følge nevrobiologisk teori er hjernens evne til relæring størst i de første ukene etter et hjerneslag [17]. Denne teorien baserer seg på dyreforsøk og vi har pr dags dato ingen studier som sammenligner intensiv trening i tidlig fase etter hjerneslag med samme type behandling i senere fase hos mennesker.

Formål og problemstilling

Hovedformålet med denne studien er:

1) Å undersøke om CI-terapi med 3 timers intensiv trening vil gi bedre funksjon enn standard behandling hos pasienter i tidlig fase etter hjerneslag. Pasienter med større hjerneslag og dårligere motorisk funksjon inkluderes og funksjonen måles 6 måneder etter hjerneslaget.

Delmål er å :

2) Sammenligne effekten av tidlig intervensjon med modifisert CI terapi med en senere

intervensjon. Tidligintervensjongruppen skal starte behandling innen 28 dager etter hjerneslaget mens senintervensjonsgruppen behandles 6 måneder etter slaget.

3) Undersøke om treningseffekten er avhenging pasientens alder.

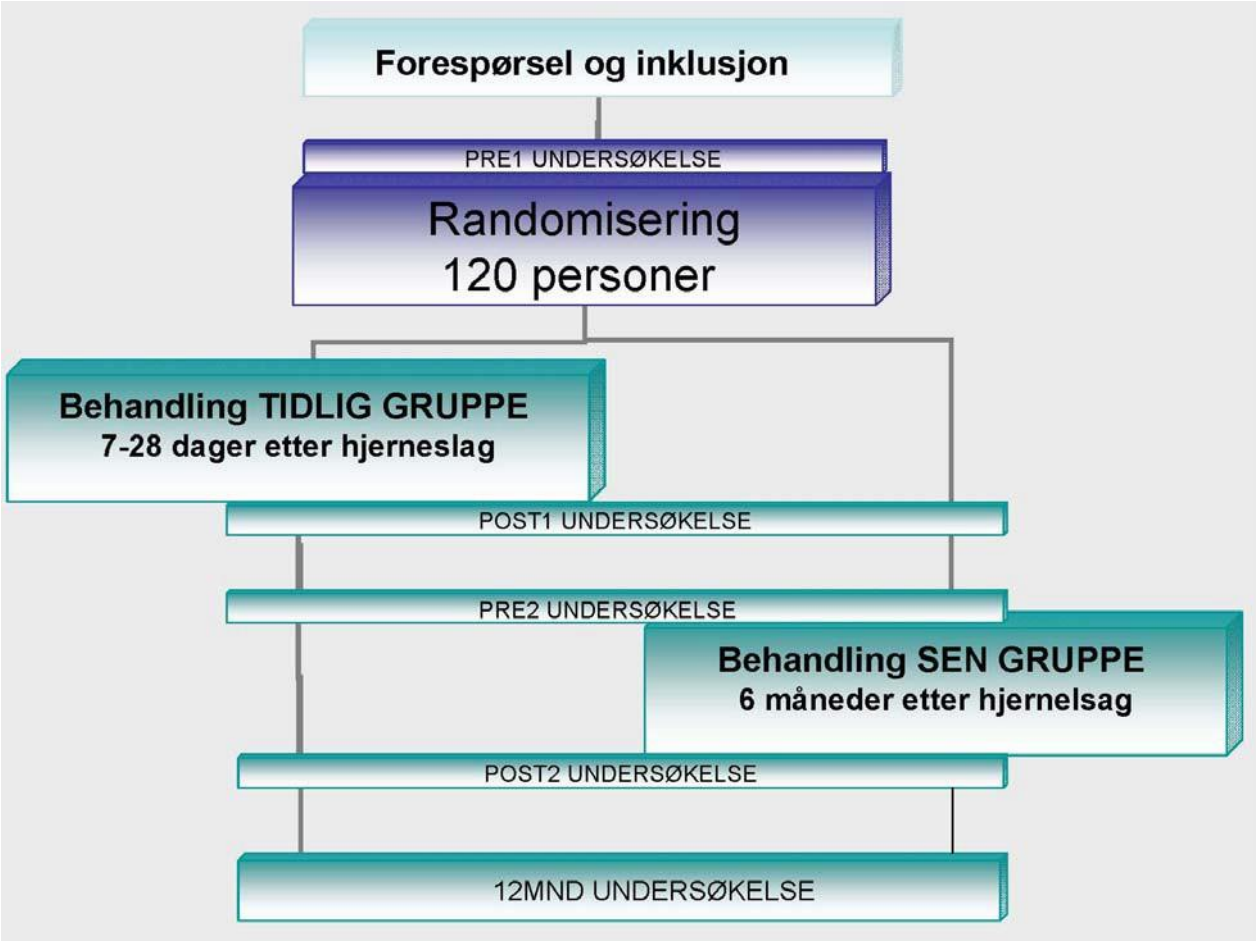
4) Undersøke hvordan behandlingseffekten korrelerer med grad av måloppnåelse.

5) Undersøke om bedringen i funksjon har betydning for pasientens livstilfredshet og oppfattelse av bedring etter hjerneslaget.

Metode

Design

Studien skal være en enkeltblindet randomisert kontrollert studie. Deltakerne blir inkludert så snart de tilfredsstiller inklusjonskriteriene. Etter evt. samtykke blir de randomisert til en tidligintervensjonsgruppe eller en senintervensjonsgruppe. Gruppen stratifiseres etter behandlingssenter i tre grupper (UNN / St. Olavs / Andre sykehus). Det brukes blokkrandomisering med 4, 6 eller 8 pasienter i hver blokk. Allokering foretas pr telefon av Klinisk Forskningsenter UNN. Figur 1 viser en skisse av den planlagte studien.



Figur 1: Flytskjema for studien

Deltakere

120 deltakere skal rekrutteres fra slagenhetene ved 6-8 norske sykehus fra september 2008 til desember 2009. Alle personer som mistenkes hjerneslag med affeksjon av overekstremitetene screenes for eventuell deltakelse. Inklusjonskriteriene går fram av tabell 1. Kriteriene er valgt fordi de er enkle å undersøke, hensiktsmessige i forhold til slagpopulasjonen, og fordi de er brukt i andre studier på subakutte slagpasienter[14, 22]. Sammenlignet med kriterier som er brukt i studier av kroniske slagpasienter tillater de inklusjon av pasienter med dårligere funksjon i overekstremitetetene og dårligere balanse.

Table 1.

Inklusjonskriterier:

- **Hjerneslag for mer enn 5 dager og mindre enn 26 dager siden**
  - Enten: Første hjerneslag
  - Eller: Andre hjerneslag uten vedvarende kraftsvikt i arm eller hånd etter første slag
- **Modified Rankin Scale 0-2 før innleggelse.**
- **Alder** over 18 år
- **Vedvarende unilateral parese i arm eller hånd**
  (SSS armmotorikk 2-5 eller SSS håndmotorikk 2-4)
- **Ekstensjonsbevegelse fingre eller håndledd:**
  - **Enten:** Kan løfte minst to fingre fra underlaget når hånden ligger pronert på et bord.
  - **Eller:** Kan ekstendere minst 10 grader i håndledd. Utgangsstilling underarm pronert, håndledd fullt flektert med støtte på underarm.
- **Evne til å følge en to stegs kommando**
- **MMS score større en 20 (eller større enn 16 kombinert med ekspressiv afasi).**

Eksklusjonskriterier:

- **Modified Rankin Scale 5 eller 6**
- **Kan ikke gi informert samtykke**
- **Stor hemispatial neglect** (Mer enn to cm på Line Bisection Test.)
- **Ikke forventet å overleve 1 år som følge av annen sykdom (for eksempel hjertesvikt eller malign sykdom)**
- **Skade i overekstremitet eller tilstand som førte til begrenset bruk av armen før hjerneslaget.**
- **Annen nevrologisk lidelse som påvirker motorisk funksjon.**

SSS = Scandinavian Stroke Scale, MMS = Mini Mental Status Examination

Intervensjon

Deltakerne i tidlig intervensjonsgruppen skal starte behandlingen mellom 7 og 28 dager etter hjerneslaget. De overføres til rehabiliteringsavdeling så snart de er inkludert i studien. De vil få tre timers CI terapi for den mest affiserte armen i 10 etterfølgende arbeidsdager.

Behandlingsopplegget har 3 hovedkomponenter. Disse er:
Målrettet trening for å forbedre funksjonsnivået i den dårligste armen. Deltakerne trener på målrettede aktiviteter som de har problemer med å gjennomføre. Aktivitetene består av bevegelsesoppgaver der vi registrer hvor lang tid det tar å gjennomføre den, eller teller hvor mange repetisjoner deltakeren klarer innen et visst tidsrom. Denne treningen settes opp etter en protokoll der vi gradvis tilnærmer oss den aktivitet vi ønsker at pasienten skal kunne gjennomføre (shaping). Det gis systematisk tilbakemelding om resultatet av gjennomføringen og på kvaliteten på bevegelsen. Vanskelighetsgraden økes etter hvert som pasienten blir bedre. I tillegg vil trenes det på en del større sammenhengende aktiviteter som skal gjøres over et lengre tidsrom (task practice).

1. Tiltak for å endre bevegelsesvaner i dagliglivet. Hensikten med dette er å få pasienten til å bevege dem dårligste armen mest mulig i løpet av de 12 dagene treningen varer. Følgende tiltak brukes for å få dette til:

- Avtale om bruk av den affiserte armen. Terapeut og pasient går igjennom hvilke aktiviteter pasienten skal bruke den dårligste armen til og i hvilke aktiviteter det ikke er trygt å bruke den dårligste armen. Det lages en detaljert dagsplan for når den dårligste armen skal brukes.
- Arbeidsoppgaver utenfor behandlingstiden. For å trene på å mestre nye oppgaver også utenfor behandlingstiden, setter vi opp et visst antall aktiviteter som pasienten skal prøve å gjøre med den dårligste armen før han kommer til neste behandling.

2. Bruk av behandlingsvott på den beste armen. For at pasienten skal huske å bruke den dårligst armen mest mulig skal han bruke en behandlingsvott på den friske armen som hindrer bruken av denne. Målet er at votten skal være på i 90% av våken tid i de 12 dagene behandlingen varer. Vi lager en detaljert plan for når votten skal brukes. I denne planen skriver vi opp aktiviteter der det kan være farlig å bruke votten, for eksempel fordi pasienten trenger den friske hånden til å holde deg fast med eller ved håndtering av varmt vann. Hver dag går vi igjennom og registrerer hvor mye votten er brukt siden forrige besøk.

Deltakerne får også fysioterapi, ergoterapi og andre intervensjoner som de vanligvis ville fått på slagenheten eller rehabiliteringsenheten. Deltakerne i senintervensjonsgruppen blir lagt inn på rehabiliteringsavdeling 6 måneder etter hjerneslaget. De mottar den samme behandlingen som tidligintervensjonsgruppen i 10 dager.

Standardisering av behandlingen

CI behandlingen skal gis av fysioterapeut og /eller ergoterapeut. Terapeutene som gir behandlingen skal enten ha vært på kurs i CI-behandling eller ha erfaring med behandlingsformen. I tillegg skal alle ha to dagers opplæring i behandlingsprotokollen til denne studien og gjennomgå en standardiseringsprosedyre. Standardiseringen går ut på at terapeuten tar videoopptak av behandlingen som sendes til prosjektkoordinator for gjennomgang. Video skal inneholde administrasjon av Motor Activity log, shaping øvelser, oppgavepraksis, avtale om bruk av vott, hjemmeoppgaver og hjemmedagbok. Videoen blir gjennomgått for å avdekke om kritiske elementer i protokollen blir

gjennomført riktig eller galt. Standardiseringsprosessen gjentas for hver terapeut, hvert halvår så lenge studien pågår.

Målinger

Alle deltakerne blir undersøkt etter inklusjon (PRE1), etter at tidliggruppen har fått sin behandling (POST1), før sen-gruppen får sin behandling (PRE2), etter at sengruppen har fått sin behandling (POST2) og 12 måneder etter hjerneslaget (12M). Undersøkelsene gjøres av uavhengige testpersoner som ikke kjenner til pasientens gruppetilhørighet. Tabell 2 viser hvilke undersøkelser som gjøres til hvilke tidspunkter.

Primært endepunkt

Wolf Motor Function test (WMFT) er det primære endepunkt for denne studien. Dette er en funksjonstest som har vært brukt i over 20 studier av CI terapi. Den består av 15 prøver hvor tid og funksjonell kvalitet blir evaluert, og 2 styrkeprøver [23]. Øvelsene er organisert etter kompleksitet, og går fra store skulderbevegelser til fingerbevegelser. Reliabilitet og validitet til testen har vært dokumentert i flere studier [23, 24].

Sekundære endepunkt

- Accelerometre – Dette er accelerasjonsmålere som deltakeren må ha på hver arm igjennom to hele dager. Disse registrerer aktiviteten i armen i to sekunders perioder. Antall perioder med aktivitet i hver arm summeres og gjøres om til et estimat for bruken av armen. Hensikten er å undersøke forholdet mellom bruken av den friske og den syke armen [25].
- Fugl Meyer funksjonsundersøkelse (FMA) – Undersøker muskelfunksjonen i arm og hånd [26].
- Motor Activity Log (MAL) - undersøker bruken av den dårligste armen i dagliglivet. Dette er et strukturert intervju og omhandler 30 forskjellige aktiviteter. Pasienten skårer

  selv hvor ofte armen brukes og hvor bra bevegelsene er [5]. Disse undersøkelsene av behandlerne på dag 1og 2 av intervensjon og på dag 9 og 10 av intervensjonen, samt på 12 måneders kontroll av uavhengig undersøker.
- Stroke Impact Scale (SIS) – Spørreskjema som undersøker hvilke følger hjerneslaget har fått for pasienten med hensyn på fysiske problemer, hukommelse, evne til å kontrollere følelser, evne til å kommunisere, hvordan du beveger deg, din livskvalitet og hvordan du klarer aktiviteter i dagliglivet. .
- LifeSat-11 - Kartlegger hvor tilfreds deltakeren er med livet som helhet, yrkessituasjon, økonomi, fritid, sosialt liv, seksualliv, familieliv og ekteskap,

fysisk helse og psykisk helse.

- Goal Attainment Scaling – Undersøker graden av måloppnåelse hos pasienten. Gjennom intervju og samtaler med pasienten blir undersøkeren enig om hvilke mål som er viktig for pasienten. Pasienten scorer så hvor viktig han mener målet er og hvor vanskelig han tror det blir å nå dette målet. Måloppnåelsen evalueres på en 5 punkts skala.

  Måloppnåelsen vektes så i forhold til viktighet og vanskelighetsgrad og vi får en tallverdi for pasientens måloppnåelse [27, 28]. Det settes opp tre mål på PRE1 undersøkelsen og nye 3 mål på PRE2 undersøkelsene. Måloppnåelse evalueres på alle etterfølgende undersøkelsespunkter.
- Nine Hole Peg test – Måler tiden det tar å plassere 9 pinner i en Peg-board.

Kofaktorer

- Sittebalansedelen fra Trunk Impairment Scale (TIS-SB)
-
- 5x Sit to stand (5STS) –
- Funksjon i underekstremitetene
- Functional Reach (FR) – Statisk balanse.
- Kognitiv funksjon. MMS
- Barthel ADL index – Undersøker i hvilken grad deltakeren klarer deg selv i dagliglivet.
- Annen fysio-, ergo, intensiv rehabilitering og treningsterapi i perioden.
- Trombolysebehandling
- NIHSS score.
- Antall behandlingstimer under intervensjonen
- Bruk av behandlingsvott.
- Borgs Skala 2x pr dag under treningen

	Inkl. Screening	PRE1	Behandling	POST1	PRE2	Behandling	POST2	12M oppfølging
Sosiodemografiske data		X						
NIHSS		X			X			X
WMFT		X		X	X		X	X
Accelerometre		X		X	X		X	X
Fugl-Meyer o-eks		X		X	X		X	X
TIS-SB, 5STS, FR		X		X	X		X	X
Goal Attainment Scaling		X		X	XX		XX	XX
MMS		X			X			X
Barthel index								
SIS					X			X
LifeSat-11								
Motor activity log			T:X-X			S:X-X		X
Annen trening / rehabilitering					X			X
Antall behandlingstimer			X			X		
Bruk av behandlingsvott			X			X		
Borgs Skala (2x pr dag)			X			X		
Arbeidsførhet					X			X

Table 2: Undersøkelser og tidspunkter. X – undersøkelsen gjennomføres. XX – PRE1og PRE2 mål evalueres. T:X-X Tidliggruppe undersøkes i starten og slutten av intervensjonen. S:X-X – Sengruppen undersøkes i starten og slutten av intervensjonen.

Dataanalyse og statistikk

Data skal analyseres etter intention-to-treat prinsippet. Alle pasientene vil bli analysert i den gruppen de opprinnelig ble allokert til uavhengig av hvilken behandling de mottar, og selv om de bytter gruppe under veis. Manglende data vil bli erstattet med den siste registrering på den aktuelle deltaker. Forandringer innenfor gruppene vil bli analysert med parrede t-tester. Forskjell mellom gruppene vil bli analysert med variansanalyse for repeterte nålinger. Det tillates 5% sannsynlighet for type 1 feil (α=0,05) med tosidige signifikanstester. WMFT har vanligvis en forskjøvet distribusjon [8] og dersom dette er tilfelle i vårt utvalg vil vi benytte en logaritmetransformering for å utvide de nederste delene av skalaen. Fishers exact test vil bli brukt ved små tall, korrelasjonsanalyser og evt. variansanalyser vil bli benyttet for å studere sammenheng mellom ulike pasientkarakteristika, intervensjon og behandlingsresultat.

Styrkeberegning

Styrken er beregnet for sannsyneligheten for å vise forskjell mellom gruppene på PRE2 undersøkelsen før kontrollgruppen får sin behandling. Vi bruker standardavvik og behandlingseffekt fra EXCITE studien [29] som grunnlag for beregningen. Vi regner med at standardavviket ved inklusjon (logWMFT=1.02) kan representere standardavviket i begge gruppene. Og at behandlingseffekten fra pre til posttest i behandlingsgruppen (logWMFT=0,55) kan representere den forskjellen vi kan regne med å få mellom gruppene på dette tidspunktet. 53 pasienter i hver gruppe vil da gi en power på 0,80 [30]. Med iberegnet frafall på 10% får vi da et minimums deltakerantall på totalt 117 personer.

Personvern, etikk og formidling av resultater

Studien er meldt til og godkjent av Personvernombudet for forskning ved Universitetssykehuset Nord-Norge. Studien er også godkjent av Regional komité for medisinsk og helsefaglig forskningsetikk i Nord-Norge. Det tas sikte på å publisere 5 artikler med utgangspunkt i de 5 problemstillingene som er beskrevet. Artiklene skal publiseres i internasjonale medisinske tidskrifter. Vi forplikter oss til å publisere positive så vel som negative resultater.

Samarbeidspartnere

En viktig del av prosjektet er å bygge opp en infrastruktur som kan brukes til å gjøre klinisk kontrollerte studier på slagrehabilitering i Norge. De kontakter og nettverk som bygges under denne studien vil kunne benyttes ved senere studier og forenkle oppstarten av disse. Vi deler opp samarbeidspartnere i forskningssykehus og leverandørsykehus. Forskningssykehusene både rekrutterer og behandler pasienter i studien mens leverandørsykehusene rekrutterer pasienter som behandles på forskningssykehusene. Universitetssykehuset Nord-Norge (UNN) har det overordnede ansvar for prosjektet. Avdelingsoverlege dr. med. Audny Anke er prosjektleder og stipendiat Gyrd Thrane har det daglige ansvaret for prosjektet. Det er opprettet en styringsgruppe som i tillegg til de overnevnte består av Bent Indredavik dr.med avdelingsoverlege, Torunn Askim PhD forsker og Roland Stock MSc fysioterapeut, alle fra St. Olavs Hospital. Nina Emaus PhD, post.doc Institutt for samfunnsmedisin Universitetet i Tromsø er tilknyttet som biveileder for stipendiat. St. Olavs hospital og UNN inngår som forskningssykehus. Nordlandssykhuset v/Avdelingsoverlege Gaute Jensen, Sykehuset Ålesund og Aker Universitetssykehus v/Overlege Sigurd Vatn er forespurt om å delta som forskningssykehus. UNN Harstad v/Seksjonsoverlege Guri Heiberg, Sykehuset Levanger v/ Avdelingsoverlege Dagfinn Thorsvik er forespurt om å delta som leverandørsykehus.

Referanser

(1) MacDonald BK, Cockerell OC, Sander JW, Shorvon SD. The incidence and lifetime prevalence of neurological disorders in a prospective community-based study in the UK. *Brain* 2000 April;123 (Pt 4):665-76.

(2) Schaechter JD. Motor rehabilitation and brain plasticity after hemiparetic stroke. *Prog Neurobiol* 2004 May;73(1):61-72.

(3) Nakayama H, Jorgensen HS, Raaschou HO, Olsen TS. Recovery of upper extremity function in stroke patients: the Copenhagen Stroke Study. *Arch Phys Med Rehabil* 1994 April;75(4):394-8.

(4) Viitanen M, Eriksson S, Asplund K, Wester PO, Winblad B. Determinants of long-term mortality after stroke. *Acta Med Scand* 1987;221(4):349-56.

(5) Taub E, Miller NE, Novack TA et al. Technique to improve chronic motor deficit after stroke. *Arch Phys Med Rehabil* 1993 April;74(4):347-54.

(6) Mark VW, Taub E. Constraint-induced movement therapy for chronic stroke hemiparesis and other disabilities. *Restor Neurol Neurosci* 2004;22(3-5):317-36.

(7) Taub E, Uswatte G, King DK, Morris D, Crago JE, Chatterjee A. A placebo-controlled trial of constraint-induced movement therapy for upper extremity after stroke. *Stroke* 2006 April;37(4):1045-9.

(8) Wolf SL, Winstein CJ, Miller JP et al. Effect of constraint-induced movement therapy on upper extremity function 3 to 9 months after stroke: the EXCITE randomized clinical trial. *JAMA* 2006 November 1;296(17):2095-104.

(9) van der Lee JH, Wagenaar RC, Lankhorst GJ, Vogelaar TW, Deville WL, Bouter LM. Forced use of the upper extremity in chronic stroke patients: results from a single-blind randomized clinical trial. *Stroke* 1999 November;30(11):2369-75.

(10) Dahl AE, Askim T, Stock R, Langorgen E, Lydersen S, Indredavik B. Short- and long-term outcome of constraint-induced movement therapy after stroke: a randomized controlled feasibility trial. *Clin Rehabil* 2008 May;22(5):436-47.

(11) Sterr A, Elbert T, Berthold I, Kolbel S, Rockstroh B, Taub E. Longer versus shorter daily constraint-induced movement therapy of chronic hemiparesis: an exploratory study. *Arch Phys Med Rehabil* 2002 October;83(10):1374-7.

(12) Dettmers C, Teske U, Hamzei F, Uswatte G, Taub E, Weiller C. Distributed form of constraint- induced movement therapy improves functional outcome and quality of life after stroke. *Arch Phys Med Rehabil* 2005 February;86(2):204-9.

(13) Dromerick AW, Lang CE, Powers WJ et al. Very Early Constraint-Induced

Movement Therapy (VECTORS): Phase II Trial Results. 2007 p. 465.

(14) Boake C, Noser EA, Ro T et al. Constraint-induced movement therapy during early stroke rehabilitation. *Neurorehabil Neural Repair* 2007 January;21(1):14-24.

(15) Hakkennes S, Keating JL. Constraint-induced movement therapy following stroke: a systematic review of randomised controlled trials. *Aust J Physiother* 2005;51(4):221-31.

(16) Bonaiuti D, Rebasti L, Sioli P. The constraint induced movement therapy: a systematic review of randomised controlled trials on the adult stroke patients. *Eura Medicophys* 2007 June;43(2):139- 46.

(17) Dobkin BH. Strategies for stroke rehabilitation. *Lancet Neurol* 2004 September;3(9):528-36.

(18) Sterr A, Szameitat A, Shen S, Freivogel S. Application of the CIT concept in the clinical environment: hurdles, practicalities, and clinical benefits. *Cogn Behav Neurol* 2006 March;19(1):48-54.

(19) Sterr A. Training-based interventions in motor rehabilitation after stroke: theoretical and clinical considerations. *Behav Neurol* 2004;15(3-4):55-63.

(20) Siegert RJ, Lord S, Porter K. Constraint-induced movement therapy: time for a little restraint?
Clin Rehabil 2004 February;18(1):110-4.

(21) Askim T. Recovery after stroke. Assesment and treatment; with focus on motor function Norwegian University of Science and Technology; 2008.

(22) Lang CE, Wagner JM, Edwards DF, Dromerick AW. Upper Extremity Use in People with Hemiparesis in the First Few Weeks After Stroke. *J Neurol Phys Ther* 2007 June;31(2):56-63.

(23) Wolf SL, Thompson PA, Morris DM et al. The EXCITE trial: attributes of the Wolf Motor Function Test in patients with subacute stroke. *Neurorehabil Neural Repair* 2005 September;19(3):194-205.

(24) Morris DM, Uswatte G, Crago JE, Cook EW, III, Taub E. The reliability of the wolf motor function test for assessing upper extremity function after stroke. *Arch Phys Med Rehabil* 2001 June;82(6):750-5.

(25) Uswatte G, Foo WL, Olmstead H, Lopez K, Holand A, Simms LB. Ambulatory monitoring of arm movement using accelerometry: an objective measure of upper-extremity rehabilitation in persons with chronic stroke. *Arch Phys Med Rehabil* 2005 July;86(7):1498-501.

(26) Fugl-Meyer AR, Jaasko L, Leyman I, Olsson S, Steglind S. The post-stroke hemiplegic patient. 1. a method for evaluation of physical performance. *Scand J*

Rehabil Med 1975;7(1):13-31.

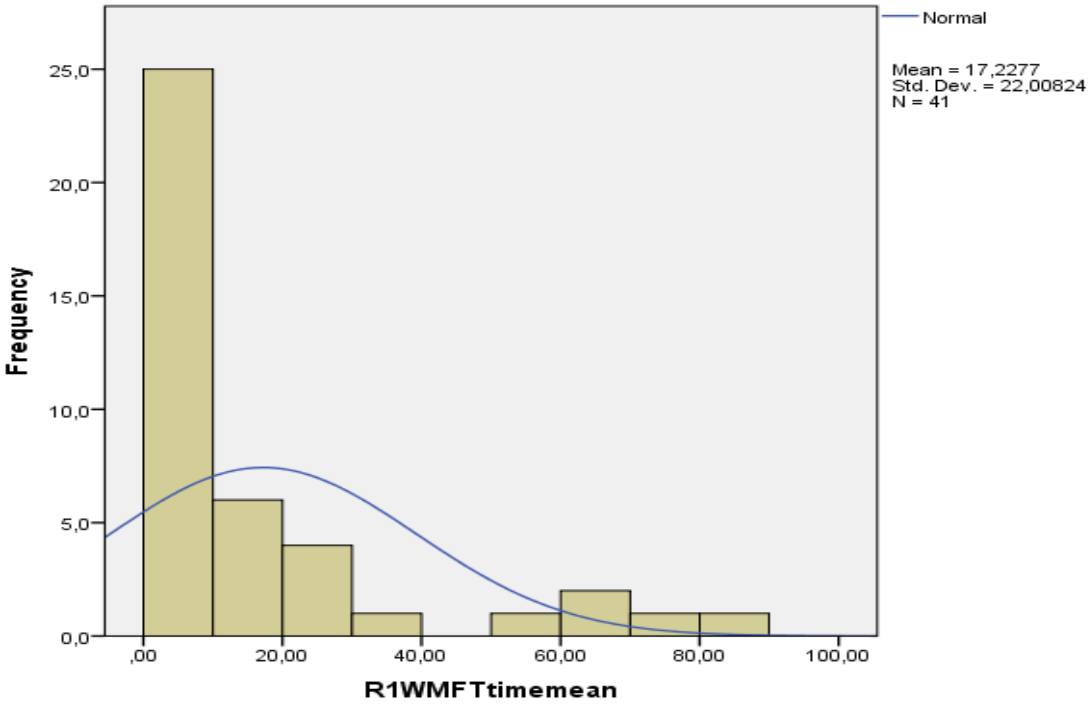
(27) Rushton PW, Miller WC. Goal attainment scaling in the rehabilitation of patients with lower- extremity amputations: a pilot study. *Arch Phys Med Rehabil* 2002 June;83(6):771-5.

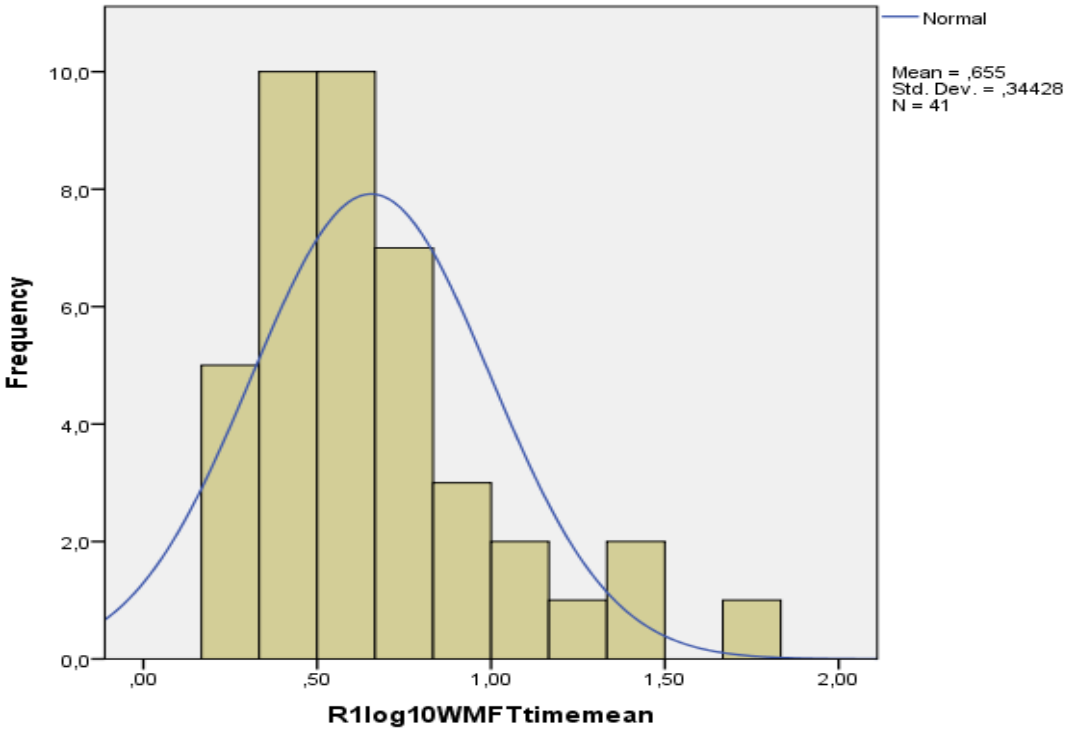
(28) Ashford S, Turner-Stokes L. Management of shoulder and proximal upper limb spasticity using botulinum toxin and concurrent therapy interventions: A preliminary analysis of goals and outcomes. *Disabil Rehabil* 2008 May 10;1-7.

(29) Thompson PA, Wolf SL. Means and standard deviations from the EXCITE trial. 2008.
Ref Type: Personal Communication

Java Aplets for Power and Sample Size [computer program]. Retrieved 10.07.2008: http://www.stat.uiowa.edu/~rlenth/Power; 2006.

Appendix 2: Histograms of distribution before and after log10-tranformation for rater 1 performance time scores




Normal
Mean = ,655
Std. Dev. = ,34428
N = 41
Frequency
10,0
8,0
6,0
4,0
2,0
0,0
,00
,50
1,00
1,50
2,00
R1log10WMFTtimemean

Appendix 3: Request for participation in research project

Samtykke til deltakelse i studien

Jeg er villig til å delta i studien

(Signert av prosjektdeltaker, dato)

Jeg bekrefter å ha gitt informasjon om studien

(Signert, rolle i studien, dato)