Validation of a Quantitative Single-Subject Based Evaluation for Rehabilitation-Induced Improvement Assessment

Marta Gandolla,[1] Franco Molteni,[2] Nick S. Ward,[3] Eleonora Guanziroli,[2] Giancarlo Ferrigno,[1] and Alessandra Pedrocchi[1]

[1]NearLab - Department of Electronics, Information and Bioengineering, Politecnico di Milano, Via G. Colombo 40, 20133 Milan, Italy; [2]Villa Beretta Rehabilitation Center, Valduce Hospital, Via N. Sauro 17, 23845 Costa Masnaga, LC, Italy; and [3]Sobell Department of Movement Neuroscience, UCL Institute of Neurology, 33 Queen Square, London, UK

(Received 12 November 2014; accepted 3 April 2015)
Published online: 17 April 2015

ABBREVIATIONS

EMG	Electromyography
EV	Endurance velocity
CGI-I	Clinical Global Impression scale for Improvement
CS	Capacity score
GV	Gait velocity
ICC	Interclass correlation coefficient
IS	Improvement score
MAS	Modified Ashworth Scale
MDC	Minimum detectable change
MRC index	Medical Research Council index
FD	Foot drop
FES	Functional electrical stimulation
PSL	Paretic step length
SEM	Standard error of measurement
TAAI	Tibialis anterior activation

INTRODUCTION

Neuro-motor rehabilitation is getting more and more importance in a social context where the aging of society and continuously improved ability to face acute intervention are enhancing the social impact of long-term disabilities. The goal of rehabilitation is not only to provide a support of a lost motor function (i.e., orthotic effect), but it is indeed to help an adequate plasticity to re-learn the lost motor function (i.e., therapeutic effect). Therefore, ideally, the outcome of a rehabilitation treatment is successful if it induces a stable improvement on the targeted task. Dealing in

Address correspondence to Marta Gandolla, NearLab - Department of Electronics, Information and Bioengineering, Politecnico di Milano, Via G. Colombo 40, 20133 Milan, Italy. Electronic mail: marta.gandolla@polimi.it

particular with neuro-motor rehabilitation trainings, different aspects of motor improvement can be nowadays accurately evaluated with a wide battery of measures, that are quantitative, reliable, and safe. Depending on the available equipment and the specific goal of the training, one or more outcome measures can be longitudinally assessed to help the clinicians in the diagnoses as well as in the therapy design.

So as to provide a clear and synthetic picture of the overall treatment outcome, a comparative and synthetic analysis of a wide variety of outcome measures is therefore required. At present, for each patient, the evaluation of significant improvement is performed every day by clinicians, taking advantage of one's own experience. However, the multi-dimensionality of the problem is very high, and it is difficult to distinguish advantages induced by the treatment or by external factors. Moreover, the evaluation is clinician-dependent, therefore subjective and strongly linked to the *a priori* expected outcomes. Inter-operator differences are indeed a relevant issue in the design of guidelines to treatment.

The improved accuracy and reliability in measuring different experimental parameters with highly accurate devices has paved the way to study automatic methods for a comprehensive evaluation. Comprehensive quantitative indices have been proposed in literature, based on the measure of how closely an individual gait pattern approaches to normal,[27,28] leveraging on mathematical methods such as principal component analysis,[10] or correspondence factor analysis.[22] However, these approaches were proposed to be only applied to kinematic parameters variation derived from gait analysis, and do not take into account the various correlation that exist in different motor domains (i.e., kinematic, kinetic, muscular, *etc.*). Another proposed approach is the principal component analysis of any set of variables to condensate their information, in order to provide a unique parameter to the correspondent analysis of neuroimages.[32] However, principal component analysis does not take into account data direction, and it is difficult to interpret. Overall these methodologies haven't really reached the bedside and clinicians still use clinical scores and instrumented measures as the base of a personal judgment which is the most diffused practice to assess whether a single subject effectively globally improved after a rehabilitation training.

In this study, we propose a quantitative comprehensive method of combining multiple measures in order to assess a stable functional improvement. This approach: (i) encompasses different aspects of motor function (e.g., kinematic, muscle activity, clinical indices, *etc.*); (ii) provides a method that can evaluate treatment-induced improvement in a single subject; (iii) is simple to interpret in that it provides a binary outcome in relation to improvement; (iv) is validated with respect to standard clinical evaluation. Outcome measures to be included for the functional improvement assessment have to be determined following the specific impaired element targeted, and consequently defined as the rehabilitation program goal. We hypothesized that a multidimensional assessment succeed in accurately quantifying the achievement of carryover effect on a single-subject base better than single measure assessment. Moreover, we hypothesized that the quantitative comprehensive assessment would map on the clinical qualitative assessment which is intrinsically a multidimensional assessment, since it is based on inputs at different levels and it is often the standard evaluation procedure before and after a neurorehabilitation treatment.

As test-bed, a functional electrical stimulation (FES) based treatment for foot drop (FD) correction—externally induced ankle dorsiflexion through peripheral FES during the swing phase of gait—performed with chronic post-stroke participants is presented. This context is particularly challenging since FES has been demonstrated to have therapeutic beneficial effects on FD.[21] This phenomenon, referred to as the 'carryover effect', has been observed in a number of subsequent studies, both as short or long lasting effect.[1,24,30] The carryover effect however is only seen in sub-groups of stroke patients treated with FES, but currently the characteristics of responders and non-responders are not clear. This is problematic when it comes to investigating the use of FES in large populations because the variability of response will dilute the carryover effect size at the group level, and this makes single-subject quantitative improvement evaluation even more crucial.

In summary, we propose and validate a consistent score along with a threshold definition based on minimum detectable change (MDC) to define a significant functional improvement on a single-subject base by merging different instrumental and clinical outcome measures. The proposed approach has been implemented in a custom-made software available along with this paper to support the use of the proposed approach in clinical practice.

MATERIALS AND METHODS

Participants

Patients were recruited from the outpatient and inpatient services at the Villa Beretta Rehabilitation Centre (Costamasnaga, LC, Italy). All patients had suffered from first-ever stroke >6 months previously, resulting in weakness of at least the tibialis anterior

muscle (to $<4+$ on the Medical Research Council (MRC) scale[23]). Exclusion criteria consisted of (i) responsiveness of less than 10° in FES-induced ankle dorsiflexion; (ii) language or cognitive deficits sufficient to impair cooperation in the study; (iii) inability to walk even if assisted; (iv) high spasticity at ankle joint plantar flexor as measured by the modified Ashworth scale index, MAS >2.[2]

Experiments were conducted with approval from the Villa Beretta Rehabilitation Centre Ethics Committee and all subjects gave informed written consent.

Fifteen patients were recruited in the study to allow possible dropouts. Eleven patients completed the 20 sessions of training, and six of them returned for the follow-up assessment. Table 1 outlines participants individual characteristics. Mean age was 43.6 ± 14 years, and mean time post-ictus was 34.7 ± 40.3 months.

Training

All patients were recruited for a specific FES-based treatment for FD correction. No other therapies for FD correction were delivered during this treatment. Patients were trained 5 times per week for 4 weeks, receiving a total of 20 sessions lasting 30 min of walking supported by a commercial electrical stimulator. Two commercial devices were available at the Villa Beretta Rehabilitation Centre: Bioness L300 (Bioness Inc.) and WalkAide (Innovative Neurotronics). Two stimulating electrodes were placed superficially along the peroneal nerve to elicit tibialis anterior muscle contraction during the swing phase of gait. Swing phase (trigger to start the stimulation) was detected on-line by wireless heel switches (Bioness) or by accelerometers (WalkAid), which is the only difference between the two stimulating devices. The more suitable commercial device was selected for each patient depending on his/her best responsiveness to stimulation, best wearability, and reliability of swing phase detection. Among the eleven patients that completed the training sessions, 3 patients used Bioness and 8 used WalkAid (Table 1). Current stimulation amplitude was selected for each participant at the beginning of each session so as to be able to elicit ankle dorsiflexion during gait, but at the same time to remain within the tolerance level.

Clinical and Instrumental Measures

Patients impairment at the time of recruitment for this study (t_1—within 5 days before the start of the treatment), after the intervention (t_2—within 5 days since the end of the treatment), and at a follow-up assessment (t_3—at least 1 month after the end of the intervention) was evaluated using a battery of clinical and instrumental tests. The patient's evaluation typically lasted 30–40 min, and included the following procedure. The patient was welcomed and seen by the referent clinician for the general visit and the MRC scale index evaluation at ankle dorsiflexion, which is a scale for grading muscle power in stages from 0 to 5. The patient's effort is graded from grade 5—muscle contracts normally against full resistance to grade 0—no movement is observed.[23] Afterwards, the patient was required to perform the gait analysis test performed following the standard Davis evaluation protocol (physical evaluation and measurement, 22 markers placement, static offset measurement, walking

TABLE 1. Participant characteristics.

Subj	Age (years)	Sex	Site of lesion	Type of lesion	Time (months)	t_2 ass	t_3 ass	Training device
pz01	53	M	L GP	H	37	No	No	WalkAid
pz02	37	F	R ACA	H	10	Yes	Yes (3 months)	Bioness
pz03	23	M	R MCA	H	23	Yes	Yes (1 months)	WalkAid
pz04	38	F	R GP	I	23	Yes	Yes (1 months)	Bioness
pz05	64	F	L MCA	H + I	13	Yes	Yes (2 months)	WalkAid
pz06	57	M	L Caudate Nucleus	I	6	No	No	WalkAid
pz07	19	M	L MCA	H	44	Yes	No	WalkAid
pz08	47	F	L GP	H	44	Yes	Yes (4 months)	WalkAid
pz09	25	F	R MCA	I	30	Yes	No	WalkAid
pz10	49	M	R MCA	I	89	No	No	WalkAid
pz11	46	M	R GP	I	13	Yes	Yes (1 months)	WalkAid
pz12	33	M	L parac lobule	I	6	Yes	No	WalkAid
pz13	61	F	R MCA	H	158	Yes	No	WalkAid
pz14	57	M	R parietal subc, BG	I	18	Yes	No	Bioness
pz15	45	F	R MCA	I	9	No	No	Bioness

Subj = subject; M = male; F = female; R = right; L = left; MCA = middle cerebral artery; ACA = anterior cerebral artery; GP = Globus Pallidus; parac = paracentral; subc = subcortical; BG = Basal Ganglia; H = haemorrhagic; I = ischemic; time = time since stroke at the time of t_1 acquisition; t_2 ass. = post-treatment assessment measure; t_3 ass = follow-up assessment measure.

performance)[11] in the Gait and Movement lab (Fig. 1), which is composed by two synchronized digital video cameras for video capturing and recording (BTS eVIXTA; BT); eight digital video cameras for kinematic acquisition (BTS SMART-DX 7000; BTS); and eight triaxial force plates for kinetic acquisition (BTS P-6000; BTS). 3D markers coordinates are computed stereometrically from the two dimensional camera data; each marker has to be identified by at least two cameras for each acquired frame, and it is tracked with operator assistance on BTS SMART-Analyzer commercial software, which is a complete solution for the biomechanical analysis of movement with three-dimensional kinematic/kinetic data. Each segment is then evaluated following a data reduction protocol detailed in the paper from Davis *et al.*[11] and implemented in the BTS SMART-Analyzer software. During the walking performance, patients muscular activity was recorded by means of a surface wireless dynamic electromyography system (BTS FREEEMG 300; BTS). Electrodes were placed on tibialis anterior, medial and lateral gastrocnemius, soleus, rectus femoralis, and hamstrings of each leg. Finally, patients' endurance was evaluated through the 6-min walking test performed indoors, along a long, flat, straight, enclosed corridor with a hard surface. The walking course was 50 m in length, and the length of the corridor was marked every 66 cm. The turnaround points were clearly marked. The 6MWT was performed following the instruction assessed in literature.[26]

From these tests, a set of outcome measures ($N = 5$) was designed to assess different aspects of patients' functional condition:

(i) GV—self-selected gait velocity as measured during the gait analysis test;
(ii) EV—endurance velocity, as calculated during the 6-min walking test;
(iii) PSL—paretic step length as measured during the gait analysis test;
(iv) TAAI—tibialis anterior activation index defined as the ratio between the EMG activity of the tibialis anterior muscle between toe off and toe strike and during the whole gait cycle;[7]
(v) MRC index at ankle joint.[23]

We selected the outcome measures to specifically address the treatment target (i.e. drop foot recovery) and how it has an effect on the complex motor controls system (i.e., gait), and therefore we selected the outcome measures both to monitor peripheral and central functions, i.e. the correct use of the peripheral function in the control of a complex task. In the specific case, our pool of patients was affected by drop foot. Drop foot can be caused by tibialis anterior muscle weakness and/or poor control or plantarflexor muscles hyperactivity. One of the participants' exclusion criteria was

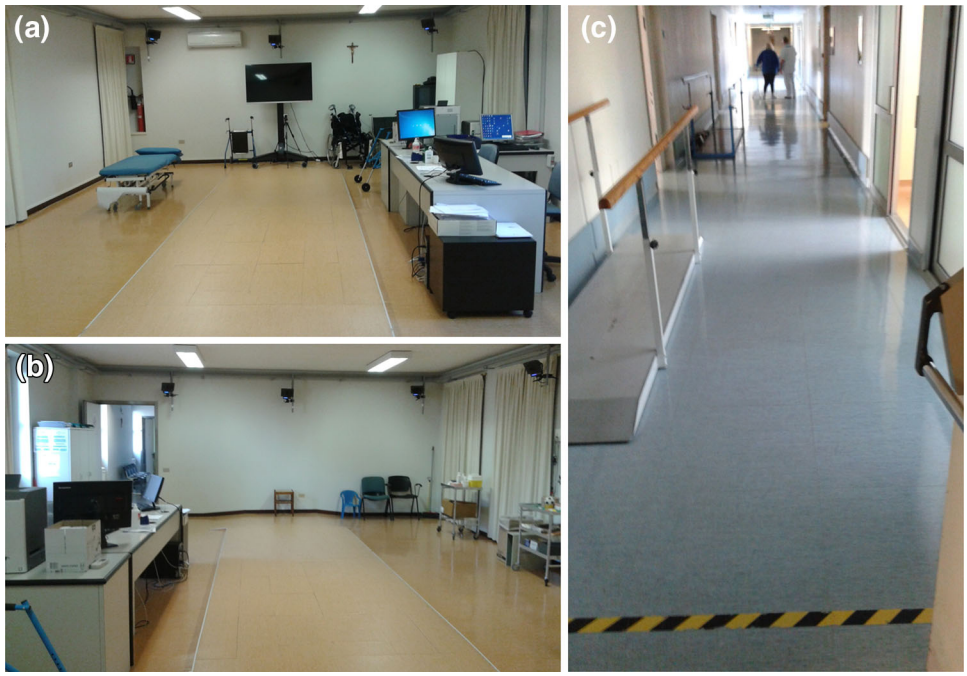

FIGURE 1. Testing environment. Testing environment in the clinical setting. (a, b) Gait analysis and movement lab; (c) corridor for 6 min walking test execution.

high spasticity at ankle joint plantar flexor as measured by the modified Ashworth scale index, and therefore our pool of patients had drop foot caused by tibialis anterior weakness and/or poor control, which is the specific target of the rehabilitation treatment. The monitored peripheral function was therefore ankle dorsiflexion recovery, and the central evaluated function gait. Ankle dorsiflexion recovery has been evaluated through paretic step length, which is a direct consequence of better foot clearance, Tibialis Anterior activation index, which indicates a correct muscle activation pattern during gait, and MRC index, which is a common clinical index on ankle residual function. Gait has been evaluated through gait velocity, and endurance velocity, common gait markers in literature.

Carryover Effect Evaluation by Clinicians

A group of 10 clinicians blindly evaluated each patient separately on the Clinical Global Impression scale for Improvement evaluation (CGI-I[8]) by looking at the videos with the patient walking (head was obscured), the gait analysis report, and the dynamic EMG report for t_1 and t_2 assessments. The carryover effect evaluation by the clinicians was performed with the same information that were used by the algorithm along with the obscured video of patients gait. It has to be noted that the clinician that was involved in scoring the patients on the MRC index was not involved in this evaluation procedure. The CGI-I scale is a global rating of the change in clinical status since the start of the treatment and ranges from 1 (very much improved) to 7 (very much worse), where 4 indicates no change. In order to obtain a carryover effect evaluation (i.e., CE = 1—carryover effect, CE = 0—no carryover effect) the mean of the GCI-I evaluation for each patient was calculated. The obtained carryover effect was assigned when the mean was less or equal to 3 (obtained improvement on the CGI-I scale).

Carryover Effect Evaluation by Algorithm

For all outcome measures an increase of the variable value reflects a functional improvement. Let OM_{Pi} be the 5-dimensional vector containing the 5 outcome measures for the ith patient at a given assessment timing. In addition let us consider a "subject zero" that represents a patient that scores zero in all outcome measures (i.e., it is the most impaired patient existent in our space). The outcome measures has been considered as equally important in the analysis, thus equally weighed in improvement evaluation algorithm. The algorithm flow is illustrated in Fig. 2.

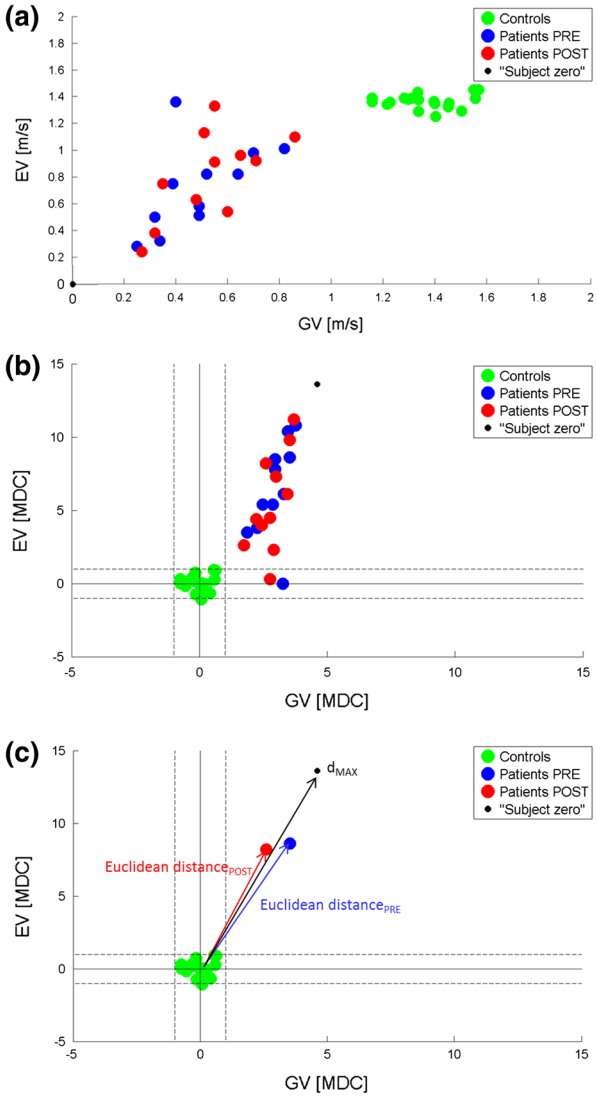

FIGURE 2. Illustration of the carryover effect evaluation algorithm. For illustration purposes, only two outcome measures are presented, gait velocity (GV), and endurance velocity (EV). (a) Plot of gait velocity in function of endurance velocity for a group of control subjects artificially generated as centered on controls mean, and for patients t_1 (i.e., pre-treatment) and t_2 (i.e., post-treatments) measures (real data). The "subject zero" (black dot) represents a patient that scores zero in all outcome measures (i.e., it is the most impaired patient existent in our space). (b) Normalization step of the algorithm. Dashed lines indicate 1 MDC from the origin (i.e., from the mean of the control subjects). (c) Illustration of the Euclidean distance calculation for a single patient. The more the patient is close to the origin, the more is close to an healthy subject control, and therefore the less is impaired. The "subject zero" is at maximum distance (d_{max}) from the origin.

Normalization

Let μ_C be the 5-dimensional vector containing the mean of a reference healthy population for each outcome measure derived from literature, and let MDC be the 5-dimensional vector containing the MDC value

for each outcome measure for the reference population (i.e., chronic post-stroke patients). A novel set of standardized outcome measures OM_{Pi_norm} can be defined as follows:

$$OM_{Pi\_norm} = \left| \frac{OM_{Pi} - \mu_C}{MDC} \right| \quad (1)$$

The standardization is useful so that the different outcome units of measure do not skew the results,[28] and so that the measure unit for each outcome variable is the appropriate MDC.

Up to now, we projected the patients in a 5-dimensional plane where the origin is where an healthy control subject would be, and therefore the faraway the patient is from the origin, the more is impaired (Fig. 2b). Eventual performances of patients better than healthy controls should be methodologically assigned at the origin, but they have never occurred. The revealed distance has to be considered significant if bigger than 1 (i.e., patient different from healthy controls of at least one MDC).

MDC is defined as $MDC = SEM * 1.96 * \sqrt{2}$, where SEM is standard error of measurement. In turn $SEM = std * \sqrt{1 - ICC}$, where std is standard deviation and ICC is interclass correlation coefficient. The sources of variability in MDC calculation are given by std and ICC. Given that std itself a measure of variability, and therefore takes into account the variability of the measure in the reference population, the source of variability can be restricted to ICC variability. In order to estimate the sensitivity of the proposed method to MDC values selection, the algorithm has been run spanning the 95% confidence interval for ICC.

Capacity Score (CS)

We projected in the defined 5-dimensional plane outcome variables values of t_1, t_2, and t_3 acquisitions for each patient, and we defined the Euclidean distance of each point from the origin (i.e., distance of the patient from the healthy control group). An increase in the Euclidean distance from the origin reflects that the patient is more impaired:

$$\text{Euclidean distance} = \sqrt{\sum_{j=1}^{5} \left(OM^{j}_{Pi\_norm}\right)^2} \quad (2)$$

$$j = 1, \ldots, 5,$$

where j indicates that the sum is performed over the 5 outcome measures (Fig. 2c).

The "subject zero" is at maximum distance (d_{max}) from the origin. We therefore obtained the capacity score, whose increment reflect a better gait capacity, as follows:

$$CS = d_{max} - \text{Euclidean distance}. \quad (3)$$

Improvement Score (IS)

The IS is defined as the difference between CS for each patient acquired at different timing (e.g. t_2 (i.e., post-treatment) $-$ t_1 (i.e., pre-treatment), t_2 (i.e., post-intervention) $-$ t_3 (i.e., follow-up assessment), *etc.*). If $IS > 0$ means that the patient improved.

Carryover Effect Definition

In particular, if $IS > 1$, the patient improved of at least one MDC unit over all the outcome measures. If $IS > 1$, we consider the patient to have achieved the carryover effect (i.e., stable functional improvement).

Carryover Effect Evaluation by Single Outcome Measure Assessment

Multidimensional assessment is useful if it allows for a better discrimination between groups (i.e., carryover effect/non-carryover effect) with respect to a single variable assessment. In order to test the ability of a single outcome measure to estimate the carryover effect, we assigned gained carryover effect to patients that presented an increase of at least one MDC relative to each single outcome measure.

Moreover, it has been tested with a non-supervised approach (k-means algorithm) if the difference between t_2 and t_1 single outcome measures lead to a distinction between carryover and non-carryover effect groups.

Statistical Analysis and Sample Size Calculation

Agreement Between Carryover Effect Evaluation by Clinicians and by Algorithm

Agreement between the proposed algorithm and clinical scoring has been evaluated with the Cohen's kappa coefficient,[9] between t_1 and t_2 assessment timing. It provides a measure of the degree to which two judges concur in their respective sorting of a number of items into mutually exclusive categories. Given that we are interested in testing whether the kappa is higher than zero (i.e., agreement better than chance), considering alpha equals 0.05, power equals to 80%, proportion of positive ratings 0.4,[24] we obtain a sample size of 10, if we want to be certain to detect a kappa equals to 0.8.[29]

TABLE 2. Improvement evaluation by clinicians on the GCI-I scale.

Part	C01	C02	C03	C04	C05	C06	C07	C08	C09	C10	Mean	CE
pz02	4	4	4	3	3	4	3	3	4	3	3.50	0
pz03	4	3	2	3	3	3	3	3	4	2	3.00	1
pz04	5	4	4	4	5	3	4	4	4	3	4.00	0
pz05	3	1	2	2	2	2	2	3	3	2	2.20	1
pz07	3	4	3	3	3	2	2	3	3	3	2.90	1
pz08	2	3	4	2	2	2	2	3	4	2	2.60	1
pz09	4	4	3	3	3	3	3	3	4	3	3.30	0
pz11	4	4	4	4	4	4	2	4	4	4	3.80	0
pz12	2	4	4	4	4	4	3	2	3	3	3.30	0
pz13	3	2	2	2	2	2	1	2	2	2	2.00	1
pz14	3	4	5	4	5	4	3	4	4	3	3.90	0

Part = participant; C = clinician; mean = mean of the CGI-I scale evaluation by clinicians; CE = carryover effect, where 1 indicates achieved and 0 non achieved carryover effect.

Agreement Among Clinicians

Agreement among clinicians has been evaluated with the Fleiss kappa,[14] a statistical measure of inter-rater reliability. Fleiss's kappa works for any number of raters giving categorical ratings and it can be interpreted as expressing the extent to which the observed amount of agreement among raters exceeds what would be expected if all raters made their ratings completely randomly. The scoring range is between 0 and 1. Since we didn't find any reference in literature about required sample size for Fleiss' kappa evaluation, we based our consideration on Cohen kappa sample size calculation.

RESULTS

Carryover Effect Evaluation by Clinicians

Fleiss kappa for evaluation agreement between the ten clinicians blindly evaluating the patients on the CGI-I scale, is equal to 0.22 (fair agreement[20]), rejecting null hypothesis that observed agreement is accidental with p value < 0.001. Mean evaluations among clinicians along with carryover effect achievement are shown in Table 2.

Controls Outcome Measures Values

Means and standard deviations for each outcome measure have been derived from literature as follows (mean ± standard deviation): GV—1.35 ± 0.21 (m s^{-1}), age range 64 ± 8;[4] EV—1.36 ± 0.24 (m s^{-1}), mean age 61.7 ± 9.9;[17] PSL—1.3 ± 0.5 (m), mean age 64 ± 10;[12] TAAI—0.70 ± 0.12, mean age 55.5 ± 16;[7] MRC—5.[23]

MDC Values for Chronic Post-Stroke Population

For each outcome measure the MDC reference value has been derived from literature as follows: GV—0.3 (m s^{-1});[15] EV—0.1 (m s^{-1});[13] PSL—26.2 (mm);[18] MRC—1.[23] For TAAI, we couldn't find the MDC value in literature, and therefore we estimated it as follows.

MDC is defined as $\mathrm{MDC} = \mathrm{SEM} * 1.96 * \sqrt{2}$, where $\mathrm{SEM} = \mathrm{std} * \sqrt{1 - \mathrm{ICC}}$. Std is standard deviation and ICC is interclass correlation coefficient, defined between 0 and 1, where 0 indicates no agreement and 1 indicates maximum agreement. We set ICC equals to 0.5 that indicate an intermediate situation, i.e., agreement by chance, and therefore we set this value as an upper bound. By using the std value equals to 0.17 of the stroke population,[7] we obtained an overestimate for the TAAI MDC equals to 0.21.

In order to estimate the sensitivity of the proposed method to MDC values selection, the algorithm has been run spanning the 95% confidence interval for ICC for gait velocity (0.682–0.936),[15] endurance velocity (0.98–0.99),[13] and paretic step length (0.992–0.998),[18] derived from literature. The TAAI MDC has been already overestimated as previously discussed, and therefore it hasn't been further analyzed. For what is concerning MRC MDC value, it is equal to 1 by definition, and therefore it does not have any source of variability.

Carryover Effect Evaluation by Algorithm

The algorithm has been implemented in an available custom-made and guided software (Fig. 3) to support the use of the proposed approach in clinical practice. Before to run the application verify the MATLAB Compiler Runtime (MCR) is installed and ensure you have installed version 8.3 (R2014a). If the MCR is not installed, or you have a different version, freely download it from the MathWorks Web site by navigating to—www.mathworks.com/products/compiler/mcr/index.html. The computational time needed to run the algorithm is less than 0.3 ms on a desktop PC (Intel

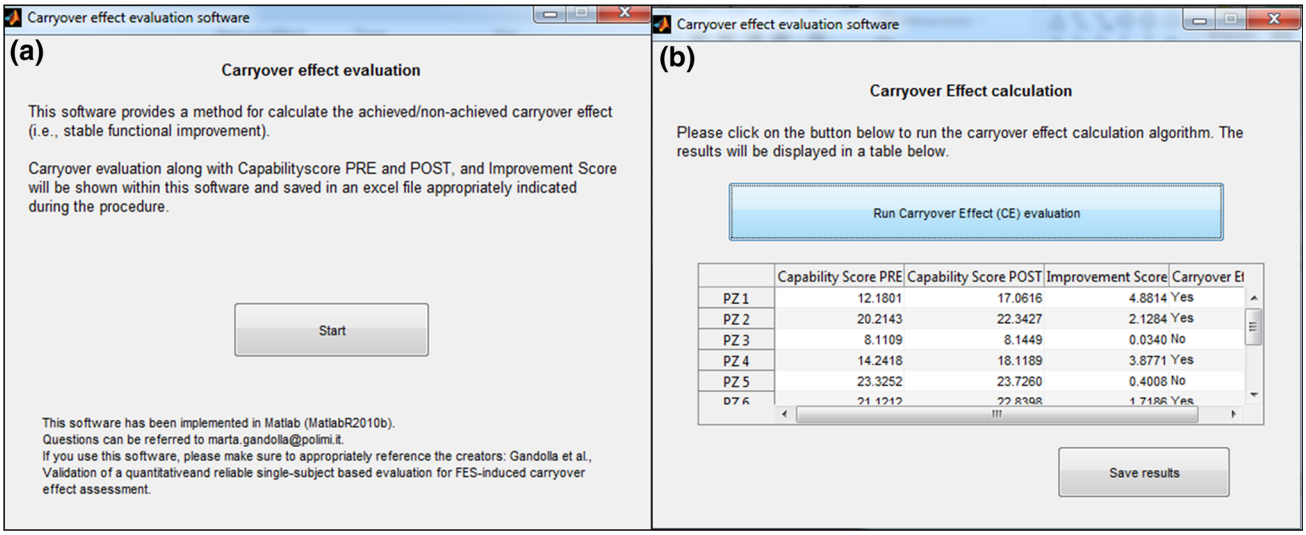

FIGURE 3. Custom-made software windows examples. Example of windows obtained during the custom-made software use. (a) introduction window; (b) carryover effect calculation window with the proper button to save the obtained results.

TABLE 3. Carryover effect evaluation by the algorithm.

Part	CS − t_1	CS − t_2	IS	CE (algorithm)	CE (clinicians)
pz04	8.15	8.17	0.02	0	0
pz02	12.20	17.19	4.99	1	0
pz05	14.33	18.15	3.82	1	1
pz14	14.93	15.06	0.13	0	0
pz12	16.25	16.05	−0.20	0	0
pz11	17.31	16.99	−0.32	0	0
pz13	18.50	19.76	1.26	1	1
pz03	20.30	21.32	1.02	1	1
pz08	21.26	23.00	1.74	1	1
pz07	23.42	23.82	0.40	0	1
pz09	24.50	22.70	−1.80	0	0

Patients are ordered from the most impaired to the less impaired following the capacity score evaluation at t_1 (i.e., before the treatment). Part = Participant; CS = gait capacity score; IS = improvement score; CE = carryover effect, where 1 indicates achieved and 0 non achieved carryover effect.

Core i7-2600 CPU @3.40 GHz). Capacity Score, Improvement Score, and carryover effect achievement obtained by the algorithm are summarized in Table 3 and Fig. 4. Five participants over eleven resulted to achieve the carryover effect. Among the six patients that have been evaluated in the follow-up assessment, the carryover effect presence/non-presence is demonstrated to be stable for all patients, but for one where the effect was no longer present (Fig. 4).

The differences in IS derived from MDC variations are equal to 0.022 ± 0.025 for GV MDC variation, 0.082 ± 0.121 for EV MDC variation, and 0.461 ± 0.485 for PSL MDC variation. Therefore, although the proposed algorithm is sensible to MDC parameter, the differences induced by spanning MDC variability can be considered as negligible with respect to the aim.

Agreement Between Clinicians and Algorithm Evaluation

Cohen's kappa for agreement between clinicians and algorithm evaluation is equal to 0.63 (substantial agreement),[24] significantly rejecting the null hypothesis that observed agreement is accidental ($p = 0.03$). If we consider agreement as a 2×2 design where first factor is improvement (two levels—improved, non-improved), and second factor is judgement (two levels—clinically algorithmically), we can observe that both judgement approaches result in respectively 5 and 6 patients assigned to improvement and non-improvement level. As it can be observed in Table 3, the difference is in two patients that are differentially assigned to the two categories.

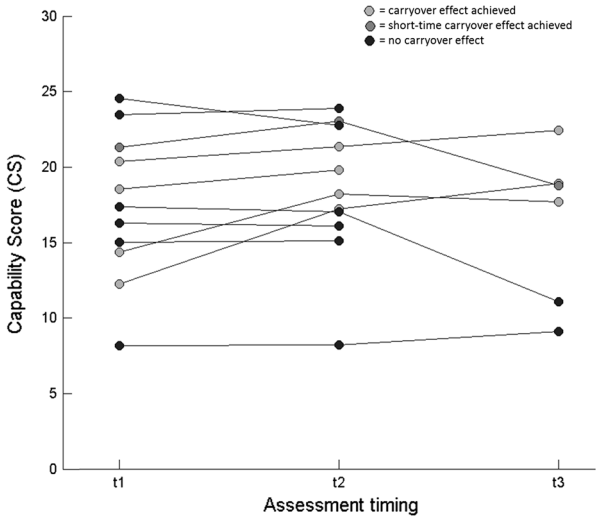

FIGURE 4. Capacity score (CS) assessment along time for all patients.

Agreement between clinicians and single outcome measure assessment

Outcome measures obtained from the patients who completed the training session are indicated in Table 4. Cohen's kappa for agreement between clinicians and single measure carryover effect assessment are equal to 0 ($p = 1$), 0.03 ($p = 0.93$), 0.44 ($p = 0.16$), −0.18 ($p = 0.72$), and 0.29 ($p = 0.41$) respectively for gait velocity, endurance velocity, paretic step length, Tibialis Anterior activation index, and MRC index. For all included outcome measures considered separately, the null hypothesis that observed agreement is accidental is not rejected.

Further, non-supervised clustering of the single outcome measures didn't show any superimposition of the revealed clusters with clinicians evaluation (Fig. 5).

DISCUSSION

The important clinical question after delivering a specific rehabilitation treatment is—has the patient functionally benefit from this treatment?" From a patient's point of view in fact, improvement is not primarily a change in specific parameters, but rather an improvement in functional execution.

According to International Classification of Functioning, Disability and Health, human dis-functioning can be evaluated under three perspectives: impairment, meaning dis-function of a body function; activity limitation, meaning dis-function of an activity; and participation restriction, meaning dis-function of the person in the social context.[17] Disability involves dis-functioning at one or more of the levels, and therefore it is important to include more dis-functioning aspects while evaluating patients' improvement following a specific treatment. This study presents a methodological approach to contemporary evaluate impairment and activity limitations, leaving for a second step the evaluation of the social context. To evaluate activity limitations, multiple factors, including one or more impairments have to be taken into account. In fact, it can't be taken for granted that the presence of a single impairment necessarily result in a degradation of function.[28]

At present, quantitative evaluation of treatment-induced effect are usually presented at group level, and the variability of response will dilute the treatment effect size in large populations. The group-level approach leverage on the hypothesis that all the patients are likely to respond to the treatment with the same mechanism, but this cannot be taken for granted, especially for those rehabilitation treatments that deal with neuro-motor approaches.

In this work, we have validated a method to provide a yes/no information on whether or not the patient has improved, encompassing multiple outcome measures to monitor different aspects of disability (e.g., kinematic, muscle activity, clinical indices, *etc.*).

When dealing with functional improvement evaluation, the first step would always be to identify the specific clinical question, the goal of the treatment and a set of outcome measures that evaluate body functions and activities. Variables selection is a crucial point in the application of the presented algorithm, and it is the result of a trade-off between the multidimensionality of the rehabilitation process, which therefore requires to be monitored by multiple measures, and an excessive number of variables that leads to assessment redundancy. The guide for outcome measures selection is the interest and the experience of the clinician that is in charge of the rehabilitation treatment, and the number of the variables to be included in the evaluation depends on the primary goal of the evaluated treatment. However, from a methodological point of view, the proposed algorithm could be applied to any pool of variables where MDC could be defined or estimated.

As test-bed in this study, we have validated the method on a specific clinical question with respect to expert clinicians' judgement. This is particularly important when dealing with comprehensive quantitative measures, since it is important that they reflect effective functional improvement and not only changes in gait or muscular patterns. In particular, the FES-based treatment goal was the restoration of walking ability, for a better independence and social integration as possible foreseen outcome. FES treatment for FD correction specifically acts on ankle dorsiflexion as underlying body function. Ankle dorsiflexion has been

TABLE 4. Outcome measures.

Part		GV (m s^{-1})	EV (m s^{-1})	PSL (mm)	TAAI	MRC
pz02	t_1	0.49	0.51	291	0.29	1
	t_2	0.51	1.13	395	0.55	3
	t_3	0.71	0.73	453	0.37	3
pz03	t_1	0.64	0.82	486	0.60	3
	t_2	0.71	0.92	508	0.38	4
	t_3	0.63	0.94	536	0.44	4
pz04	t_1	0.25	0.28	194	0.48	2
	t_2	0.27	0.24	196	0.40	3
	t_3	0.3	0.28	217	0.39	3
pz05	t_1	0.32	0.50	345	0.56	3
	t_2	0.60	0.54	445	0.36	3
	t_3	0.58	0.5	432	0.43	4
pz07	t_1	0.82	1.01	561	0.44	3
	t_2	0.86	1.10	567	0.53	4
pz08	t_1	0.52	0.82	513	0.46	3
	t_2	0.55	0.91	554	0.44	4
	t_3	0.37	0.72	450	0.49	4
pz09	t_1	0.70	0.98	591	0.62	3
	t_2	0.65	0.96	544	0.71	3
pz11	t_1	0.49	0.58	420	0.43	3
	t_2	0.48	0.63	410	0.22	3
	t_3	0.44	0.61	250	0.65	3
pz12	t_1	0.34	0.32	410	0.82	3
	t_2	0.32	0.38	400	0.71	3
pz13	t_1	0.40	1.36	430	0.69	2
	t_2	0.55	1.33	460	0.75	3
pz14	t_1	0.39	0.75	350	0.68	2
	t_2	0.35	0.75	350	0.68	4

Outcome measures obtained from the patients who completed the training session. t_1 = assessment before the intervention, within 5 days before the start of the treatment; t_2 = assessment after the intervention, within 5 days since the end of the treatment; t_3 = follow-up assessment, at least 1 month after the end of the intervention; GV = gait velocity; EV = endurance velocity; PSL = paretic step length; TAAI = Tibialis Anterior Activation Index; MRC = Medical Research Council index.

evaluated through paretic step length, tibialis anterior activity, and MRC index, whereas activity has been evaluated through gait and endurance velocity.

With respect to the specific clinical question, controversial results are shown in literature, where walking speed is often used as activity limitation marker. Several studies reported significant improvement in walking speed ranging between 6 and 22% when wearing the stimulation device (i.e., orthotic).[5,6,30] Less consistent results are reported for what is concerning gait velocity improvement even when the stimulation was turned off (i.e. carryover effect). Burridge *et al.*[6] and Kottink *et al.*[19] reported no therapeutic effect in stroke survivors after respectively 3 and 6.5 months of FES use. On the other hand, other studies reported therapeutic effect ranging from 12 to 16% in walking speed increase.[5,30]

A possible reason of these controversial results is that FES, even if patients show a similar baseline disability conditions[25] acts differently on a single-patient base, depending on mechanisms of action not yet known. Some hypotheses has been made,[3,16] but they still remain not fully proved.

The proposed algorithm allows to contemporary quantitatively evaluate the improvement of body functions and activity, and to define if the improvement is clinically significant on a single-subject base. Moreover, the proposed multimodal approach has demonstrated to better model the achieved carryover effect with respect to improvement evaluated on the base of a single outcome measure. Indeed, the improvement singularly evaluated on the outcome measures has demonstrated not to be a descriptor of the carryover effect phenomenon, while the multimodal approach achieved the agreement between algorithm and clinicians evaluation, which was considered as the gold standard.

A single study reported that FES for FD correction has orthotic effect for 76% of patients, and therapeutic value relevant in 40% of such cases, relying on expert clinician evaluation.[24] We observed the carryover effect in 45% of patients, agreed between a pool of expert clinicians and an automatic method, which can be used to support the subjective evaluation of the single clinician in everyday activity. A possible explanation of the FES effect reported only by 45% of pa-

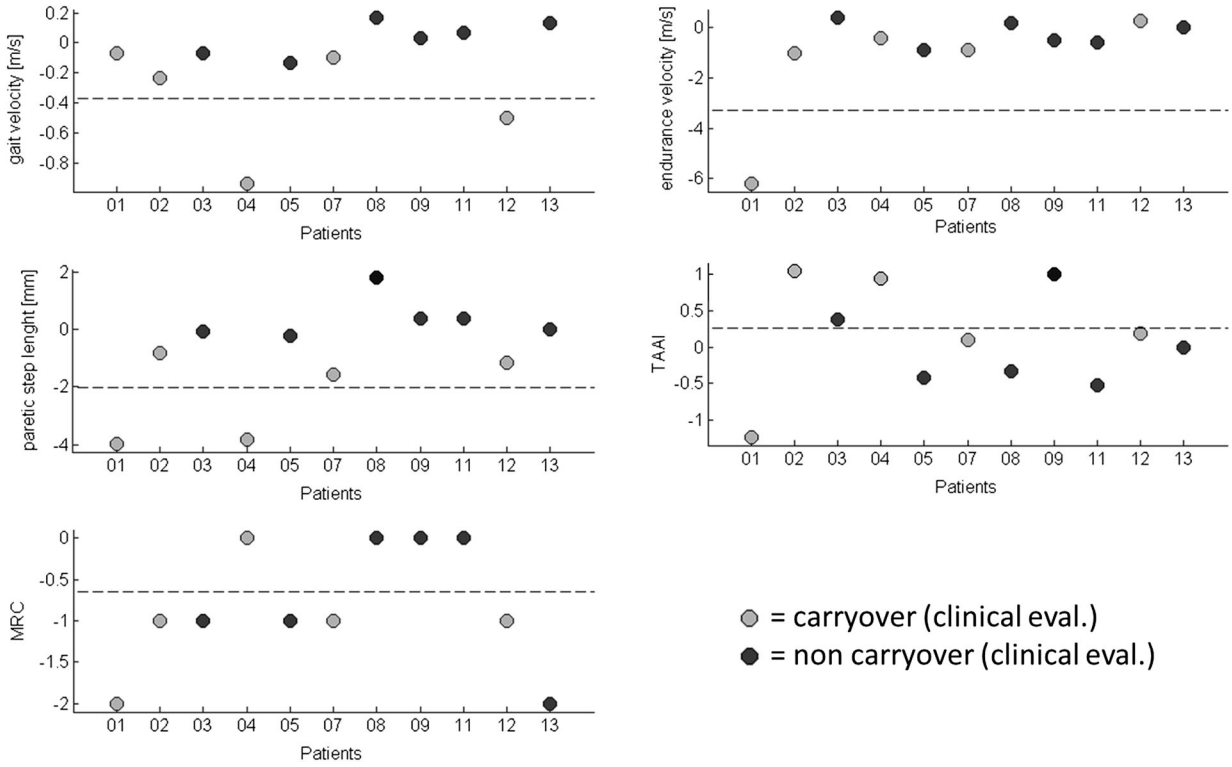

FIGURE 5. Non-supervised *k*-means clustering on normalized single outcome measures. *k*-means clustering with Euclidean distance. Dashed line indicates the division between the two clusters for each variable. Each patient is represented as the difference between the single outcome measure normalized following Eq. (1), after (t_2) and before (t_1) the treatment.

tients is that the treatment was lasting only 4 weeks and it has been demonstrated in literature that the longer the use, the better the benefit.[31] Anyway, it is known in the literature that the FES mechanism of action for recovery is primarily at central nervous system level, and that not all patients might benefit from FES at the same way.

Two patients resulted to be misclassified between algorithm and clinicians evaluation. Pz02 was classified as achieving the carryover effect by the algorithm, but not by clinicians evaluation. Indeed, by looking at the specific outcome measures values (Table 4), pz02 achieves an improvement in endurance velocity and paretic step length, which are difficult to be evaluated at sight during a clinical exam. Moreover, as it can be observed in Table 2, this patient has been evaluated as improved by five clinicians over ten. Therefore, this case highlights the support that a multimodal index as the one here proposed could be in clinical evaluation context. Whereas, pz07 was classified as achieving the carryover effect by the clinicians, but not by the algorithm. In fact this patient presents a better ankle control (i.e., better paretic step length, TAAI, and MRC indices), but not an overall functional improvement. Pz07 had already a good gait and endurance velocity that therefore didn't result to be overall improved, since the goal of the better ankle control was on gait quality rather than on gait functionality.

A limitation of the present study is the number of participants, which was quite limited even though it reached the sample size needed for agreement among clinicians and between algorithm and clinicians judgment. On the contrary, further studies are required to evaluate the incidence of carryover on a target population, to disclose FES-induced mechanisms of action, and possible predictive biomarkers.

CONCLUSIONS

The present study has rather a methodological approach—it combines different outcome measures to define an improvement score that informs about the carryover effect. The algorithm allows to quantitatively evaluate the functional improvement following every rehabilitation treatment on a yes/no base, and to singularly evaluate each patient going toward treatment personalization. The method has been implemented in a custom-made software to support the use of the proposed approach in clinical practice. This quantitative measure has been validated against clinicians' judgment with respect to a specific clinical question and it has been demonstrated to be more adequate in

describing the carryover effect phenomenon with respect to single outcome measure evaluation.

ELECTRONIC SUPPLEMENTARY MATERIAL

The online version of this article contains supplementary material, which is available to authorized users.

ACKNOWLEDGMENTS

This work was made possible thanks to the patients that agreed to participate to the project.

CONFLICT OF INTEREST

The authors report no declaration of interests.

REFERENCES

[1]Ambrosini, E., S. Ferrante, A. Pedrocchi, G. Ferrigno, and F. Molteni. Cycling induced by electrical stimulation improves motor recovery in postacute hemiparetic patients: a randomized controlled trial. *Stroke* 42:1068–1073, 2011.

[2]Ashworth, B. Preliminary trial of carisoprodol in multiple sclerosis. *Practitioner* 192:540–542, 1964.

[3]Bergquist, A. J., J. M. Clair, O. Lagerquist, C. S. Mang, Y. Okuma, and D. F. Collins. Neuromuscular electrical stimulation: implications of the electrically evoked sensory volley. *Eur. J. Appl. Physiol.* 111:2409–2426, 2011.

[4]Bohannon, R. W., A. W. Andrews, and M. W. Thomas. Walking speed: reference values and correlates for older adults. *J. Orthop. Sports Phys. Ther.* 24:86–90, 1996.

[5]Burridge, J. H., and D. L. McLellan. Relation between abnormal patterns of muscle activation and response to common peroneal nerve stimulation in hemiplegia. *J. Neurol. Neurosurg. Psychiatry* 69:353–361, 2000.

[6]Burridge, J. H., P. N. Taylor, S. A. Hagan, D. E. Wood, and I. D. Swain. The effects of common peroneal stimulation on the effort and speed of walking: a randomized controlled trial with chronic hemiplegic patients. *Clin. Rehabil.* 11:201–210, 1997.

[7]Burridge, J. H., D. E. Wood, P. N. Taylor, and D. McLellan. Indices to describe different muscle activation patterns, identified during treadmill walking, in people with spastic drop-foot. *Med. Eng. Phys.* 23:427–434, 2001.

[8]Busner, J., and S. D. Targum. The clinical global impression scale: applying a research tool in clinical practice. *Psychiatry* 4:28–37, 2007.

[9]Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20:37–46, 1960.

[10]Cretuala, A., K. Berveta, and L. Ballaza. Gillette gait index in adults. *Gait Posture* 32:307–310, 2010.

[11]Davis, R. B., S. Ounpuu, D. Tyburski, and G. R. Gage. A gait analysis data-collection and reduction technique. *Hum. Mov. Sci.* 10:575–587, 1991.

[12]Eng, J. J., A. S. Dawson, and K. Chu. Submaximal exercise in individuals with stroke: test-retest reliability and concurrent validity with VO_{2max}. *Arch. Phys. Med. Rehabil.* 85:113–118, 2004.

[13]Flansbjer, U., A. M. Holmback, D. Downham, C. Patten, and J. Lexell. Rehabilitation og gait performance tests in men and women with hemiparesis after stroke. *J. Rehabil. Med.* 37:75–82, 2005.

[14]Fleiss, J. L. Measuring nominal scale agreement among many raters. *Psychol. Bull.* 76:378–382, 1971.

[15]Fulk, G., and J. Echternach. Test-retest reliability and minimal detectable change of gait speed in individuals undergoing rehabilitation after stroke. *J. Neurol. Phys. Ther.* 32:8–13, 2008.

[16]Gandolla, M., S. Ferrante, F. Molteni, G. Guanziroli, T. Frattini, A. Marteganin, G. Ferrigno, K. Friston, A. Pedrocchi, and N. S. Ward. Re-thinking the role of motor cortex: context-sensitive motor outputs? *Neuroimage* 91:366–374, 2014.

[17]Geyh, S., A. Cieza, J. Schouten, H. Dickson, P. Frommelt, Z. Omar, N. Kostanjsek, H. Ring, and G. Stucki. ICF score sets for stroke. *J. Rehabil. Med.* 44:135–141, 2004.

[18]Kesar, T. M., S. A. Binder-Macleod, G. E. Hicks, and D. S. Reisman. Minimal detectable change for gait variables collected during treadmill walking in individuals post-stroke. *Gait Posture* 33:314–317, 2011.

[19]Kottink, A. I., H. J. Hermens, A. V. Nene, M. J. Tenniglo, C. G. Groothuis-Oudshoorn, and M. J. Ijzerman. Therapeutic effect of an implantable peroneal nerve stimulator in subjects with chronic stroke and footdrop: a randomized controlled trial. *Phys. Ther.* 88:437–448, 2008.

[20]Landis, J. R., and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics* 33:159–174, 1977.

[21]Liberson, W. T., H. J. Holmquest, D. Scot, and M. Dow. Functional electrotherapy: stimulation of the peroneal nerve synchronized with the swing phase of the gait of hemiplegic patients. *Arch. Phys. Med. Rehabil.* 42:101–105, 1961.

[22]Loslever, P., E. M. Laassel, and J. C. Angue. Combined statistical study of joint angles and ground reaction forces using component and multiple correspondence analysis. *IEEE Trans. Biomed. Eng.* 41:1160–1167, 1994.

[23]Medical Research Council/Guarantors of Brain. Aids to the Examination of the Peripheral Nervous System. London: Baillière Tindall, 1986.

[24]Merletti, R., A. Andina, M. Galante, and I. Furlan. Clinical experience of electronic peroneal stimulators in 50 hemiparetic patients. *Scand. J. Rehabil. Med.* 11:111–121, 1979.

[25]Rossini, P. M., C. Calautti, F. Pauri, and J. C. Baron. Post-stroke plastic reorganisation in the adult brain. *Lancet Neurol.* 2:493–502, 2003.

[26]Rostagno, C., and G. F. Gensini. Six minutes walk test: a simple and useful test to evaluate functional capacity in patients with heart failure. *Intern. Emerg. Med.* 3:205–212, 2008.

[27]Schwartz, M. H., and A. Rozumalski. The Gait Deviation Index: a new comprehensive index of gait pathology. *Gait Posture* 28:351–357, 2008.

[28]Shutte, L. M., U. Narayanan, J. L. Stout, P. Selber, and J. R. Gage. An index for quantifying deviations from normal gait. *Gait Posture* 11:25–31, 2000.

[29]Sim, J., and C. C. Wright. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys. Ther.* 85:257–268, 2005.

[30]Taylor, P. N., J. H. Burridge, A. L. Dunkerley, D. E. Wood, J. A. Norton, C. Singleton, and I. D. Swain. Clin-

ical use of the Odstock dropped foot stimulator: its effect on the speed and effort of walking. *Arch. Phys. Med. Rehabil.* 80:1577–1583, 1999.
[31]Von Schroeder, H. P., R. D. Coutts, P. D. Lyden, E. Billings, and V. L. Nickel. Gait parameters following stroke: a practical assessment. *J. Rehabil. Res. Dev.* 32:25–31, 1995.
[32]Ward, N. S., M. M. Brown, A. J. Thompson, and R. S. Frackowiak. Neural correlates of outcome after stroke: a cross-sectional fMRI study. *Brain* 126:1430–1448, 2003.