Quantitative Evaluation of Performance during Robot-assisted Treatment

E. Peri1✉; E. Biffi2; C. Maghini2; F. Servodio Iammarrone2; C. Gagliardi2; C. Germiniasi2; A. Pedrocchi1; A. C. Turconi2; G. Reni2

1 Neuroengineering and medical robotic laboratory (NEARlab), Electronics, Information and Bioengineering Department (DEIB), Politecnico di Milano, Milan, Italy; 2 Scientific Institute, IRCCS Eugenio Medea, Bosisio Parini (LC), Italy

Methods: Fourteen children affected by cerebral palsy (CP) performed a training with Armeo®Spring. Each session was evaluated with P, a simple parameter that depends on the overall performance recorded, and median and interquartile values were computed to perform a group analysis.

Results: Median (interquartile) values of P significantly increased from 0.27 (0.21) at T₀ to 0.55 (0.27) at T₁. This improvement was functionally validated by a significant increase of the Melbourne Assessment of Unilateral Upper Limb Function.

Conclusions: The parameter described here was able to show variations in performance over time and enabled a quantitative evaluation of motion abilities in a way that is reliable with respect to a well-known clinical scale.

Keywords
Assessment, robotics, rehabilitation, upper limb, cerebral palsy

Summary
Introduction: This article is part of the Focus Theme of Methods of Information in Medicine on “Methodologies, Models and Algorithms for Patients Rehabilitation”.

Objectives: The great potential of robots in extracting quantitative and meaningful data is not always exploited in clinical practice. The aim of the present work is to describe a simple parameter to assess the performance of subjects during upper limb robotic training exploiting data automatically recorded by the robot, with no additional effort for patients and clinicians.

1. Introduction

In clinical practice, it is substantial to identify among the multiple variables involved in rehabilitation treatments which ones might have a larger impact on outcomes and influence recovery. Such evaluations require the use of quantifiable, valid, and sensitive tools to guarantee reliable between-study comparisons and greatly improve the understanding of key treatment effects [1].

The main purposes of valid assessment methods have been largely discussed in the literature over the years. A useful test should be effective, efficient, comparable and predictable [2]. Unfortunately, many assessment methods commonly used today barely highlight the effectiveness of therapy treatments [3] since they are based on subjective impressions and could depend both on operators’ ability and on patients’ personality, and attitude. Further, they do not provide assessment during the training.

During the past few years, technologies gains an important role in rehabilitation [4] and robot-assisted rehabilitation has become a very active area of research [5, 6], since it provides controlled, intensive task-specific training that is goal directed and cognitively engaging.

Measures derived from robots can contribute to the understanding of how different treatment variables (e.g., dosage, amount, and type of assistance provided) influence motor learning and recovery [7]. Between other advantages, robot-based data have higher resolution, better inter-rater and intra-rater reliability with respect to clinical scales and are able to quantify some physical quantities (i.e. kinematics and force data) that might be useful to investigate the neurorecovery. First attempts to extract meaningful information from these data are now present [3, 7, 8] and show that robot-based measurements can be correlated with clinical scales providing clinicians with data they need for assessing the patient’s capability, progresses during therapy, and ongoing therapy needs [3, 7].

Previous studies have developed some ad-hoc assessment tools to extract outcome measures of patients’ performance [8, 9], often requiring additional time for patients and clinicians. Others have exploited the built-it technology to extract indexes of task precision, movement smoothness and velocity [10]. Only few assessment meth-
methods were validated with functional scales [7,11].

However the achievement of meaningful information from these data is still an ongoing challenge [3].

In this work we describe a simple parameter that can be easily derived from data saved by the robot and that quantifies the subjects’ performance. It can be used to follow the trend of a robot-aided treatment, to describe changes in performance before and after a rehabilitation and thus to investigate the effects of variations in the therapy on patients’ motor and functional recovery.

2. Materials and Methods

Fourteen inpatients (8–16 y) affected by cerebral palsy (CP) performed a training with the paediatric ArmeoSpring (▶ see Figure S1 in the supplementary online material), a five degree of freedom exoskeleton which guarantees passive arm weight support by means of springs, with a pressure-sensitive handgrip and virtual reality visual feedback [12]. The research protocol was approved on March 2010 by the Ethics Committee of IRCCS E. Medea.

2.1 Experimental Design

Every subject underwent 15–20 sessions lasting 30 minutes over 3–4 weeks of training depending on clinical suggestions. During each session, subjects performed a customized pull of exercises with the supervision of a physiotherapist. Eight exercises were selected to evaluate subjects’ performance over different joints and that require movements in different spaces (1D, 2D and 3D), accordingly to the indication of physiotherapists and clinicians. In particular, we evaluated one exercise performed in a 1D space (goalkeeper), five 2D exercises (egg cracking, fruit shopping, stove cleaning, moorhuhun and vertical catching) and two in a 3D space (chase balloon and reveal panorama). Details about the exercises are reported elsewhere [12]. During each training session, information about the exercise such as the scheduled difficulty level, the presence of the automatic grasping function, the thresholds to modulate the difficulty in grasping, the score obtained by the subject and the time required to perform the exercise were automatically recorded by the system, with no additional effort for the physiotherapists.

All these data have been included in a comprehensive performance parameter (p_i) computed for the i-th exercise during each session, as in ▶ Equation 1:

\[ p_i = \frac{S_i / S_{i,TOT}}{T_i / T_{i,TOT}} D_i \]  

where \( S_i \) is the score obtained during the i-th exercise, \( S_{i,TOT} \) is the maximum score obtainable, \( T_i \) is the time required to complete the i-th exercise, \( T_{i,TOT} \) is the maximum time available to finish the i-th exercise and \( D_i \) is the difficulty coefficient that considers the level of the i-th exercise and variation in autogrip and control threshold for each subject during the training.

In order to compare different exercises, \( p_i \) has been divided over the maximum performance achieved in the i-th exercise by the group of subjects (P_i).

The median value of P_i over the eight selected exercises (P) was used to follow the training of each subject session by session. P is an index of the overall performance.

Moreover, the median value of P within the first week (T_0), between the 12th and the 16th days (T_1/2) and within the fourth week (T_1) of training was computed for each subject and finally the median value over the 14 subjects was calculated at the three time points.

Moreover we performed the Melbourne Assessment (M), which evaluates the level of impairment of upper limb motor function with a well-established inter-intrarater reliability [13], and we computed its the median value within the group at T_0 and T_1.

A validation of P has been proposed by comparing, for each subject, P variations with M variations referred to the maximum value obtained by the group.

2.2 Statistical Analysis

A non-parametric Friedman test for paired samples was performed on P values between T_0, T_1/2, T_1. If these values were found significantly different (p < 0.05), a post-hoc analysis (Wilcoxon test with the Bonferroni correction) was performed comparing paired groups. Wilcoxon test was performed on M at T_0, T_1.

The Pearson’s correlation coefficient R was computed to evaluate the strength of the linear relationship between variations of M and P, while the root mean square error (RMSE) and the slope of the linear fitting were used to give details about the deviation from linearity and the sensitivity of the P parameter, respectively. Results are reported as median values (IQ), where IQ is the difference between the 3rd and the 1st quartiles.

3. Results

3.1 Evaluation of the Training

The effect of the training measured in terms of P was different for each subject but an overall improving trend was observed (linear fitting of the median curve over all the subjects, R = 0.95, ▶ Figure 1A). Moreover, differences in the trend of P corresponded to different M scores at the beginning and at the end of the training. ▶ Figure 1B reports an example of the training on two subjects (S1 and S2): in terms of P, S1 showed a linear increasing trend with a high Pearson’s correlation coefficient (R = 0.89) while S2 had an almost flat and stable performance (R = 0.31). This was denoted by changes in terms of M: both subjects obtained a score of 89 at T_0 but S1 reached a total score of 97 while S2 did not vary his performance after the training. Oscillation of the results of single session may depend on motivation, effort and by fatigue of the patient at that time.

3.2 Evaluation at T_0, T_1/2, T_1

▶ Table 1 reports medians and IQ values of the performance parameter P_i (for every selected exercise) and P, at the beginning, the middle and the end of the training over all the subjects.

P_i varied significantly between T_0 and T_1 in all the exercises (p < 0.05) except for egg cracking, but the trend is different in some cases. Goal keeper, stove cleaning, reveal panorama and vertical catching did
not significantly differ between $T_{1/2}$ and $T_1$, showing a plateau during the second half of the training. Differently, fruit shopping and egg cracking needed a longer training to vary and no significance was obtained between $T_0$ and $T_{1/2}$. Finally, the other exercises significantly increased both between $T_0$ and $T_{1/2}$ and $T_{1/2}$ and $T_1$.

Deepening the analysis for the four exercises that give the possibility to train the grasping function (stove cleaning, reveal panorama, egg cracking and fruit shopping), the first two were mainly performed with no active grasping (9 out of 12 and 10 out of 13 subjects respectively used the robot autogrip function during the training sessions) and a significant improvement was recorded between $T_0$ and $T_{1/2}$. In contrast, egg cracking and fruit shopping were performed with the autogrip function disabled for 12 out of 14 and for 9 out of 11 subjects respectively, thus training also the grasping function, with improvements only between $T_{1/2}$ and $T_1$.

Further, a statistically significant variation of $P$ was obtained over time (last row of Table 1).

### 3.3 Functional Validation of $P$

$M$ showed a significant improvement between $T_0$ and $T_1$ (median 85, IQ 7 vs. median 87, IQ 6, $p = 0.002$), highlighting progresses in terms of upper limb functionality. These improvements were related to the increased performance observed with $P$ (median 0.29, IQ 0.11 at $T_0$ and median 0.54, IQ 0.24 at $T_1$, $p = 1.5 \times 10^{-5}$). Indeed, the comparison between the variation of $M$ and $P$ ($\Delta M$ vs. $\Delta P$) for each subject showed a linear correlation with $R$ equal to 0.4 ($p = 0.123$), with a slope of the linear fitting equal to 2.3 and a RMSE between data and the linear fitting curve equal to 0.04.

### 4. Discussion

The need of reliable, quantitative and repeatable evaluations of the training effectiveness is an up-to-date theme in the clinical practice. In fact, in some cases clinical scales are operator-dependent and not sensitive enough to highlight changes felt by patients and their parents [8]. Some quantitative evaluations could be obtained by the use of ad-hoc technology (e.g. optoelectronic analysis of the kinematics, sensorized robots) but these methods are time and money consuming.

Here we propose a parameter $P$ that takes into account the time needed to finish an exercise, the scores obtained during the exercise and the level of difficulty. $P$ is computed in a simple and quick way from data automatically acquired by the robot during the training, and it does not require extra expensive devices. We used the pediatric version of Armeo Spring (Hocoma, AG) and we evaluated a pull of eight exercises. However, the value of the extracted parameter does not depend on the robot employed and on the exercise performed thus the presented method could be extended for other contexts.

Our results shows that an increase of $P$ corresponds to a functional improvement in terms of $M$, clinically validating the results obtained. A similar approach had

---

**Figure 1** Results of $P$ obtained, session-by-session, during the eight exercises considered. (A) Comparison between performance parameter $P$ obtained session by session by all the subjects. Median and interquartile values are shown as bold and colored area. (B) Comparison between performances over time of two selected subjects. The trend line is also shown.
been used in [11] where the validation of some motor indices with the Fugl Mayer clinical scale was described. The positive correlation observed between P and M was considered moderate as suggested by [14] and the not significance may be ascribed to the size of the sample. The RMSE value tends to zero and thus it was considered adequate. The slope suggests that P is more sensitive to variations during the training with respect to M. Anyway it should be considered that M is an outcome measure that gives a functional indication while the parameter P evaluates a trained task, thus being influenced by the learning process.

All the exercises analyzed showed significant variations of P over time. P allowed to follow the training of each subject, session by session which is a meaningful advantage with respect to the low temporal resolution of standard clinical evaluations. Moreover P can give indications about the performance changes between the beginning and the end of the training for each i-th exercise, thus being valuable for a pre-post evaluation.

The analysis of P could be also used to optimize the design of the training in terms of temporal planning, as suggested by our results.

First, some exercises led to the best performance achievement in only two weeks of training (e.g. goalkeeper, stove cleaning, reveal panorama and vertical catching) and indeed no statistically significant difference was found between T_{1/2} and T_1. However, these results were independent from the joint trained: in another exercise that trains the same body district of stove cleaning and reveal panorama (i.e. fruit shopping) subjects did not show any significant improvement by the middle of the therapy. This can be due to the training of the hand grip-release function, which could be responsible of the slower growth of the performance curve in fruit shopping. Indeed, the other exercise that was used to train the grasping function (eggs cracking) confirms this observation, as the trend revealed lower variations between T_0 and T_{1/2} with respect to the significant variations between T_{1/2} and T_1 (relative increase of 23.1% and 76.9% respectively). This suggests that the hand grip-release function requires more time to obtain motor improvements, thus ad-hoc exercises should be introduced early in the training to maximize their effectiveness.

Second, exercises with complex virtual environment (e.g. chase balloon and moorhuhn) needed longer training to achieve the best performance. Therefore, exercises with difficult virtual settings should be included at the beginning of the intervention since the visual feedback may be a crucial aspect of the training timeline.

Beyond the many details that P could provide to a therapy evaluation and planning, the overall performance of a subject or a group of patients was showed by P, that took into account motor performance of all the body districts trained in the different spaces in a comprehensive manner. Further, it highlighted variations in subject’s motor performance even at the middle of the training, that could not be retrieved taking in consideration the functional evaluations that are usually performed only before and after the training.

To conclude, in this work we proposed a simple parameter that is a useful tool to follow subjects’ performance during the training and thus to provide meaningful information with minimal additional effort for both patients and clinicians. These results confirm that variations in terms of P are coherent with functional clinical improvements, thus giving information on functional recovery. Moreover, P may be used not only to plan a rehabilitative therapy but also to customize and update the strategy patient by patient. Finally, the parameter may be used for the validation of the effectiveness of new exercises.

Further steps will be the increase of the patient group size and the construction of an age-matched healthy control group to normalize the parameter over the maximum performance achievable during the exercises and to analyze the importance of the learning process in the increments of P.

<table>
<thead>
<tr>
<th>Exercise</th>
<th>T_0 (#)</th>
<th>T_{1/2} (#)</th>
<th>T_1 (#)</th>
<th>p-value Friedman (#)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goalkeeper</td>
<td>0.25 (0.09)</td>
<td>0.33 (0.31)</td>
<td>0.66 (0.18)</td>
<td>5.4 10^{-4}</td>
</tr>
<tr>
<td>N Goalkeeper</td>
<td>12</td>
<td>12</td>
<td>10</td>
<td>*</td>
</tr>
<tr>
<td>Moorhuhn</td>
<td>0.32 (0.14)</td>
<td>0.45 (0.40)</td>
<td>0.72 (0.25)</td>
<td>*</td>
</tr>
<tr>
<td>N Moorhuhn</td>
<td>13</td>
<td>13</td>
<td>11</td>
<td>*</td>
</tr>
<tr>
<td>Stove Cleaning</td>
<td>0.25 (0.18)</td>
<td>0.53 (0.20)</td>
<td>0.54 (0.10)</td>
<td>0.014</td>
</tr>
<tr>
<td>N Stove Cleaning</td>
<td>12</td>
<td>12</td>
<td>8</td>
<td>*</td>
</tr>
<tr>
<td>Fruit Shopping</td>
<td>0.45 (0.12)</td>
<td>0.43 (0.26)</td>
<td>0.62 (0.24)</td>
<td>*</td>
</tr>
<tr>
<td>N Fruit Shopping</td>
<td>11</td>
<td>11</td>
<td>9</td>
<td>*</td>
</tr>
<tr>
<td>Egg Cracking</td>
<td>0.13 (0.19)</td>
<td>0.16 (0.14)</td>
<td>0.26 (0.12)</td>
<td>4.8 10^{-4}</td>
</tr>
<tr>
<td>N Egg Cracking</td>
<td>14</td>
<td>14</td>
<td>11</td>
<td>*</td>
</tr>
<tr>
<td>Reveal Panorama</td>
<td>0.10 (0.08)</td>
<td>0.17 (0.15)</td>
<td>0.30 (0.47)</td>
<td>0.002</td>
</tr>
<tr>
<td>N Reveal Panorama</td>
<td>13</td>
<td>12</td>
<td>10</td>
<td>*</td>
</tr>
<tr>
<td>Chase Balloon</td>
<td>0.26 (0.15)</td>
<td>0.47 (0.24)</td>
<td>0.68 (0.26)</td>
<td>0.001</td>
</tr>
<tr>
<td>N Chase Balloon</td>
<td>13</td>
<td>12</td>
<td>10</td>
<td>*</td>
</tr>
<tr>
<td>Vertical Catching</td>
<td>0.35 (0.11)</td>
<td>0.50 (0.15)</td>
<td>0.54 (0.14)</td>
<td>4.9 10^{-4}</td>
</tr>
<tr>
<td>N Vertical Catching</td>
<td>13</td>
<td>13</td>
<td>10</td>
<td>*</td>
</tr>
<tr>
<td>P</td>
<td>0.29 (0.11)</td>
<td>0.44 (0.22)</td>
<td>0.54 (0.24)</td>
<td>1.5 10^{-5}</td>
</tr>
<tr>
<td>N</td>
<td>14</td>
<td>14</td>
<td>12</td>
<td>*</td>
</tr>
</tbody>
</table>
Acknowledgment

We would like to thank the patients that underwent the clinical trial and the physiotherapists who performed the training with the patients.

References