Predicting Upper-Extremity Function Recovery From Kinematics in Stroke Patients Following Goal-Oriented Computer-Based Training

Fabrizio Antenucci (fabrizio.antenucci@saddlepointscience.com)
CNR: Consiglio Nazionale delle Ricerche https://orcid.org/0000-0001-5070-7408
Belén Rubio Ballester
IBEC: Institut de Bioenginyeria de Catalunya
Martina Maier
IBEC: Institut de Bioenginyeria de Catalunya
Anthony C.C. Coolen
Radboud Universiteit Nijmegen: Radboud Universiteit
Paul F. M. J. Verschure
IBEC: Institut de Bioenginyeria de Catalunya

Research Article

Keywords: Rehabilitation, Stroke, Interactive feedback, Upper extremities, Posture monitoring, Motion sensing, Motion classification, Multivariate regression

Posted Date: June 18th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-591866/v1

RESEARCH

Predicting upper-extremity function recovery from kinematics in stroke patients following goal-oriented computer-based training

Fabrizio Antenucci[2*], Belén Rubio Ballester[1*], Martina Maier[1*], Anthony C. C. Coolen[2*] and Paul F. M. J. Verschure[1,3*]

Abstract

Background: After a stroke, a wide range of deficits can occur with varying onset latencies. As a result, predicting impairment and recovery are enormous challenges in neurorehabilitation. Body function and structure, as well as activities, are assessed using clinical scales. For functional deficits of the upper extremities these include the Fugl-Meyer Assessment for the Upper Extremity (FM-UE), the Chedoke Arm and Hand Activity Inventory (CAHAI) and Barthel Index (BI), administered by clinicians. Although these scales are generally accepted for the evaluation of the motor and functional impairment of the upper-limbs, they are time-consuming, show high inter-rater variability, have low ecological validity, and are vulnerable to biases introduced by compensatory movements and action modifications. For these reasons, alternative methods need to be developed for efficient and objective assessment. Computer-based motion capture and classification tools have the potential to collect and process kinematic data to estimate impairment, function and recovery while overcoming these limitations.

Methods: We present a method for estimating clinical scores from movement parameters that are entirely extracted from kinematic data recorded during unsupervised rehabilitation sessions performed with the Rehabilitation Gaming System (RGS). RGS is a rehabilitation technology that uses image-based motion capture, goal-oriented individualised training, virtual reality content delivery, and restricts compensatory trunk movements through feedback. The main protocol considered in this study asks patients to use their upper limbs to intercept spheres that are presented in a 3 dimensional virtual reality display. RGS maps the planar physical arm movements onto matching movements by an avatar presented in a first-person perspective. In this analysis, we performed a multivariate regression using clinical data from 98 stroke patients who completed a total of 191 sessions with RGS.

Results: Our multivariate regression model reaches an accuracy of R^2 : 0.38, with an error (σ : 12.8), in predicting FM-UE scores. We analyse our model by assessing reliability ($r = 0.89$ for test-retest), sensitivity to clinical improvements (95% true positive rate) and generalisation to other tasks that involve planar reaching movements (R^2 : 0.39). The model achieves comparable accuracy also for the CAHAI (R^2 : 0.40) and BI scales (R^2 : 0.35).

Conclusions: Our results highlight the clinically relevant predictive power of kinematic data collected during unsupervised goal-oriented motor training combined with automated inference techniques and provide new insight into factors underlying recovery and its biomarkers.

Keywords: Rehabilitation; Stroke; Interactive feedback; Upper extremities; Posture monitoring; Motion sensing; Motion classification; Multivariate regression

*Correspondence: fabrizio.antenucci@saddlepointscience.com; belen.rubio.ballester@gmail.com; martinamaier@gmx.ch; ton.coolen@saddlepointscience.com; pverschure@ibecbarcelona.eu
[2]Saddle Point Science Ltd, 10 Lincoln Street, York, UK
[1]Laboratory of Synthetic, Perceptive, Emotive and Cognitive Systems (SPECS), Institute for Bioengineering of Catalonia (IBEC), The Barcelona Institute of Science and Technology (BIST), Baldiri Reixac 10-12, 08028, Barcelona, Spain
[3]Institució Catalana de Recerca, Estudis Avançats (ICREA), Barcelona, Spain
Full list of author information is available at the end of the article

Background

Stroke is the second major cause of death and disability worldwide, with about 15 million new cases every year [1]. One third of these cases lead to persistent cognitive and motor disabilities [2]. About 80% of stroke survivors present weakness and partial loss of voluntary control in the upper-extremities [3], or hemiparesis, which is often associated with other sensorimotor alterations, such as hypertonia or tremor.

Although hemiparesis is a highly prevalent symptom and severely limits the independence of affected patients, its causes and recovery dynamics are not fully understood [4]. Recent literature converges onto the core idea that it is mainly due to a combination of residual corticospinal tract capacity and an upregulation of the reticulospinal tract [5, 6]. Further, recovery seems to follow a temporal structure where most of the improvement occurs during the first months post-stroke [7, 8]. However, one dilemma in understanding the mechanism of hemiparesis and its recovery is that it is based on assessment methods such as the Fugl-Meyer and ARAT scales which have in turn their own limitations. Indeed, a recent systematic review [9] investigating a total of 225 studies (N=6197) using 151 different kinematic metrics found that kinematic assessments of upper limb sensorimotor function are poorly standardised and rarely investigate clinimetrics in an unbiased manner. Specifically, using descriptors of accuracy, efficacy, efficiency, movement planning, precision, spatial posture, speed, temporal posture, and range of movement together with clinimetric properties of these descriptors (i.e., reliability, measurement error, convergent validity, and responsiveness), the authors show that the studies analysed exclusively focused on finding correlations between measures of impairment, and only two of the studies reported explicit responsiveness metrics such as correlations in change. Actually, both cross-sectional and longitudinal construct validity have been supported by the relationship between kinematic measures of reaching and the degree of sensorimotor impairment or scores obtained with clinical rating scales [10]. However, there is very limited information regarding test-retest reliability, i.e reproducibility, and responsiveness of kinematic outcome measures of reaching performance.

In order to advance our knowledge about the hemiparesis phenotype and its progression it seems necessary to find alternative methods and measures of motor function and recovery that are objective, reliable and sensitive. One proposal by Murphy et al. [11] explored responsiveness in a number of kinematic descriptors and found a significant covariation of the ARAT scores with movement time (R^2=0.36), smoothness (R^2=0.31), and trunk displacement (R^2=0.35). Although these results are promising, the study involves a limited number of subjects (N=24) from an highly homogeneous sample (i.e., acute patients only). Further, the ARAT clinical scale presents poor robustness to compensation and is especially vulnerable to the use of explicit strategies to boost performance. Majeed et al. [12] explored the application of models based on LASSO regression to predict changes in motor ability (FM-UE) and motor function (WMFT). These models propose that recovery in both scales can be approximated by the patient's age, the patient's motor control during the execution of fast movements, and other demographic and clinical features, altogether accounting for 65% and 86% of the variability for the FM-UE and WMFT scales respectively. Although these models reach exceptional accuracy, their utility is limited because they make use of kinematic data obtained during the supervised execution of very specific pointing movements, and are based on generic unbounded linear models, with the consequence that their predicted values could be largely outside the meaningful range of the scale (e.g. predict FM-UE scores larger than 66 points).

We propose a new approach towards using kinematic data obtained in unsupervised rehabilitation sessions to predict level of impairment and functional recovery. Data is obtained from patients engaging with goal-oriented embodied individualised training with the Rehabilitation Gaming system (RGS). We first explore the clinimetric properties of hand movements that were collected during unsupervised RGS sessions. Secondly, we build and analyse the performance (i.e., external validity, robustness, responsiveness, sensitivity, and generalisation) of a predictive model of motor recovery (1-3 months post-baseline) based on data of acute to chronic stroke survivors.

Methods

Subjects

Our retrospective analysis uses data of 191 RGS sessions from 98 hemiplegic patients (age in range [23, 87], mean 63; days post-stroke in range [5, 3045], mean 400; cf. Table 1) who were recruited between 2010 and 2015 to participate in studies conducted in Barcelona and Tarragona, Spain [13, 14, 15]. Participants met the following inclusion criteria: 1) ischemic strokes (middle cerebral artery territory) or hemorrhagic strokes (intracerebral); 2) mild-to-moderate upper limb hemiparesis (Medical Research Council scale for proximal muscles > 2) after a first-ever stroke; 3) age between 20 and 90 years old; and 4) the absence of any significant cognitive impairment (Mini-Mental State Evaluation > 22).

Table 1 Characteristics of the 191 samples composing the main dataset (single session, Spheroids scenario). The r columns refer to the Pearson correlation coefficients with the FM-UE, CAHAI, and BI clinical scales, respectively. Correlations below the significance threshold $r \sim 0.081$ (cf. Fig. 12 in Appendix) are in grey. The clinical scales are measured no more than 4 days before or after the coupled RGS session. The 'instantaneous' variables (6-12) obtained directly from RGS log-files are in Italic type. The work area is computed as the area of the complex hull of the hand movements using standard methods, e.g. Jarvis' Algorithm [16]. The distance covered refers to the total length of the hand paths during training. The performance success rate is defined as the number of spheres intercepted over the total released during the RGS session. The 'Smoothness' and 'TGDM' are defined in the main text, cf. Sec. 'Identification of variables'.

Variables	Range [min, max]	Mean	SD	r(FM-UE)	r(CAHAI)	r(BI)	Missing
FM-UE score	[4, 66]	43	16	1	0.89	0.34	-
CAHAI score	[13, 91]	52	26	0.89	1	0.50	-
BI score	[10, 100]	80	21	0.34	0.50	1	15
1. Gender	female/male	73/118	-	0.17	0.11	0.036	-
2. Age	[23, 87]	63	12.8	-0.015	-0.10	-0.30	-
3. Dominant side more affected	yes/no	72/119	-	-0.013	0.039	0.21	-
4. Time since stroke (days)	[5, 3045]	400	625	-0.25	-0.17	0.22	-
5. Sessions completed so far	[2, 49]	10	11	0.31	0.38	0.32	-
6. Work area (m^2)	[0.011, 1.8]	0.38	0.35	0.29	0.27	0.14	-
7. Distance covered (m)	[2.6, 240]	56	34	0.18	0.21	0.26	-
8. Performance (% success)	[0.37, 0.94]	0.68	0.105	0.33	0.37	0.33	-
9. Maximum reaching speed (m/s)	[2.8, 88]	18	16	0.17	0.13	0.061	-
10. Difficulty level reached	[-0.16, 0.89]	0.46	0.23	0.45	0.52	0.46	-
11. Smoothness (mm)	[0.17, 3.7]	1.2	0.55	0.42	0.39	0.28	-
12. TGDM (m)	[0.011, 0.12]	0.062	0.021	0.52	0.58	0.43	-

Protocol

Participants followed a rehabilitation protocol including 3-5 weekly sessions of 30 minutes each for 3-12 weeks using the Rehabilitation Gaming System (RGS) shown in Fig. 1. The joint movements of the user's head, shoulders and elbows are tracked and mapped onto an avatar through a biomechanical model using a custom developed vision based motion capture system. Arm movements are displayed on a screen from a first-person perspective, realising a rehabilitation paradigm that combines goal-oriented embodied and situated action execution, motor imagery, and action observation.

For the RGS sessions in the main dataset, cf. Table 1, the participants are instructed to intercept virtual spheres that move towards them by executing horizontal bimanual movements over the surface of a table ('Spheroids' protocol). The task parameters (the frequency of sphere appearance, their speed, their range, and size) are combined in a single parameter ('difficulty level') and automatically adjusted during the session in order to maintain the user's performance between 70% and 80% success rate [17, 18]. The system allows for the storage and extraction of performance parameters as well as hand path trajectories derived from joints' positions and rotations recorded at about 100 Hz.

During the rehabilitation patients are evaluated using standard clinical scales: Fugl-Meyer Assessment for the upper extremity (FM-UE), Chedoke Arm and Hand Activity Inventory (CAHAI) and Barthel Index (BI). When collecting the 191 samples (Table 1), the following measures are taken to improve data quality:

- The clinical score measurements (FM-UE, CAHAI, BI) are coupled to the RGS session closest in time, with a maximum time separation of 4 days between the measurement and the RGS session;
- The first two RGS sessions of a patient at the start of the rehabilitation trajectory are excluded to ensure that patients are familiar with the RGS environment for all collected samples.

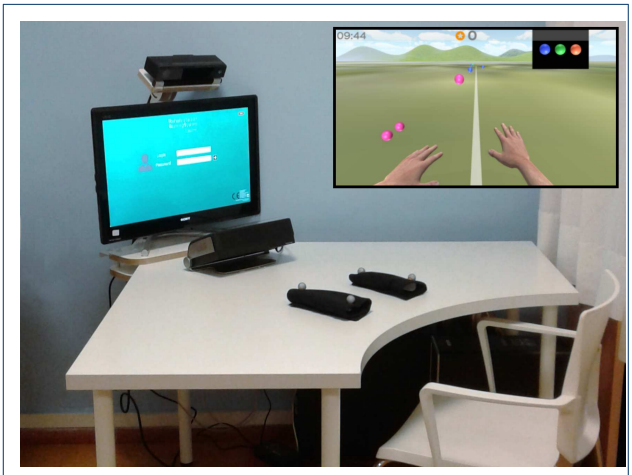

Figure 1 The Rehabilitation Gaming System (RGS). The system consists of a PC, a 17 inch LCD touch display, a image-based motion capture device (Kinect 360, Microsoft) positioned on top of the screen [14]. In the inset we show a screenshot from the 'Spheroids' activity during a RGS session. The virtual tasks logic and graphics are implemented using the Torque 3D and Unity 3D game engines.

Outcomes and Analysis

To analyse the potential of RGS-derived movement descriptors for capturing both impairment and recovery in standardised clinical scales, we first extract a set of

variables that are known to correlate with the severity of hemiparesis [9, 12]. Next, to execute a convergent validation analysis, we generate a model that combines the information of several variables to estimate the patient's score on a clinical scale. The model includes an estimation of the prediction error. We use repeated cross-validation to avoid overfitting, allocating 50% of the samples to the training set and 50% to the validation set. We conduct a robustness analysis in order to explore test-retest reliability. Additionally, we perform a sensitivity analysis by computing correlations between the change in the movement descriptors and clinical improvements. For this analysis we only consider pairs of RGS sessions from the same patient for which the time-difference between the two sessions is larger than 16 days, for a total of 54 samples (108 RGS sessions and their associated demographical descriptors and clinical assessments).

Finally, we explore the potential of the model to generalise to other tasks that involve bimanual 2D planar reaching movements. To study this last property we identify a second dataset of 37 samples from a previous study in which 19 subacute stroke patients with hemiparesis were rehabilitating with a different RGS training scenario which is a variation on the well known arcade game 'Whac-A-Mole' [14].

The prediction analysis we present will focus on the FM-UE score, and we will briefly discuss the generalisation to the CAHAI and BI scales.

Identification of variables

We consider 31 variables in total, 12 are first order variables while the remaining 19 are second order variables and obtained as functions of the first order ones.

We identify two groups of first order variables: 1) demographic and physiological data at recruitment, cf. variables 1-5 in Table 1, and 2) kinematic descriptors extracted for the more affected limb during training sessions, cf. variables 6-12 in Table 1. For the evaluation of all kinematic descriptors, the first and last two minutes of each training session are discarded to avoid the interference of behaviours or events related to the start and the ending of the training session (revision of instructions, postural adjustments, exposure to the final score screen, etc.).

Second order variables include chronicity (i.e. acute, sub-acute and chronic categories) and the difference between the less and the more affected upper-limb in each of the quantitative first order variables, as well as their logarithmic transformations. The descriptive statistics of second order variables are given in Table 5 in the Appendix.

Among the above mentioned variables, most are obvious and/or well-known [9] (cf. caption of Table 1). However, we introduce also the new descriptors 'Smoothness' and 'Total-goal direct movement' (TGDM). Specifically, to extract information on UE motor function we introduce an original kinematic descriptor, $J(\sigma)$, to assess the patient's movements at a specific temporal resolution, σ. This metric allows us to isolate goal-oriented movements from the hand trajectory in a certain direction, assumed to be stored over time as a function $f(t)$. For the main dataset, we have considered the left/right direction, as it is the principal axes in the movement dynamics of 'Spheroids' protocol. We assume measurements are taken at discrete time points $t_i = i\Delta$ for $i = 0, 1, ..., T-1$, with Δ being the timestep (for the 'Spheroids' dataset we have $\Delta \simeq 0.01$ s). We define the total hand displacement during goal-oriented movements $J(\sigma)$ as the difference between the actual movements and a smoothened version of the discrete movements. The smoothened hand path $f_\sigma(t)$ is obtained using a Gaussian smoothing process with parameter σ

$$f_\sigma(t_i) = \frac{\sum_{j=0}^{T-1} f(t_j) \exp\left[-\frac{(t_i-t_j)^2}{2\sigma^2}\right]}{\sum_{j=0}^{T-1} \exp\left[-\frac{(t_i-t_j)^2}{2\sigma^2}\right]} \tag{1}$$

where the parameter σ defines how smooth the new trajectory will be, see an example in Fig. 2. Therefore $J(\sigma)$ is obtained as

$$J(\sigma) = \sqrt{\frac{\sum_{i=0}^{T-1} \left[f_\sigma(t_i) - f(t_i)\right]^2}{T}}\,. \tag{2}$$

Following this analysis method we derive the two new variables corresponding to the value of $J(\sigma)$ at the two peaks in the σ-dependent Pearson correlation with the clinical scales, cf. Fig. 3: 'Smoothness' in correspondence to the high-frequency peak, and 'TGDM' in correspondence to the low-frequency peak. The location of two peaks is weakly dependent on the clinical scale considered, yet it appears to be related to the data structure: the high-frequency peak is linked to the time resolution of the data ($\Delta \simeq 0.01$ s), while the low-frequency peak is related to the typical timescale of the Spheroids protocol, i.e. a spheroid is launched every ~ 10 s and moving towards opposite sides of the tasks space. We will further interpret these two new variables in the following analysis.

Predictive Models

To combine variables for the prediction of clinical scores of impairment and recovery, we introduce a model that allows for the presence of noise on both the variables $\boldsymbol{Z}$ and the score S, and we hence name it a double-noise parametric model. Its generative func-

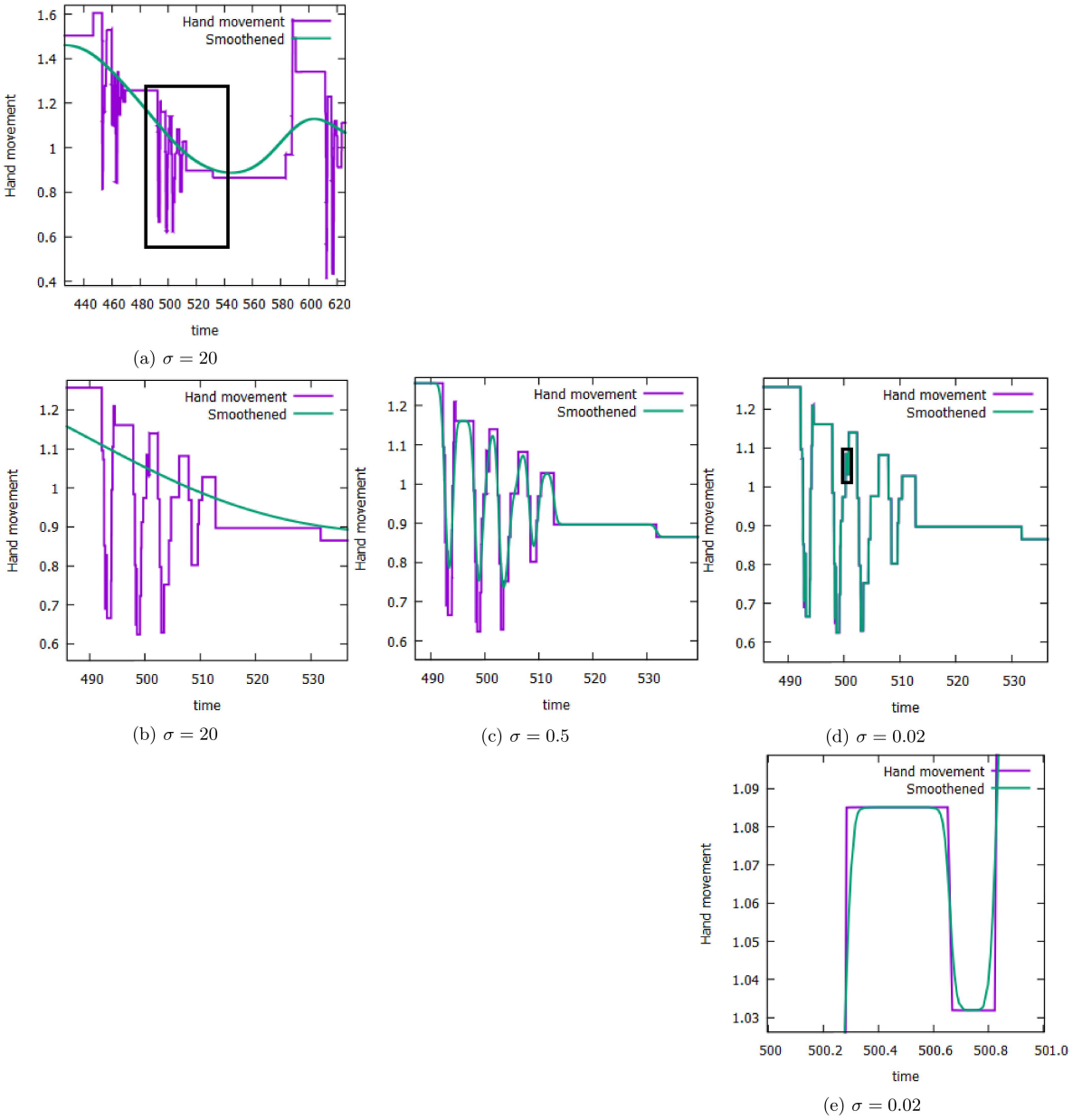

Figure 2 An example of the smoothed trajectory $f_\sigma(t)$ Eq. 1 (green line) for lateral (left/right) direction hand movements with $\sigma = 20s$ (left) $\sigma = 0.5s$ (center) and $\sigma = 0.02s$ (right) compared to the real trajectory as recorded by the camera (purple line).

tional form is

$$S(\boldsymbol{Z}|\boldsymbol{\theta}) = \frac{B-A}{2}\tanh\left(\boldsymbol{\beta}\cdot\boldsymbol{Z}+\beta_0+\sigma_1 u\right)+ \\ + \frac{B+A}{2}+\sigma_2 v \qquad (3)$$

where $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \beta_0, A, B, \sigma_1, \sigma_2\}$ are the $p+5$ model hyperparameters to be inferred: $\boldsymbol{\beta}$ (association parameters of p active variables), β_0 (parametric offset), A e B (range offsets), σ_1 and σ_2 (noise strengths). The sources of noise u (the covariate noise) and v (the score noise) are both assumed to be standard normally distributed. Since $\tanh(-x) = -\tanh(x)$, we remove the resulting parameter sign ambiguity by enforcing $B \geq A$. Note that the saturation of the sigmoidal function captures the boundedness of the clinical scores, so that the average over the noise of the predicted score S is constrained in the interval $[A, B]$.

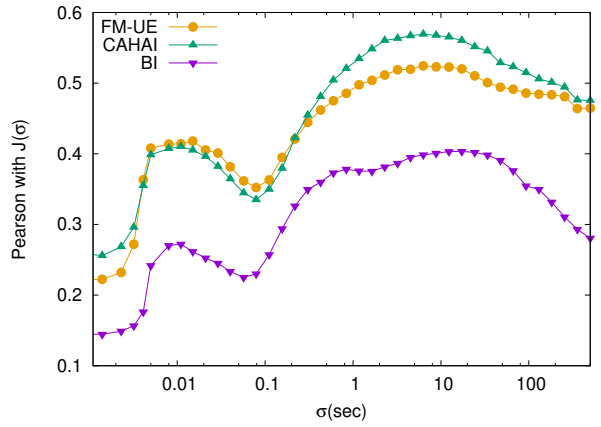

Figure 3 Pearson correlation coefficient between clinical scales (FM-UE, CAHAI, BI) and $J(\sigma)$ (cf. Eq. (2)) obtained for the lateral (left/right) direction of the paretic hand trajectory at different values of σ for the 'Spheroids' dataset (cf. Table 1). The high-frequency peak is at about $\sigma = 0.01$ s for all three scales. The low-frequency peak is at $\sigma = 8.8$ s for FM-UE, $\sigma = 5.9$ s for CAHAI, and $\sigma = 17$ s for BI.

The probability of a particular score S, given the variables and the hyperparameters, is given by

$$p(S|\boldsymbol{Z},\boldsymbol{\theta}) \tag{4}$$
$$= \frac{1}{\sigma_2\sqrt{2\pi}} \int Dv \; \mathrm{e}^{-\frac{1}{2\sigma_2^2}[S - a\tanh(\boldsymbol{\beta}\cdot\boldsymbol{Z}+\beta_0+\sigma_1 v)-b]^2}$$

with the short-hands $Dv = (2\pi)^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{2}v^2}dv$, $a = (B-A)/2$, $b = (B+A)/2$. We use the following (improper) prior distribution over the parameters $\boldsymbol{\theta}$:

$$p(\boldsymbol{\theta}) = Z^{-1}\mathrm{e}^{-\frac{1}{2}d\boldsymbol{\beta}^2} p(\sigma_1)\, p(\sigma_2). \tag{5}$$

We take $p(\sigma_1) \sim \mathrm{e}^{-q/\sigma 1}$ and similar for σ_2 with q being a very small number, typically of the order of the accuracy of the numerical work, e.g. $q \sim 10^{-10}$. These priors guarantee that the log-likelihood function is bounded from below.

We adopt Maximum A Posteriori (MAP) inference: given the dataset $\mathcal{D} = \{(\boldsymbol{Z}_1, S_1), \ldots, (\boldsymbol{Z}_n, S_n)\}$, the optimal parameters $\boldsymbol{\theta}$ correspond to the maximum of the posterior probability or, alternatively, to the minimum of the regularised log-likelihood function:

$$\Omega(\boldsymbol{\theta}) =$$
$$-\sum_{i=1}^{n} \log \int Dz \; \mathrm{e}^{-\frac{1}{2\sigma_2^2}[S_i - a\tanh(\boldsymbol{\beta}\cdot\boldsymbol{Z}_i+\beta_0+\sigma_1 z)-b]^2}$$
$$+ \frac{1}{2}d\boldsymbol{\beta}^2 + n\log(\sigma_2) + \frac{q}{\sigma_2} + \frac{q}{\sigma_1}\,. \tag{6}$$

The errors on the inferred parameters $\boldsymbol{\theta}$ are estimated from the curvature of the regularised log-likelihood function at the minimum.

Two simpler models derived from (6) can be considered corresponding to having either score noise only or covariate noise only:

- *Score noise model*, $\sigma_1 = 0$. Taking the limit $\sigma_1 \to 0$ in (6) gives

$$\Omega_{\mathrm{sco}}(\boldsymbol{\theta}) =$$
$$\frac{1}{2\sigma_2^2}\sum_{i=1}^{n}[S_i - b - a\tanh(\boldsymbol{\beta}\cdot\boldsymbol{Z}_i+\beta_0)]^2 +$$
$$+ n\log(\sigma_2) + \frac{1}{2}d\boldsymbol{\beta}^2 + \frac{q}{\sigma_2}, \tag{7}$$

- *Covariate noise model*, $\sigma_2 = 0$ Taking the limit $\sigma_2 \to 0$ in (6) gives

$$\Omega_{\mathrm{cov}}(\boldsymbol{\theta}) =$$
$$\frac{1}{2\sigma_1^2}\sum_{i=1}^{n}\left[\tanh^{-1}\left(\frac{S_i-b}{a}\right) - \boldsymbol{\beta}\cdot\boldsymbol{Z}_i - \beta_0\right]^2 +$$
$$+ \sum_{i=1}^{n}\log\left[a^2-(S_i-b)^2\right] + n\log\left(\frac{\sigma_1}{a}\right) +$$
$$+ \frac{1}{2}d\boldsymbol{\beta}^2 + \frac{q}{\sigma_1} \tag{8}$$

  provided $|S_i - b| < a$ for all i, otherwise $\Omega_{\mathrm{cov}}(\boldsymbol{\theta}) = \infty$. This implies that for each b the minimisation over a is to be carried out strictly over the open interval $a > \max_i |S_i - b|$.

Results

Identification of features

To explore the convergent validity of RGS-derived kinematic descriptors in comparison to standardised clinical assessments, we compute Pearson correlations between all the variables (the RGS-derived descriptors and the baseline characteristics) and the clinical scores, cf. Fig. 4. By comparing these to a randomised outcome distribution we identify a significance threshold of $r \simeq 0.081$ for all the Pearson correlations variable-scores (cf. Fig. 12 in Appendix). Several of the kinematic descriptors, in particular 'TGDM' and 'difficulty level reached' ('difficulty') correlate highly with all clinical scales. Generally there is a high level of consistency between correlations of the variables with the different clinical scales examined. The main exceptions are 'age' and 'time since stroke' ('days-s'). The former does not display a relevant correlation with FM-UE and CAHAI scores but it is negatively correlated with BI ($r = -0.30$, $p < .0001$), while the latter correlates negatively with the FM-UE ($r = -0.25$,

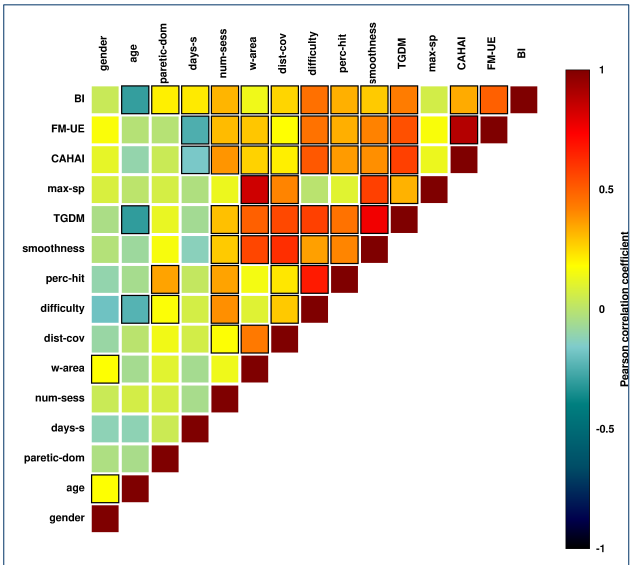

Figure 4 Correlogram of clinical scales and variables, cf. Table 1. The scale indicates the value of the correlation coefficients, going from -1 (full negative correlation) to 1 (full positive correlation). Black bordered squares indicate significant correlations ($p < 0.05$).

$p = .00049$) and CAHAI ($r = -0.17$, $p = .019$) scores but positively with BI ($r = 0.22$, $p = .0033$). The correlations between FM-UE and CAHAI scores is high ($r = 0.89$, $p < .0001$), while the consistency with BI is significantly lower: ($r = 0.34$, $p < .0001$ with FM-UE; $r = 0.5$, $p < .0001$ with CAHAI). Generally, the correlations of the kinematic descriptors are higher with FM-UE and CAHAI than with BI.

All kinematic variables show consistent inter-variables correlation. In particular, 'maximum reaching speed' ('max-sp') correlates highly with 'work area' ('w-area') ($r = 0.76$, $p < .0001$). This is interesting as these two variables are measured in a very different way, as the maximum reaching speed is obtained in a single instant while the work area is a function of the whole trajectory generated over the RGS session. The variable 'age' correlates negatively with 'TGDM' ($r = -0.31$, $p < .0001$) and 'difficulty' ($r = -0.24$, $p = .00098$), while it is uncorrelated with FM-UE and CAHAI scores, but not to BI ($r = -0.30$, $p < .0001$). This suggests that age-related technology proficiency may affect the kinematic descriptors. Nevertheless, this effect is relatively weak, i.e., the correlations of 'difficulty' and 'TGDM' with the clinical scores are significantly higher than the ones with 'age'.

In order to better understand the meaning of the two variables obtained with the smoothing techniques ('smoothness' and 'TGDM'), we also extract finite time-windowed variables for comparison. Specifically, we compute the maximum range of movement (left/right direction) within overlapping time windows of size σ, where σ is again fixed by the condition of maximum correlation with the clinical scale of interest, and we average the measurements across all possible windows along the whole RGS session. The resulting values show a very high correlation with the TGDM variable ($r = 0.98$, $p < 0.01$) and show very similar Pearson coefficients with the FM-UE ($r = 0.54$, $p < 0.01$), CAHAI ($r = 0.57$, $p < 0.01$) and BI ($r = 0.44$, $p < 0.01$) clinical scales. These results suggest that the TGDM is capturing information about the typical range of movement associated with the scenario events occurring within relevant time windows (i.e. about 10 seconds in the 'Spheroids' scenario). Following the same method, we extract time-windowed maximum reaching speed. The resulting values show a very high correlation with the smoothness variable ($r = 0.87$, $p < 0.01$) together with very similar Pearson coefficients with the FM-UE ($r = 0.42$, $p < 0.01$), CAHAI ($r = 0.39$, $p < 0.01$) and BI ($r = 0.21$, $p < 0.01$) clinical scales. These results suggest that smoothness is linked to the ability of the patient to perform fast movements to complete the RGS tasks.

External validity: prediction of instantaneous scores

In the following we combine information from several variables to estimate clinical scores associated with a single patient's RGS session.

In this section we will focus mostly on the FM-UE scale. The other clinical scales will be discussed in Sec. 'Generalisation to CAHAI and BI scales'. Nevertheless, it is useful to anticipate here that the results for CAHAI are very similar to FM-UE. This similarity is expected as the two scales have high relative correlations and comparable correlations to most of the variables, cf. Tables 1,5. To be quantitative, we can consider the case in which the FM-UE score S_i^{FM} of the 'Spheroids' dataset is estimated by simply rescaling the corresponding CAHAI value S_i^{CAHAI} by 66/91; this leads to the standard error

$$\frac{1}{191}\sqrt{\sum_{i=1}^{191}[(S_i^{\mathrm{FM}} - (66/91)S_i^{\mathrm{CAHAI}})^2]} \simeq 10.1 \qquad (9)$$

and a R^2 value of $1 - \sigma^2_{\mathrm{FM,CAHAI}}/\sigma^2_{\mathrm{FM}} \simeq 0.62$. The prediction of BI scores from kinematic descriptors is instead generally harder. This last point can be explained by the lower correlation of BI scores with most of the variables (cf. Tables 1,5). If we estimate the FM-UE score by a simple rescaling of the BI value by 66/100 we get a standard error of

$$\frac{1}{176}\sqrt{\sum_{i=1}^{176}[(S_i^{\mathrm{FM}} - (66/100)S_i^{\mathrm{BI}})^2]} \simeq 15.5 \qquad (10)$$

Table 2 The association parameters β of the optimal variable set for the prediction of 'instantaneous FM-UE' and 'FM-UE change' (ΔFM-UE), 'Spheroids' protocol. Note that for 'FM-UE change' the value of a variable refers to the difference between the two sessions. The values listed here refer to the normalised variables, so that the values of the different βs are directly comparable.

Covariate	β(**FM-UE**)	β(Δ**FM-UE**)
Difficulty	0.186(0.051)	0.194(0.062)
TGDM	0.049(0.087)	0.184(0.090)
Diff. Distance covered	0.027(0.044)	-0.220(0.080)
Diff. TGDM	-0.197(0.057)	-0.043(0.070)
Log. work area	0.073(0.059)	-0.108(0.064)
Log. smoothness	0.086(0.079)	-0.070(0.078)

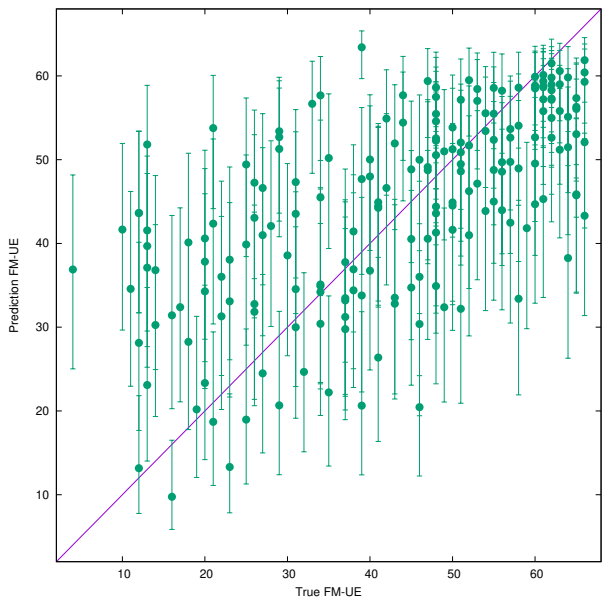

Figure 5 True FM-UE versus predicted FM-UE for the 191 samples of Table 1 ('Spheroids' scenario), using the covariate noise model with association parameters given in Table 2.

and a R^2 value of $1 - \sigma^2_{\mathrm{FM,BI}}/\sigma^2_{\mathrm{FM}} \simeq 0.12$. So we see again that BI carries different information than FM-UE (or CAHAI).

In the following, we adopt the covariate noise model for predictions of the FM-UE scale, cf. Eq. (8). Indeed we found that the three models presented in Sec. 'Predictive models' (double noise model, score noise model, covariate noise model) offer comparable performance in prediction of FM-UE scale on the dataset in Table 1. The typical error in the final prediction is of order ~ 10, while the inferred noise score error σ_2 is typically ~ 0.1. In this sense the score noise has little impact in this situation and so we prefer the covariate noise model to decrease the number of parameters to be inferred.

Here we aim to estimate the *instantaneous* FM-UE scores. By instantaneous we mean that we use only patient's baseline characteristics and RGS-derived movement descriptors (extracted from a single session logfile). In particular, we rule out the two variables 'session completed so far' and 'time since stroke' and the second order variables obtained from them. In this way, the prediction is intended to anticipate the clinical status of the patient at a given moment without knowledge of the rehabilitation history.

To avoid overfitting, we perform repeated cross-validation with 50% of samples for training and 50% for validation, obtaining the optimal active set of variables possible for our dataset[1]. The active variables are 6 and shown in Table 2. In total we have 10 parameters (6 association parameters and 4 hyperparameters) inferred from the 191 sessions. The most important variables for the prediction of the instantaneous FM-UE score are 'difficulty' and 'Diff. TGDM' (difference in TGDM between the non-paretic arm and the paretic arm). We note that the resulting active variable set does not contain patient's baseline variables and the estimation of scores on clinical scales are solely obtained from (unsupervised) RGS-derived data. This also means that predictions made by this model for different RGS sessions of the same patient are considered independent measurements. The hyperparameters of the model that predicts FM-UE are given by $a = 32.72(0.85)$, $b = 34.55(0.73)$, $\sigma_1 = 0.551(0.043)$ and $\beta_0 = 0.370(0.051)$. To evaluate the accuracy of the final regression model on unseen data, we consider the Leave-one-out cross-validation (LOOCV) for which we obtain an accuracy on the training set of 0.755 while the accuracy on the validation set is 0.746[2], so that the difference is about 1.2%.

Eventually, we compare the true and predicted score values, shown in Fig. 5. The model predicts the FM-UE score with an average error of $E_{\mathrm{FM}} \sim 12.8$, Pearson r true-predicted of 0.63 and a coefficient of determination $R^2 = 0.38$. These values are intermediate between an estimation based on simple rescaling of CAHAI and BI scores, cf. Eqs. (9),(10).

Robustness: test-retest reliability

To evaluate the test-retest reliability of our model for prediction of the instantaneous FM-UE score, we consider two unseen datasets each composed of 921 RGS sessions, for a total of 1842 unseen RGS sessions. Each session in the first dataset 'test' is associated with a session in the second dataset 'retest' and obtained from the same patient within less than 48 hours. The small

[1] We have enforced the presence of 'Diff. distance covered' in the active variable set since it is relevant in the prediction of clinical change, cf. Tables 3,6.

[2] Here we define the accuracy as the percentage of points that are correctly estimated above or below the median FM-UE score 47.

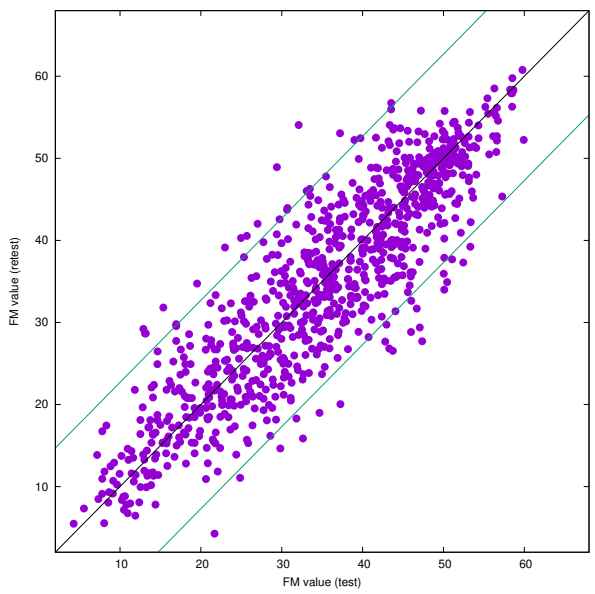

Figure 6 Predicted FM-UE score of first session (test) vs predicted FM-UE score of second session (retest) for unseen 921 couples of RGS sessions. Each couple is recorded from a same patient within 48 hours. The predictions are obtained using the covariate noise model with association parameters given in Table 2.

time frame makes it plausible that the clinical state of the patient is unchanged between the two test-retest sessions, and so they can be used to assess the reliability of the regression models. These data were collected in the same trials as the main dataset, but they correspond to rehabilitation sessions for which we do not have an associated measurement of a clinical scale (so they cannot be used for training).

The Pearson correlation between the 'test' and 'retest' data sets is 0.89, that is "Good correlation" according to Cronbach's alpha score, cf. Fig. 6. In Fig. 6 we also show the interval defined by the standard error of the regression $E \simeq 12.7$: the large majority of values are within this interval. Indeed we measure an average retest error $\sqrt{\sum_i (S_i^{\text{test}} - S_i^{\text{retest}})^2/N}$ equal to 5.9. This value gives an upper bound to the real value and it is less than half of the average error observed on the true score prediction.

These results support the internal consistency of this assessment method to predict the instantaneous clinical scores of the patients.

Responsiveness: prediction of improvement

Starting from the original dataset ('Spheroids', Table 1), we design a new dataset ('responsiveness' dataset, Table 3) composed of 54 samples where each sample represents a couple of sessions of the same patient for which the time lapsed between the two is larger than 16 days. We observe that the Pearson correlation between change in FM-UE (ΔFM-UE) and change in CAHAI (ΔCAHAI) is $r = 0.68$, Pearson between ΔFM-UE and change in BI (ΔBI) is $r = 0.67$, while the Pearson between ΔCAHAI and ΔBI is $r = 0.72$. For each sample we consider now as variables the change (between the two sessions) of the original variables.

By comparing it to a randomised outcome distribution, we identify a significance threshold of $r \sim 0.14$ for the Pearson correlations variable-scores. Several of the variables correlate highly with the change in all three clinical scores, cf. Fig. 7. The highest correlated variable is 'Change in TGDM' ($r = 0.48$, $p = .00024$ with ΔFM-UE; $r = 0.55$, $p < .0001$ with ΔCAHAI; $r = 0.49$, $p = .0011$ with ΔBI). Note that, in comparison to the prediction of a single session's score, the variables 'age' and 'chronic' are more correlated with the outcome when predicting the score change (cf. Tables 1,3 and Tables 5,6 in the Appendix). The initial scores (the clinical scores at the first session) have high correlation with change because of ceiling effect (cf. Table 6 in Appendix). In Fig. 7 we show the correlogram of all the variables and clinical scales. We observe that generally the correlations between kinematic descriptors in a single session (cf. Fig. 4) are preserved also when considering the change between sessions; for example the highest inter-variables correlation is for 'Change in w-area' and 'Change in max-sp' at ($r = 0.87$, $p < .0001$).

We utilise this dataset to analyse the responsiveness of the previous model (obtained for predictions of instantaneous scores) in detecting changes of clinical status in the same patient. We therefore adopt here the same set of active variable used for the prediction of instantaneous clinical scores. The association parameters of the model that predicts ΔFM-UE are shown in Table 2. The corresponding model hyperparameters are $a = 23.3(3.3)$, $b = 20.9(3.2)$, $\sigma_1 = 0.430(0.060)$ and $\beta_0 = -0.68(0.12)$.

We compare the true and predicted ΔFM-UE in Fig. 8. The Pearson correlation between true ΔFM-UE and predicted ΔFM-UE is 0.76. The value of the coefficient of determination R^2 is 0.57. These results show that the model has a good responsiveness to clinical change with a precision comparable to the one obtained for the instantaneous score.

Sensitivity: prediction of recovery

In order to evaluate the sensitivity (i.e. the true positive rate) of the model to predict ΔFM-UE we first identify 38 out of the 54 samples from the 'responsiveness' dataset (cf. Table 3) for which the associated ΔFM-UE values exceed an MDC of 4 points.

Table 3 Characteristics of the 54 samples composing the responsiveness dataset. We select a number of sessions of same patient with a delay of at least 16 days from the main dataset, cf. Table 1. The last three columns report the Pearson coefficient correlation between the variable and the change of clinical score between the two session. Correlations below the significance threshold $r \sim 0.14$ are in grey. Characteristics of second order variables for this dataset are shown in Table 6 in Appendix.

Variables	Range [min,max]	Mean	SD	$r(\Delta$FM-UE$)$	$r(\Delta$CAHAI$)$	$r(\Delta$BI$)$	Missing
Change in FM-UE score	[-2, 35]	9.1	9.1	1	0.68	0.67	-
Change in CAHAI score	[-1, 75]	25	19	0.68	1	0.72	-
Change in BI score	[-6, 69]	19	22	0.67	0.72	1	13
1. Gender	female/male	24/30	-	-0.0046	0.14	0.29	-
2. Age	[42, 84]	65	13	-0.25	0.061	0.051	-
3. Dominant side more affected	yes/no	24/30	-	-0.21	0.0065	-0.043	-
4. Time since stroke (days)	[1.5, 3.4]	2.1	0.46	-0.35	-0.37	-0.42	-
5. Sessions completed so far (at first)	[5, 46]	23	14	-0.45	-0.33	-0.46	-
6. *Change in work area* (m^2)	[-1.2, 1.2]	0.066	0.43	0.15	0.048	0.12	-
7. *Change in distance covered* (m)	[-68, 75]	14	23	0.21	0.22	0.19	-
8. *Change in performance (% success)*	[-0.17, 0.40]	0.076	0.11	0.075	0.12	0.22	-
9. *Change in max. reaching speed* (m/s)	[-51, 77]	3.3	21	0.087	0.048	0.063	-
10. *Change in difficulty*	[-0.12, 0.91]	0.21	0.20	0.38	0.39	0.36	-
11. *Change in smoothness* (mm)	[-0.51, 1.8]	0.22	0.50	0.37	0.35	0.39	-
12. *Change in TGDM* (m)	[-0.046, 0.14]	0.024	0.035	0.48	0.55	0.49	-

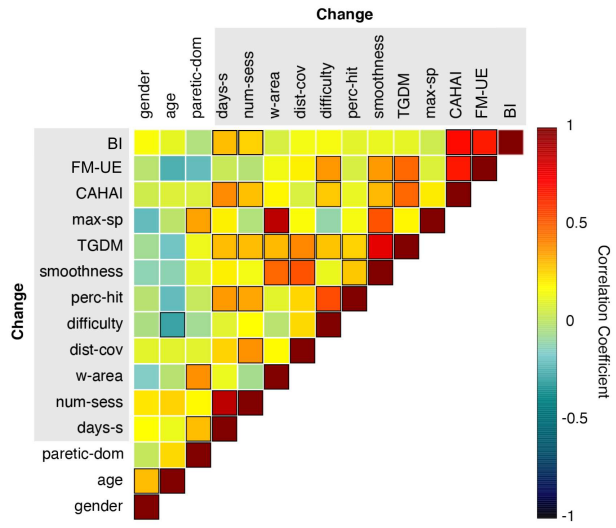

Figure 7 Correlogram of change of clinical scale and change in variables between couples of sessions of the same patient, cf. Table 3. The scale indicates the value of the correlation coefficients, going from -1 (full negative correlation) to 1 (full positive correlation). Black bordered squares indicate significant correlations ($p < 0.05$).

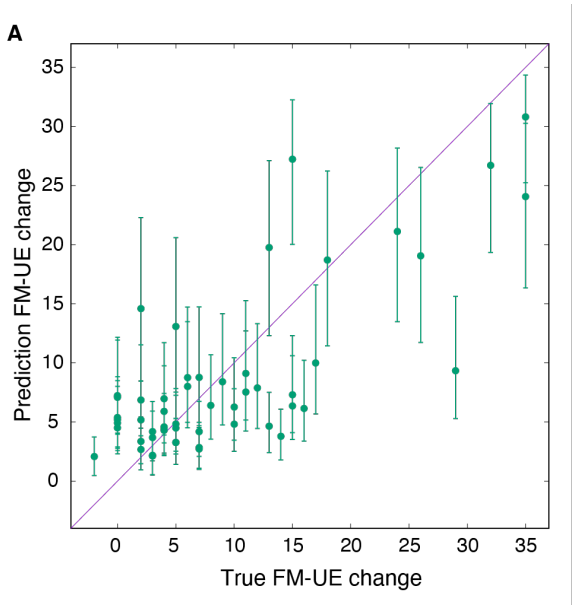

Figure 8 True versus predicted ΔFM-UE for 54 data points of 'responsiveness' dataset (cf. Table 3), using the covariate noise model with association parameters given in Table 4.

Given this subsample, with mean ΔFM-UE of 12.4 (SD: 9.4), our model predicts recovery in 36 of the cases (ΔFM-UE ≥ 4), indicating a TPR of 95% (mean of predicted ΔFM-UE is 10.3, SD: 7.9), cf. Fig. 8.

Generalisation

For the generalisation analysis, we consider an additional 37 RGS sessions from 19 hemiparetic participants that trained in a different 2D VR-based motor rehabilitation protocol derived from 'Whac-A-Mole' [14]. The observed FM-UE scores in this dataset are in the range $[5, 60]$, with mean 37 and SD 14.

Unlike the 'Spheroids' protocol, the gameplay of 'Whac-A-Mole' requires movements on the full 2d plane. In response, we utilise the smoothing technique in both cardinal axes of the task. i.e. front/back and left/right directions. Pearson correlations between the clinical scales and the variable $J(\sigma)$ reveal a similar pattern to the one observed in the Spheroids scenario, with a peak of the Pearson coefficients at about $1s$ corresponding to the variable TGDM in each direction (cf. Fig. 13 in Appendix). The location of the main peak is again close to the typical timescale of the protocol (that is faster than 'Spheroids'). For the FM-UE

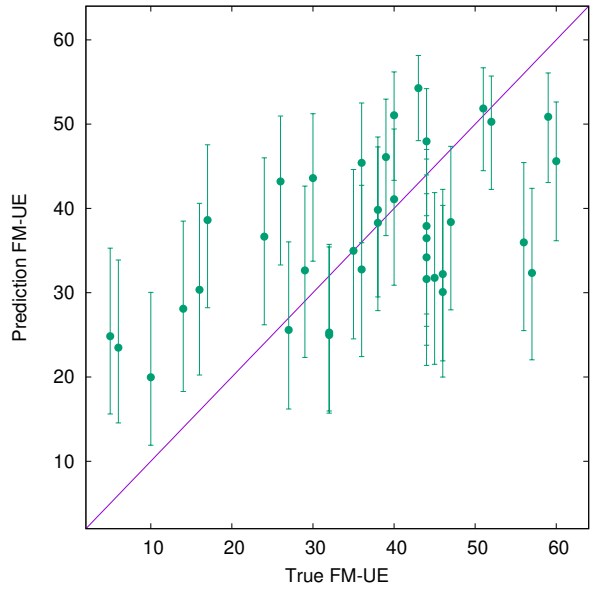

Figure 9 True versus predicted FM-UE for 37 samples of the 'Whac-A-Mole' dataset, using the covariate noise model with two active variables (TGDMfb and TGDMlr).

score, the highest Pearson coefficient is observed in the frontal direction ($r = 0.54$ for $\sigma = 1.3s$); the lateral hand displacement peak is ($r = 0.50$ at $\sigma = 1.1s$).

When predicting clinical scales, we use now only 2 active variables in order to limit overfitting: the variables TGDMfb (Total-goal directed movement for front/back direction) and TGDMlr (Total-goal directed movement for left/right direction). We then infer 6 parameters (2 association parameters + 4 hyperparameters) from the 37 RGS sessions. The two association parameters are (for normalised variables) $\beta_{\mathrm{TGDMfb}} = 0.15(0.13)$ and $\beta_{\mathrm{TGDMlr}} = 0.18(0.14)$. The hyperparameters of the model that predicts FM-UE for 'Whac-A-Mole' scenario are given by $a = 31.9(3.5)$, $b = 31.0(2.5)$, $\sigma_1 = 0.49(0.11)$, and $\beta_0 = 0.21(0.13)$.

The FM-UE predictions are shown against the true values in Fig. 9. The Pearson correlation between true FM-UE and predicted FM-UE is 0.63. Average error is $E \sim 11.2$. The value of the coefficient of determination R^2 is 0.39.

Generalisation to CAHAI and BI scales

In the previous sections we focused on the FM-UE scale but the CAHAI and BI scores are also available in the main dataset (cf. Table 1); we can then gain some insights on the differences between the three clinical scales from the point of view of the RGS kinematics.

Overall, the CAHAI scale has similar properties to FM-UE in relation to the kinematic descriptors, cf. Fig. 4. To stress the generalisation potential of the model, we can then adopt the same model introduced in Table 2 for the prediction of instantaneous FM-UE scores also for the prediction of the instantaneous CAHAI score. The association parameters for CAHAI are reported in Table 4. The most important variables are 'difficulty' and 'TGDM'. The hyperparameters of the covariate noise model that predicts CAHAI scores are $a = 39.158(0.099)$, $b = 51.962(0.079)$, $\sigma_1 = 0.953(0.064)$ and $\beta_0 = 0.0319(0.071)$. The predicted scores are plotted against the true CAHAI values in Fig. 10. The model predicts the CAHAI score with an average error of $E_{\mathrm{CAHAI}} \sim 20.1$, Pearson r true-predicted of 0.66 and a coefficient of determination $R^2 = 0.40$. This accuracy is close to what we obtained for the prediction of the instantaneous FM-UE, cf. Fig. 5.

In Fig. 11 we compare FM-UE and CAHAI, both for the true scores and the predicted scores. We note that

Table 4 The association parameters β for the prediction of the instantaneous CAHAI, 'Spheroids' scenario. The active variables are the same as in Table 2. The values refer to the normalised variables, so that the values of the different βs are directly comparable.

Covariate	β(**CAHAI**)
Difficulty	0.333(0.087)
TGDM	0.36(0.15)
Diff. Distance covered	0.016(0.088)
Diff. TGDM	-0.187(0.098)
Log. work area	0.09(0.10)
Log. smoothness	-0.04(0.14)

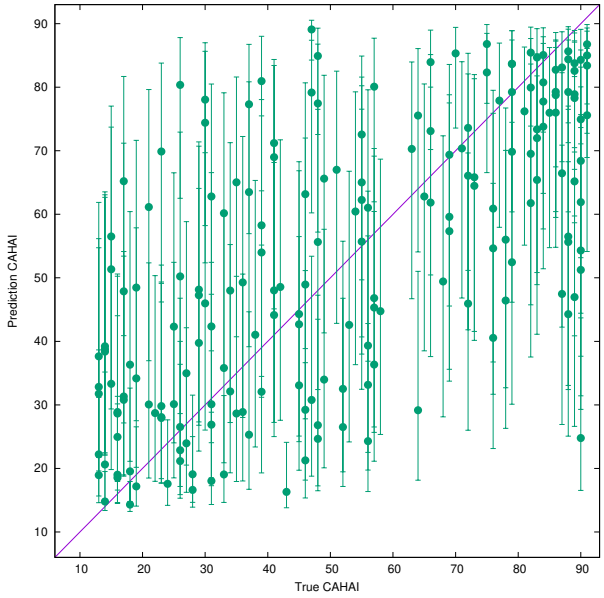

Figure 10 True versus predicted CAHAI for 191 data points of 'Spheroids' dataset (cf. Table 1), using the covariate noise model with association parameters given in Table 4.

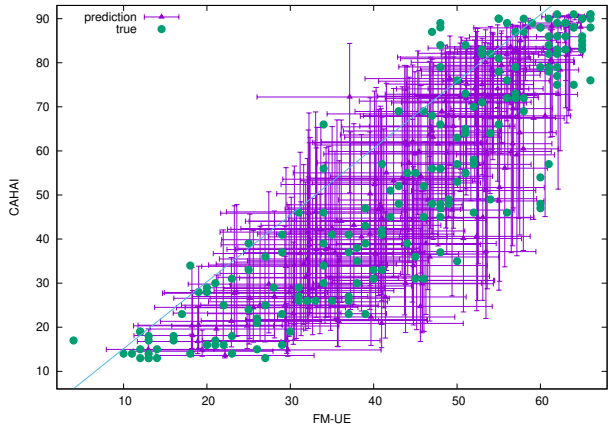

Figure 11 True FM-UE versus true CAHAI (green dots) and predicted FM-UE versus predicted CAHAI (purple triangles with errorbars) for 191 data points of 'Spheroids' database (cf. Table 1). We use the covariate noise model with association parameters given in Table 2 for FM-UE and Table 4 for CAHAI.

the relationship between FM and CAHAI is generally well preserved in the predictions; for example the Pearson between FM-UE and CAHAI scores is $r = 0.89$ for true values and $r = 0.88$ for predictions. The fact that the variability in the true FM-UE vs true CAHAI is seemingly comparable to the one in the prediction model, reinforces the idea that the precision we achieve is similar to the one of estimating FM-UE directly from CAHAI, as we estimated using Eq. 9.

Finally, we observe that the model that predicts FM-UE and CAHAI scores does not work well for the BI. Most kinematics variables have significantly smaller correlation with the BI (in particular 'work area', 'TGDM', 'smoothness') while baseline information and clinical history of the patient are comparatively more relevant (for example the patient's age, cf. Table 1 and Fig. 4). We then devise a different model for the prediction of the BI considering all variables, including not instantaneous ones (such as 'time since stroke' or 'session completed so far'). We use again repeated 50-50 cross-validation to avoid overfitting and select optimal active variable set. The active variable set for BI is composed by 5 variables (β values for normalised variables): 'age' ($\beta = -0.21(0.15)$), 'sessions completed so far' ($\beta = 0.24(0.19)$), 'difficulty' ($\beta = 1.21(0.72)$), 'Log. time since stroke' ($\beta = 0.14(0.14)$), 'Log. Difficulty' ($\beta = -0.91(0.70)$). Using a double noise model, cf. Eq. 6, we infer the hyperparameters $a = 51.56(0.10)$, $b = 49.22(0.12)$, $\sigma_1 = 0.5948(0.0072)$, $\sigma_2 = 0.178(0.027)$ and $\beta_0 = 1.025(0.011)$. Comparing true and predicted BI scores, we measure an average standard error of $E_{\rm BI} \sim 16.8$, a Pearson correlation true-prediction of 0.62 and a coefficient of determination R^2 of 0.35. This accuracy is comparable to the one achieved by the models for FM-UE and CAHAI scores. Nevertheless, the dataset Table 1 is very unbalanced towards high BI scores (mean score 80, with only 3 samples with a score below 25), so that the previous prediction performance will not generalise well to homogeneous unseen BI data (i.e., the precision for low scores is relatively poor).

Discussion and Conclusions

Our understanding of post-stroke motor recovery depends on our capacity to evaluate and characterise impairment and disability. Current standardised assessment methods rely on human criteria and present relevant unsystematic variability due to differences in the evaluators' training, lack of systematicity in the administration of the assessments, and often are excessively focused on one single aspect of the impairment and/or disability.

Different rehabilitation approaches show a preference for using (and even targeting) specific assessment methods for the evaluation of their therapeutic efficacy, and often these methods have been developed by the same team of authors. For example, the effectiveness of Constraint-Induced Movement Therapy [19] is usually evaluated using the Wolf Motor Function Test [20] and the Motor Analog Scale [21], while the effectivity of occupational therapy has been frequently assessed using the Barthel Index [22] and the Functional Independence Measure [23].

Thus, there is an urgent need to establish alternative methods for a common evaluation protocol and characterisation of the hemiparesis phenotype, thus allowing us to identify specific impairment features that could advance our understanding of the recovery dynamics and guide the design of effective rehabilitation therapies. In pursuing this objective, we have conducted a careful analysis of the raw kinematic data from the upper-extremities of 191 individuals with post-stroke hemiparesis, and we have constructed a predictive model of instantaneous function and recovery. Our results reveal a new digital biomarker of upper-limbs motor impairment, the Total Goal Directed Movement (TGDM), which relates to the patients range of motion during the execution of meaningful goal-oriented reaching movements. The TGDM strongly correlates with the level of impairment captured by the FM-UE and the level of disability captured by the CAHAI, and also presents predictive power about the patients' progress, showing high correlations with the magnitudes of improvement and deterioration estimated by both scales. This result is especially interesting given the current limited evidence about the responsiveness

of kinematic outcome measures of reaching performance in people with hemiparesis after stroke [24]. According to a recent systematic review on the clinimetric properties of kinematic upper limb assessments [9] only two papers captured responsiveness (i.e., ability to capture longitudinal changes in the measured construct), and just nine parameters showed enough evidence to predict recovery (i.e., number of velocity peaks, trunk displacement, task/movement time). The quality of evidence however was very low for all metrics. Further, current recommendations point out that wearables with integrated Inertial Measurement Units and vision-based tracking systems are insufficient to measure the quality of movement and improvement in motor function. However, our findings, together with the growing evidence supporting distance travelled as an accurate and responsive digital biomarker of recovery [25], suggest the opposite.

We built a model of instantaneous motor impairment and recovery as measured by FM-UE that capitalises on the information captured by the TGDM variable, and we validated it in terms of external validity (R^2 : 0.38), robustness (test-retest reliability) ($r > 0.89$), responsiveness (R^2 : 0.57), sensitivity (TPR : 95%) and generalisation ($r > 0.43$). The relevance of our results is emphasised by their consistency across clinimetric properties and by their generalisation potential, relying on a large and heterogeneous dataset of patients at different stages post-stroke. We believe that the applicability of the TGDM and its derived models to evaluate impairment and motor recovery is promising for a number of reasons: 1) it can be derived from unimanual displacements executed in the horizontal plane, 2) it does generalise to other tasks involving two-dimensional horizontal reaching movements towards targets, and 3) it can be estimated during unsupervised motor training. Our results provide an early example of how fully digital biomarkers of deficits and recovery post-stroke can provide new digital health methods and technologies for neurorehabilitation that can generalise beyond the clinic and serve continuous high-resolution diagnostics, prognostics and intervention.

Appendix

In Table 5 we report the descriptive statistics of the secondary variables (functions of primary variables in Table 1) for dataset 'Spheroids'.

The distribution of the Pearson correlation coefficients of the FM-UE score with all the variables in Tables 1, 5 is shown in Fig. 12. This is compared with the distribution obtained with randomised outcome, whose standard deviation is $r \simeq 0.081$.

In Table 6 we report the descriptive statistics of the secondary variables (functions of primary variables in Table 3) for the 'responsiveness' dataset (change of variables and clinical scales between two RGS sessions of the same patient with a delay of at least 16 days).

In Fig. 13 we show the Pearson's r between $J(\sigma)$ Eq. 2 and the clinical scales as function of the timescale σ for the database 'Whac-A-Mole', cf. Sec. 'Generalisation'. This is a dataset composed of 37 RGS sessions with limited range of the clinical scales: FM-UE range of observations $[5, 60]$, mean 37, SD 14; CAHAI range of observations $[7, 49]$, mean 32, SD 15; BI range of observations $[48, 100]$, mean 85, SD 13. The 'Whac-A-Mole' scenario requires movements in all 2D place (unlike 'Spheroids' scenario, cf. Table 1), so we define two independent $J(\sigma)$: one associated to the front/back trajectory and one associated to the lateral (left/right) trajectory. In Fig. 13 we show that in both directions there is a peak in the Pearson's r at about $\sigma = 1$ s for all three clinical scales. In correspondence of these two peaks, we define two variables: TGDMfb (Total-goal directed movement in front/back direction) and TGDMlr (Total-goal directed movement in left-right direction). We stress that the timescale of the peak is roughly equivalent to the timescale of the gameplay of the 'Whac-A-Mole' scenario. Finally, we note that for the same analysis done in the 'Spheroids' scenario we observe two clear peaks in the lateral direction (cf. Fig. 3). In 'Whac-A-Mole' instead the high-frequency peak is not clearly visible. One factor that may affect this difference is that the gameplay of 'Whac-A-Mole' is faster than 'Spheroids' (roughly 1 s instead of 10 s), so that the separation between the two potential peaks is smaller. Other factors that may influence the absence of the second peak are the fact that the gameplay is 2D (so that the speed in one direction is less informative) or the inadequate time-resolution of the camera (different from the one used for 'Spheroids').

Funding

This study was supported by the European Commission (EC), the European Research Council under grant agreement 341196 (cDAC), EC H2020 project socSMCs (H2020EU.1.2.2. 641321), and by the RGS@home, EIT Health project 19277. EIT Health is supported by EIT, a body of the European Union.

Abbreviations

- RGS: Rehabilitation Gaming System
- FM-UE: Fugl-Meyer Assessment for Upper Extremities
- CAHAI: Chedoke Arm and Hand Activity Inventory
- BI: Barthel Index
- SD: Standard Deviation
- TGDM: Total-Goal Directed Movement
- max-sp (Figs. 4,7): maximum reaching speed
- perc-hit (Figs. 4,7): performance (percentage hit)
- dist-cov (Figs. 4,7): distance covered
- w-area (Figs. 4,7): work area
- num-sess (Figs. 4,7): sessions completed so far
- days-s (Figs. 4,7): time since stroke (days)
- paretic-dom (Figs. 4,7): dominant side more affected
- MDC: Minimal Detectable Change
- TPR: True Positive Rate

Table 5 Characteristics of the secondary variables for the 191 samples composing the main dataset (obtained from first order variables, cf. Table 1). The r columns refers to the Pearson correlation coefficients with the FM-UE, CAHAI, and BI clinical scales, respectively. Correlations below the significance threshold $r \sim 0.081$ (cf. Fig. 12 in Appendix) are in grey. From the time since stroke we obtain the categories Acute (5-90 days), Sub-acute (3-12 months) and Chronic (over 1 year). The 'instantaneous' variables obtained directly from game log-file are in Italic type. The 'Diff.' variables are obtained as the difference between the value observed for the less affected arm and the value for the more affected one. The 'Log.' variables are obtained as the natural logarithm of the corresponding first order variables.

Variables	**Range [min, max]**	**Mean**	**SD**	r**(FM-UE)**	r**(CAHAI)**	r**(BI)**
15. Chronic	yes/no	57/134	-	-0.25	-0.17	0.14
<Subacute>	yes/no	74/117	-	-	-	-
16. Acute	yes/no	60/131	-	0.32	0.18	-0.15
17. *Diff. work area* (m^2)	[-1.3, 1.6]	0.25	0.50	-0.31	-0.29	-0.069
18. *Diff. distance covered* (m)	[-160, 120]	11	34	-0.32	-0.26	-0.12
19. *Diff. performance (% success)*	[-0.22, 0.53]	0.087	0.13	-0.26	-0.22	-0.16
20. *Diff. maximum reaching speed* (m/s)	[-61, 98]	8.1	23	-0.22	-0.22	-0.15
21. *Diff. difficulty level reached*	[-0.47,0.63]	0.11	0.19	-0.22	-0.15	-0.058
22. *Diff. smoothness* (mm)	[-2.6, 4.4]	0.34	0.75	-0.29	-0.28	-0.20
23. *Diff. TGDM* (m)	[-0.035, 0.084]	0.015	0.021	-0.52	-0.47	-0.24
24. Log. time since stroke (days)	[1.6, 8.0]	4.8	1.7	-0.24	-0.076	0.36
25. Log. sessions completed so far	[0.0, 3.9]	1.8	0.9	0.29	0.38	0.42
26. *Log. work area*	[-4.3, 0.62]	-1.3	0.9	0.40	0.38	0.19
27. *Log. distance covered*	[0.92, 5.5]	3.8	0.7	0.26	0.32	0.34
28. *Log. maximum reaching speed*	[1.0, 4.5]	2.6	0.73	0.26	0.20	0.045
29. *Log. difficulty level reached*	[-0.16, 0.60]	0.25	0.14	0.44	0.50	0.45
30. *Log. smoothness*	[-1.7, 1.4]	0.012	0.48	0.48	0.46	0.33
31. *Log. TGDM*	[-4.7, -2.1]	-2.9	0.48	0.52	0.56	0.43

Table 6 Characteristics of the secondary variables for the 54 samples composing the responsiveness dataset (obtained from first order variables, cf. Table 3). We select couple of sessions of same patient with a delay of at least 16 days from the main dataset, cf. Table 1. The r columns refers to the Pearson correlation coefficients with the FM-UE, CAHAI, and BI clinical scales, respectively. Correlations below the significance threshold $r \sim 0.14$ are in grey. The 'instantaneous' variables obtained directly from game log-file are in Italic type. The 'Diff.' variables are obtained as the difference between the value observed for the less affected arm and the value for the more affected one. The 'Log.' variables are obtained as the natural logarithm of the corresponding first order variables.

Variables	**Range [min,max]**	**Mean**	**SD**	$r(\Delta$**FM-UE**$)$	$r(\Delta$**CAHAI**$)$	$r(\Delta$**BI**$)$
15. Chronic	yes/no	10/44	-	-0.30	-0.44	-0.38
<Subacute>	yes/no	10/44	-	-	-	-
16. Acute	yes/no	34/20	-	0.19	0.44	0.53
17. *Change in diff. work area* (m)	[-1.5, 1.8]	-0.13	0.67	-0.30	-0.060	-0.24
18. *Change in diff. distance covered* (m)	[-94, 150]	3.1	40	-0.45	0.23	-0.30
19. *Change in diff. performance (% success)*	[-0.22, 0.53]	0.087	0.13	0.18	0.10	0.17
20. *Change in diff. maximum reaching speed* (m/s)	[-76, 71]	-5.1	30	-0.28	-0.16	-0.32
21. *Change in diff. difficulty*	[-0.46, 0.62]	0.0095	0.21	0.064	0.12	0.14
22. *Change in diff. smoothness* (mm)	[-2.7, 1.4]	-0.17	0.85	-0.41	-0.26	-0.47
23. *Change in diff. TGDM* (m)	[-0.12, 0.088]	-0.0050	0.040	-0.51	-0.43	-0.56
24. Log. time since stroke (at first)	[0.41, 1.2]	0.73	0.16	-0.36	-0.37	-0.40
25. Log. sessions completed so far (at first)	[0, 1.2]	0.52	0.39	-0.55	-0.36	-0.55
26. *Change in Log. work area*	[-6.4, 1.5]	-0.79	1.6	0.062	0.057	0.11
27. *Change in Log. distance covered*	[-0.90, 5.6]	3.0	2.0	0.18	0.16	0.17
28. *Change in Log. maximum reaching speed*	[-1.8, 5.6]	1.4	2.3	0.15	0.30	0.17
29. *Change in Log. difficulty*	[-3.8, 0]	-1.4	0.90	0.064	0.13	0.024
30. *Change in Log. smoothness*	[-2.0, 3.0]	0.63	1.1	0.31	0.32	0.37
31. *Change in Log. TGDM*	[-10, 0.5]	-1.4	1.8	0.035	0.17	0.037
Days between sessions	[17,89]	49	26	0.037	0.31	0.39
Initial FM score	[13,66]	44	15	-0.44	-0.20	-0.13
Initial CAHAI score	[14,90]	48	22	-0.50	-0.50	-0.38
Initial BI score	[31,100]	72	23	-0.60	-0.64	-0.82

Ethics approval and consent to participate
All research on human subjects reported in this manuscript was prospectively approved by the Institutional Review Board of Hospital Vall d'Hebron, Hospital del Mar i l'Esperança from Barcelona, and Hospital Joan XXIII from Tarragona, and all participants provided written informed consent.

Competing interests
P. F. M. J. Verschure leads the research group SPECS that developed RGS, and is the CEO/founder of the spin-off company Eodyne Systems, SL, which commercializes RGS with the goal to achieve a large-scale distribution of low-cost science-based rehabilitation technologies.

Availability of data and materials
The datasets used and analysed in this study are available from the corresponding author on reasonable request.

Consent for publication
Not applicable.

Authors' contributions
Contributors BR, FA, and AC planned the analysis. BR, MM and PV contributed to data acquisition. FA, and AC analysed the data. All authors wrote the draft paper. PV submitted the study and is responsible for the overall content as guarantor.

Author details
[1]Laboratory of Synthetic, Perceptive, Emotive and Cognitive Systems (SPECS), Institute for Bioengineering of Catalonia (IBEC), The Barcelona Institute of Science and Technology (BIST), Baldiri Reixac 10-12, 08028, Barcelona, Spain. [2]Saddle Point Science Ltd, 10 Lincoln Street, York, UK. [3]Institució Catalana de Recerca, Estudis Avançats (ICREA), Barcelona, Spain.

References

1. Thrift, A.G., Thayabaranathan, T., Howard, G., Howard, V.J., Rothwell, P.M., Feigin, V.L., Norrving, B., Donnan, G.A., Cadilhac, D.A.: Global stroke statistics. International Journal of Stroke **12**(1), 13–32 (2017)
2. Shah, S., Vanclay, F., Cooper, B.: Improving the sensitivity of the barthel index for stroke rehabilitation. Journal of clinical epidemiology **42**(8), 703–709 (1989)
3. Lai, S.-M., Studenski, S., Duncan, P.W., Perera, S.: Persisting consequences of stroke measured by the stroke impact scale. Stroke **33**(7), 1840–1844 (2002)
4. Krakauer, J.W., Cortés, J.C.: A non-task-oriented approach based on high-dose playful movement exploration for rehabilitation of the upper limb early after stroke: a proposal. NeuroRehabilitation **43**(1), 31–40 (2018)
5. Asboth, L., Friedli, L., Beauparlant, J., Martinez-Gonzalez, C., Anil, S., Rey, E., Baud, L., Pidpruzhnykova, G., Anderson, M.A., Shkorbatova, P., *et al.*: Cortico–reticulo–spinal circuit reorganization enables functional recovery after severe spinal cord contusion. Nature neuroscience **21**(4), 576–588 (2018)
6. Makin, T., Diedrichsen, J., Krakauer, J.: Reorganisation in adult primate sensorimotor cortex: Does it really happen. The cognitive neurosciences: Vol. sixth edition. Cambridge, MA: MIT Press. Find this resource (2020)
7. McDonnell, M., Hillier, S.: Vestibular rehabilitation for unilateral peripheral vestibular dysfunction. Cochrane Database of Systematic Reviews (1) (2015). doi:10.1002/14651858.CD005397.pub4
8. Ballester, B.R., Duff, A., Maier, M., Cameirao, M., Bermudez, S., Duarte, E., Cuxart, A., Rodríguez, S., Verschure, P.F.: Revealing an extended critical window of recovery post-stroke. bioRxiv, 458745 (2018)
9. Schwarz, A., Kanzler, C.M., Lambercy, O., Luft, A.R., Veerbeek, J.M.: Systematic review on kinematic assessments of upper limb movements after stroke. Stroke **50**(3), 718–727 (2019)
10. Wegner, C., Filippi, M., Korteweg, T., Beckmann, C., Ciccarelli, O., De Stefano, N., Enzinger, C., Fazekas, F., Agosta, F., Gass, A., *et al.*: Relating functional changes during hand movement to clinical parameters in patients with multiple sclerosis in a multi-centre fmri study. European journal of neurology **15**(2), 113–122 (2008)
11. Murphy, M.A., Willén, C., Sunnerhagen, K.S.: Kinematic variables quantifying upper-extremity performance after stroke during reaching and drinking from a glass. Neurorehabilitation and neural repair **25**(1), 71–80 (2011)
12. Abdel Majeed, Y., Awadalla, S.S., Patton, J.L.: Regression techniques employing feature selection to predict clinical outcomes in stroke. PloS one **13**(10), 0205639 (2018)
13. da Silva Cameirão, M., Bermudez i Badia, S., Duarte, E., Verschure, P.F.: Virtual reality based rehabilitation speeds up functional recovery of the upper extremities after stroke: a randomized controlled pilot study in the acute phase of stroke using the rehabilitation gaming system. Restorative neurology and neuroscience **29**(5), 287–298 (2011)
14. Ballester, B.R., Maier, M., Mozo, R.M.S.S., Castañeda, V., Duff, A., Verschure, P.F.: Counteracting learned non-use in chronic stroke patients with reinforcement-induced movement therapy. Journal of neuroengineering and rehabilitation **13**(1), 1–15 (2016)
15. Ballester, B.R., Maier, M., Domingo, D.A., Aguilar, A., Mura, A., Pareja, L.T., Esteve, M.F.G., Verschure, P.F.M.J.: Adaptive vr-based rehabilitation to prevent deterioration in adults with cerebral palsy. In: 2019 International Conference on Virtual Rehabilitation (ICVR), pp. 1–7 (2019). doi:10.1109/ICVR46560.2019.8994754
16. Jarvis, R.A.: On the identification of the convex hull of a finite set of points in the plane. Information Processing Letters **2**(1), 18–21 (1973). doi:10.1016/0020-0190(73)90020-3
17. Cameirão, M.S., i Badia, S.B., Oller, E.D., Verschure, P.F.: Neurorehabilitation using the virtual reality based rehabilitation gaming system: methodology, design, psychometrics, usability and validation. Journal of neuroengineering and rehabilitation **7**(1), 1–14 (2010)
18. Nirme, J., Duff, A., Verschure, P.F.: Adaptive rehabilitation gaming system: on-line individualization of stroke rehabilitation. In: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 6749–6752 (2011). IEEE
19. Wolf, S.L., Winstein, C.J., Miller, J.P., Taub, E., Uswatte, G., Morris, D., Giuliani, C., Light, K.E., Nichols-Larsen, D., EXCITE Investigators, f.t., *et al.*: Effect of constraint-induced movement therapy on upper extremity function 3 to 9 months after stroke: the excite randomized clinical trial. Jama **296**(17), 2095–2104 (2006)
20. Wolf, S.L., Catlin, P.A., Ellis, M., Archer, A.L., Morgan, B., Piacentino, A.: Assessing wolf motor function test as outcome measure for research in patients after stroke. Stroke **32**(7), 1635–1639 (2001)
21. Taub, E., Morris, D.M., Crago, J., King, D.K., Bowman, M., Bryson, C., Bishop, S., Pearson, S., Shaw, S.E.: Wolf motor function test (wmft) manual. Birmingham: University of Alabama, CI Therapy Research Group (2011)
22. Mahoney, F.I., Barthel, D.W.: Functional evaluation: the barthel index: a simple index of independence useful in scoring improvement in the rehabilitation of the chronically ill. Maryland state medical journal (1965)
23. Ottenbacher, K.J., Hsu, Y., Granger, C.V., Fiedler, R.C.: The reliability of the functional independence measure: a quantitative review. Archives of physical medicine and rehabilitation **77**(12), 1226–1232 (1996)
24. Veerbeek, J.M., Langbroek-Amersfoort, A.C., Van Wegen, E.E., Meskers, C.G., Kwakkel, G.: Effects of robot-assisted therapy for the upper limb after stroke: a systematic review and meta-analysis. Neurorehabilitation and neural repair **31**(2), 107–121 (2017)
25. Wagner, J.M., Rhodes, J.A., Patten, C.: Reproducibility and minimal detectable change of three-dimensional kinematic analysis of reaching tasks in people with hemiparesis after stroke. Physical therapy **88**(5), 652–663 (2008)

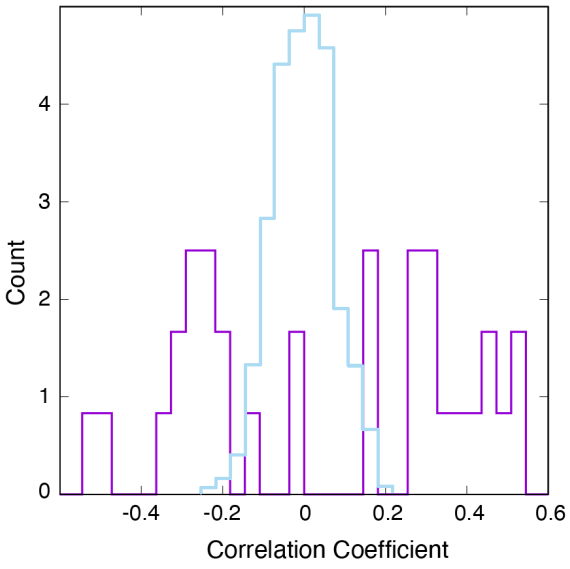

Figure 12 Distribution of Pearson correlation coefficients measured between the FM-UE scores and all the variables in Tables 1,5. The light blue line is the distribution obtained for random outcomes.

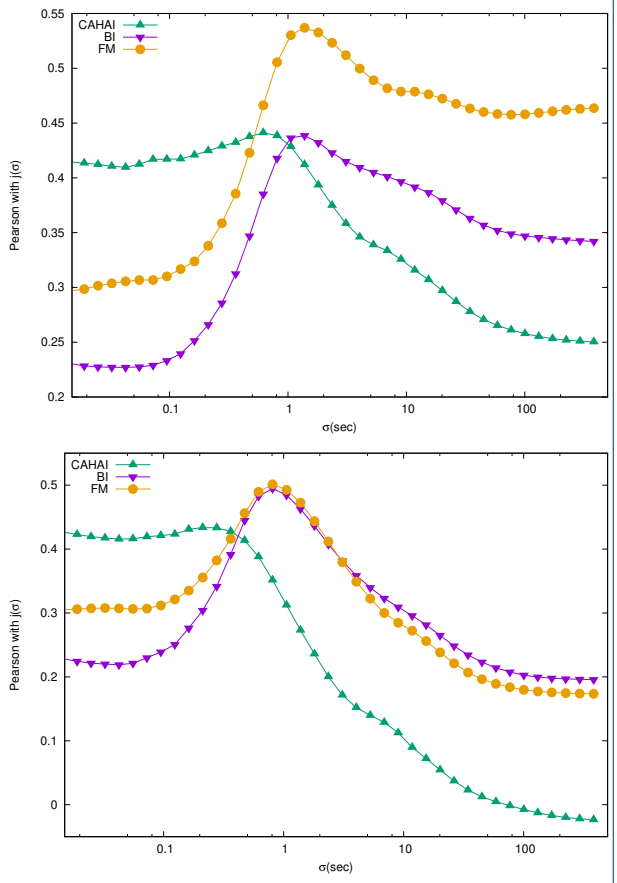

Figure 13 Pearson correlation coefficient between clinical scales (FM-UE, CAHAI, BI) and $J(\sigma)$ Eq. 2 in the front/back direction (top) and left/right direction (bottom) at different values of the timescale σ on dataset "Whac-A-Mole" ($N = 37$, cf. Sec. 'Generalisation'). For the FM-UE score, the peaks are at ($r = 0.54$ for $\sigma = 1.3s$) for the frontal direction and at ($r = 0.50$ at $\sigma = 1.1s$) for the lateral direction. For the CAHAI score, the peaks are at ($r = 0.44$ for $\sigma = 0.78s$) for the frontal direction and at ($r = 0.43$ at $\sigma = 0.31s$) for the lateral direction. For the BI, the peaks are at ($r = 0.44$ for $\sigma = 1.2s$) for the frontal direction and at ($r = 0.49$ at $\sigma = 1.1s$) for the lateral direction.