Spinal Cord (2016) 54, 1145–1151

www.nature.com/sc

ORIGINAL ARTICLE

Predicting task performance from upper extremity impairment measures after cervical spinal cord injury

J Zariffa[1,2,3], A Curt[4], MC Verrier[1,3,5], MG Fehlings[6,7], S Kalsi-Ryan[5,6], GRASSP Cross-Sectional Study Team and Ontario GRASSP Longitudinal Study Team

Background: Automated sensor-based assessments of upper extremity (UE) function after cervical spinal cord injury (SCI) could provide more detailed tracking of individual recovery profiles than is possible with existing assessments, and optimize the delivery and assessment of new interventions. The design of reliable automated assessments requires identifying the key variables that need to be measured to meaningfully quantify UE function. An unanswered question is to what extent measures of sensorimotor impairment can quantitatively predict performance on functional tasks.

Objective: The objective was to define the predictive value of impairment measures for concurrent functional task performance in traumatic cervical SCI, as measured by the Graded Redefined Assessment of Strength, Sensibility and Prehension (GRASSP).

Setting: Retrospective analysis.

Methods: A data set of 138 GRASSP assessments was analyzed. The Strength and Sensation modules were used as measures of impairment, whereas the concurrent Prehension Performance module was used as the surrogate measure of function. Classifiers were trained to predict the scores on each of the six individual tasks in the Prehension Performance module. The six scores were added to obtain a total score.

Results: The Spearman's ρ between predicted and actual total Prehension Performance scores was 0.84. Predictions using both the Strength and Sensation scores were not found to be superior to predictions using the Strength scores alone.

Conclusions: Measures of UE motor impairment are highly predictive of functional task performance after cervical SCI. Automated sensor-based assessments of UE motor function after SCI can rely on measuring only impairment and estimating functional performance accordingly.

Spinal Cord (2016) **54,** 1145–1151; doi:10.1038/sc.2016.77; published online 31 May 2016

INTRODUCTION

Upper extremity (UE) function is fundamental to most activities of daily living. Cervical spinal cord injury (SCI) may result in UE paralysis or paresis, with devastating consequences for independence, quality of life and community participation. However, even small amounts of UE functional recovery after SCI can have significant implications for regaining independence.[1] Developing treatment interventions that may help restore UE function is therefore of the utmost importance and is the focus of multiple active lines of research.[2–6] An integral part of the development of new interventions is the ability to reveal their effectiveness, which requires a suitable toolbox of clinical assessments.

A number of specialized clinical assessments are currently available for measuring different aspects of UE function after SCI. Following the terminology of the International Classification of Functioning, Disability and Health (ICF),[7] assessments are often categorized as measuring (i) body structures and functions (that is, characterizing the impairment), (ii) activity (that is, characterizing the ability to accomplish functional tasks) or (iii) community participation. In SCI, the predominant assessment of body structures and function is the International Standards for the Neurological Classification of Spinal Cord Injury (ISNCSCI).[8] SCI-specific assessments relevant to UE activity include the Graded Redefined Assessment of Strength, Sensibility and Prehension (GRASSP),[9] the Capabilities of the Upper Extremity Test,[10] the Toronto Rehabilitation Institute Hand Function Test[11] and the Spinal Cord Independence Measure (SCIM).[12] As for the applicability of participation measures in SCI, a comprehensive toolkit known as the Participation and Quality of Life toolkit has been developed.[13] Taken together, these clinical assessments can provide a thorough description of UE status after SCI and help infer the interrelationships between ICF domains. However, they are all limited in terms of their frequency of administration and the time of administration post SCI. None of the tools listed would typically be administered more often than every few weeks. Furthermore, assessment outside of a clinical or laboratory setting is difficult and most often relies on self-report using instruments such as the Capabilities of Upper Extremity Questionnaire[14] or interviews such as the SCIM.

Technological approaches that can automate the assessment of UE function after SCI based on combinations of sensors could assist with overcoming both of the limitations outlined: low frequency of

[1]Toronto Rehabilitation Institute–University Health Network, Toronto, Ontario, Canada; [2]Institute of Biomaterials and Biomedical Engineering, University of Toronto, Toronto, Ontario, Canada; [3]Rehabilitation Sciences Institute, University of Toronto, Toronto, Ontario, Canada; [4]Spinal Cord Injury Center, Balgrist University Hospital, Zurich, Switzerland; [5]Department of Physical Therapy, University of Toronto, Toronto, Ontario, Canada; [6]Krembil Neuroscience Centre–University Health Network, Toronto, Ontario, Canada and [7]Department of Surgery, University of Toronto, Ontario, Canada
Correspondence: Dr J Zariffa, Toronto Rehabilitation Institute–University Health Network, 550 University Avenue, #12-102, Toronto, Ontario, Canada M5G 2A2.
E-mail: jose.zariffa@utoronto.ca
Received 6 November 2015; revised 12 April 2016; accepted 17 April 2016; published online 31 May 2016

assessment and lack of assessments in the community. Automated sensor-based UE assessments (hereafter referred to as 'automated assessments') could 'fill in the gaps' left by the current assessment tools and have a triple benefit for advancing patient care: (i) improved tracking of recovery in the sub-acute phase, providing the data needed to support new strategies in individualized rehabilitation programs;[15,16] (ii) integration with telerehabilitation interventions, to ensure appropriate therapy progression even in the absence of regular contact with a clinician; and (iii) more efficient design of clinical trials, through enhanced characterization of recovery profiles longitudinally.[17]

To date, only a small number of studies have proposed automated UE assessments specifically for the SCI population. These have relied primarily on robotic rehabilitation devices[18–20] and inertial measurement units.[21,22] Although these studies have shown that data obtained from automated systems can have strong relationships with validated manual clinical assessments, this field of research is still in its early stages.

A fundamental question relevant to the development of automated assessments is to what extent will the data recorded by a collection of sensors be able to predict performance on a functional task? In most cases, the quantities directly recorded by sensors will relate most closely to body functions and impairment—for example, muscle activation or range of motion. Task performance is a more complex construct that depends on the integrated function of many body systems, as well as a host of other factors (for example, environment, compensation strategies, motivation and so on). In this study, our objective was to evaluate how well task performance can be predicted from measures of body function.

The GRASSP served as a framework for our investigation. The GRASSP is a manual (that is, not sensor-based) assessment designed specifically for UE function after cervical SCI.[9,23] We use it here because it simultaneously captures multiple constructs of interest, which allows us to study the relationships between them. Specifically, the GRASSP consists of three domains and five subtests. The Strength domain (GR-str) consists of motor testing of 10 UE muscles; the Sensation domain consists of the Dorsal Sensation (GR-ds) and Palmar Sensation (GR-ps) subtests; the Prehension domain consists of the Prehension Ability subtest that examines a set of grasp patterns (GR-pa) and the Prehension Performance subtest (GR-pp) that examines a set of functional tasks.[9] By using the Strength and Sensation (GR-ds+GR-ps = GR-sens) scores to reflect body functions, and the GR-pp scores to reflect performance on functional tasks, we can investigate the predictive relationships between these two concepts. This study does not, in itself, describe a novel automated assessment; rather, we seek to elucidate the underlying relationships that will inform the development of such technology.

METHODS

GRASSP

Detailed information on the GRASSP and its properties can be found in previous publications.[9,23] We briefly review the main features of the assessment:

- The Strength domain is evaluated through manual muscle testing of 10 UE muscles, namely the anterior deltoid, elbow flexors, elbow extensors, wrist extensors, extensor digitorum, opponens pollicis, flexor pollicis longus, finger flexors, finger abductors and first dorsal interossei. Each receives a score from 0 to 5, for a total GR-str score between 0 and 50 for each of the right and left sides.
- The Sensation domain is evaluated by testing three palmar and three dorsal finger locations with Semmes Weinstein monofilaments. Each location receives a score from 0 to 4, resulting in GR-ps and GR-ds between 0 and 12 each and a total GR-sens score of up to 24, for each side.
- The Prehension Ability domain is evaluated by asking the individual to perform three prehension patterns (cylindrical grasps, lateral key pinch and tip to tip pinch). Each is scored from 0 to 4, based on active vs passive positioning of the wrist and fingers. This results in a GR-pa score between 0 and 12 for each side.
- The Prehension Performance domain is evaluated based on six functional tasks: namely, pouring water from a bottle, unscrewing lids from jars, performing a pegboard task, using a key, manipulating coins and placing nuts onto screws.[24] Each task is scored from 0 to 5, for a total GR-pp score between 0 and 30.

Data set

A retrospective analysis was conducted on GRASSP assessments collected during previous longitudinal and cross-sectional studies.[9,25] Where available, comments included in the study documentation were examined prior to analysis. We excluded from analysis any assessment in which comments indicated factors that might have altered the relationships between muscle strength and performance on the functional tasks—for example, pain or additional support required for the arm to conduct the measurement. Note that comments were not available for all records. For the longitudinal study records, three time points were available: 4–6 weeks, 3 months and 6 months post injury. A single time point per participant was included in the data set. We used the 4–6-week time point, except where an examination was missing or comments suggested that examination should be excluded, in which case we used the 3-month time point. For cross-sectional study records, a single time point was available, and all participants had chronic injuries (ranging from 6 months to 20 years post injury). The use of mixed time points is beneficial in building a heterogeneous data set that will produce robust classifiers (see next section). The right arm was used in all cases, such that a single assessment was used per individual, with no preference for hand dominance.

Classifier design

Hereafter, we refer to measures corresponding to the 'body functions and structures' component of the ICF classification as measures of 'impairment' and to measures corresponding to the 'activities' component of the ICF as measures of 'task performance'.

We used a machine learning approach to capture the relationships between impairment and task performance. Our objective was to design a classifier that accepts measures of impairment as input and produces a task performance score as output. Thus, we selected as our inputs the GR-str scores (10 manual motor testing values, see above) and the GR-sens scores (6 values from monofilament testing).

The outputs were the GR-pp scores from the six tasks of the GR-pp module.[9] We built a separate classifier for each of the six tasks. Each task is scored on a scale from 0 to 5, and therefore each classifier had to assign a given input to one of the six classes (that is, a score of 0, 1, 2, 3, 4 or 5). The outputs of all six classifiers were summed to produce a total GR-pp score. The total GR-pp score has been shown previously to correlate strongly with independence in activities of daily living, as measured by the SCIM Self-Care sub-score.[26] A diagram summarizing the prediction process is shown in Figure 1.

We constructed two versions of each classifier: one using only the Strength scores as inputs, and one using both Strength and Sensation scores as inputs. A comparison of these two approaches was conducted to determine whether information about sensory function was essential for predicting task performance.

Random Forest classifiers[27] were used, using 50 trees. The minimum number of observations per leaf was set to 2. A Random Forest is a collection of decision trees trained on random subsets of data; to classify a new observation, a majority vote is used among all the trees. Although individual trees have only limited prediction abilities, the ensemble of trees as a whole forms an effective classifier. Random Forests are a popular method because of their relative simplicity combined with good performance on a wide variety of problems.

Classifier training and evaluation

Leave-one-out cross-validation was used to train the classifiers and test their performance. The following performance metrics were used:

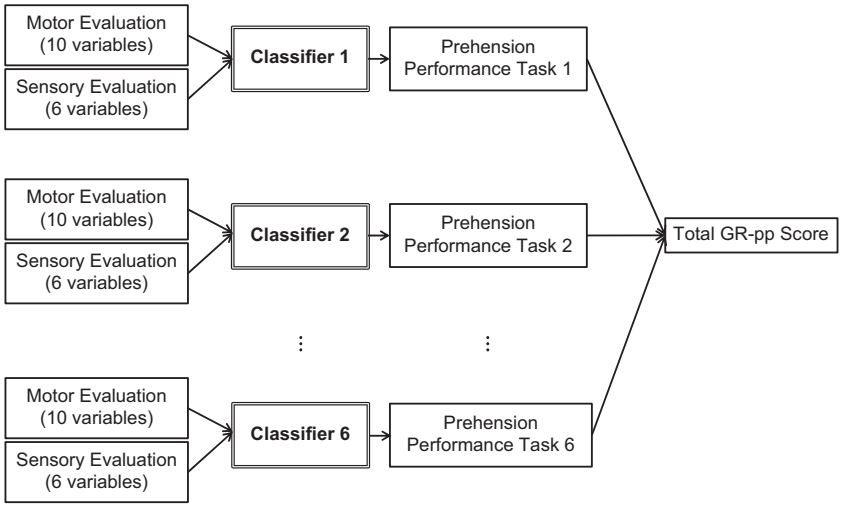

Figure 1 Flow diagram for prediction of GR-pp sub-scores and total scores. Note that the same motor and sensory inputs are provided to all six classifiers. Each classifier produces an output with a value of 0, 1, 2, 3, 4 or 5. The total score is the sum of these outputs and is thus an integer value between 0 and 30.

- Error between predicted and actual scores in each of the six individual classifiers.
- Error between predicted and actual total GR-pp scores.
- Spearman's rank correlation coefficient (Spearman's ρ) between predicted and actual total GR-pp scores.

Post hoc analysis

As noted above, not all entries in the data set contained complete comments. Information may thus have been missing about factors that could have affected the impairment-to-task performance relationship (for example, pain). Our next step was designed to minimize the impact of such confounding factors. We assumed that if a confounding factor was present, the performance of the classifiers would suffer. We therefore identified outliers in the results of the analysis described above, removed them from the data set and re-trained the classifiers on this reduced data set. This process was repeated until no outliers remained. For the purposes of eliminating entries from the data set, an outlier was defined as a data point for which the absolute error between the predicted and actual total GR-pp score was greater than 15 points (that is, half the maximum possible score). The analysis on the final reduced data set is referred to as the *post hoc* analysis.

Comparison with predictions from neurological and motor levels

Finally, to confirm that the prediction of functional task performance does indeed require fine-grained information about impairment, we compared the prediction performance of the classifiers described above with the performance that would be achieved simply based on the ISNCSCI neurological level of injury or the right-side motor level. For each individual in the data set, the predicted score for each of the six GR-pp tasks was set to the rounded mean score for that task among all individuals with the same neurological level in the data set. The total GR-pp scores were then computed, and the Spearman's ρ between these predicted scores and the true total GR-pp scores was determined. The analysis was repeated using the right-side motor level instead of the neurological level.

RESULTS

Data set

The retrospective analysis of the available data yielded records from 138 study participants, 53 from the longitudinal study and 85 from the cross-sectional study. After examining the available comment fields for these assessments, nine participants were excluded. Reasons included shoulder restrictions due to injury, pain or both; additional support provided to the arm during the manipulation tasks (for example, right arm supported by the left arm); motor scores evaluated on a restricted range of motion; significant spasticity in the UE; and scores of 0 assigned without the participants having actually attempted the task. In an additional two participants, the 4–6-week assessment was missing; hence, the 3-month assessments were used. As a result, 129 assessments were included in the data set, each corresponding to a separate individual. The neurological level of injury of the included participants ranged from C1 to T1, and their severities from American Spinal Injury Association Impairment Scale (AIS) A to D.[9,25]

Prediction of prehension performance sub-scores

Figure 2 shows the distribution of errors obtained in predicting the GR-pp sub-scores using only the GR-str scores as inputs. Out of 774 predictions (129 assessments × 6 sub-scores, obtained by collecting all of the test sets from the cross-validation process), 444 were correct (57.4%), 158 (20.4%) differed from the correct score by 1 point and 172 (22.2%) differed by 2 or more points. Thus, 77.8% were within 1 point of the correct score on a 0 to 5 scale.

When both the GR-str and GR-sens scores were included as inputs to the classifiers, 448 predictions were correct (57.9%), 153 (19.8%) differed from the correct score by 1 point, and 173 (22.3%) differed by 2 or more points. A χ^2 test comparing the error distributions when using only the GR-str scores and when using both the GR-str and GR-sens scores revealed no significant difference ($P=0.95$).

Prediction of prehension performance total scores

Figure 3a shows a scatter plot of the predicted vs actual total GR-pp scores, obtained by summing the outputs of the six classifiers. The Spearman's ρ between the actual and predicted scores was 0.84. Figure 3b shows the errors between actual and predicted total scores for all entries in the data set. The median error was 0 (IQR = − 2–3) on a 30-point scale. The data in these figures correspond to using only the GR-str scores as the classifier input.

When both the GR-str and GR-sens scores were included as inputs to the classifiers, the median error was 0 (IQR = − 2–4), which was not significantly different than when GR-str only was used ($P=0.66$ using a Wilcoxon rank sum test). The Spearman's ρ between the actual and predicted scores when using both GR-str and GR-sens was 0.84.

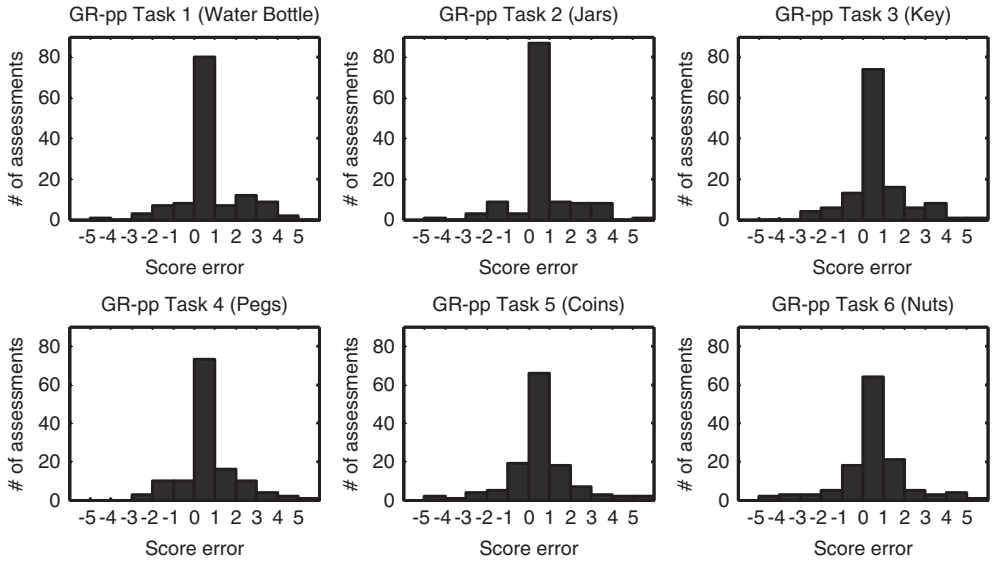

Figure 2 Histograms of errors between the actual and predicted sub-scores, for each of the six classifiers. The results in this figure correspond to the pre-planned analysis. A full color version of this figure is available at the *Spinal Cord* journal online.

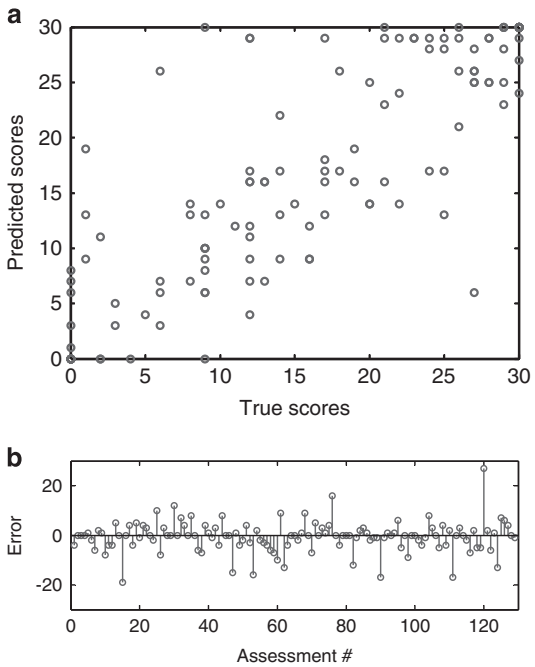

Figure 3 (**a**) Predicted vs actual total GR-pp scores. (**b**) Error between the predicted and actual total score for each of the assessments in the data set. The results in this figure correspond to the pre-planned analysis. A full color version of this figure is available at the *Spinal Cord* journal online.

Post hoc analysis

After the analysis of the total score predictions, six outliers were identified and removed, as per the definition in the Methods. These outliers were based on the classifiers using only GR-str scores as inputs. The analysis was repeated on the reduced data set, and the results contained two outliers. These were again removed and the analysis repeated. After this third pass, no more outliers were found. The final reduced data set therefore contained 121 assessments, compared with 129 in the original data set (that is, 6.2% of the data were removed). The GR-str and GR-pp scores of the removed outliers are provided in the Supplementary Materials (Supplementary Table S1).

Figure 4 provides the analogous results to Figure 2 for the classifiers re-trained on the reduced data set, using only the GR-str scores as inputs. Out of 726 predictions (121 assessments × 6 sub-scores), 452 were correct (62.3%), 148 (20.4%) differed from the correct score by 1 point and 126 (17.4%) differed by 2 or more points, resulting in 82.7% predictions within 1 point of the correct score. When both GR-str and GR-sens scores were used as inputs, the error distribution was 60.3% with no error, 20.7% with a 1-point error and 19.0% with an error of 2 or more points. Once again, a χ^2 test revealed no significant difference between the performance of the two input strategies ($P=0.92$).

Figure 5 provides the analogous results to Figure 3 for the classifiers re-trained on the reduced data set, using only the GR-str inputs. In this case, the median error was 0 (IQR = − 2–3), and the Spearman's ρ was 0.92. When both the GR-str and GR-sens scores were included as inputs to the classifiers, the median error was 0 (IQR = − 2–4), which was not significantly different than when GR-str only was used ($P=0.94$ using a Wilcoxon rank sum test). The Spearman's ρ when using both GR-str and GR-sens was 0.91.

Comparison with predictions from neurological and motor levels

The breakdown of neurological levels in the data set used was 8 entries with C1, 8 with C2, 8 with C3, 39 with C4, 23 with C5, 20 with C6, 6 with C7, 2 with C8 and 1 with T3. GR-pp total score predictions based only on the neurological level yielded a Spearman's ρ of 0.16.

When using the right motor level, the breakdown in the data set was 3 entries with C1, 3 with C2, 2 with C3, 12 with C4, 25 with C5, 33 with C6, 21 with C7, 5 with C8, 8 with T1, 1 with L5 and 2 with S1, and the Spearman's ρ obtained was 0.35.

Fourteen entries out of 129 were excluded from these analyses of the neurological and motor level predictive abilities, because of missing ISNCSCI data.

DISCUSSION

This study demonstrated that functional task performance can be predicted with high accuracy from upper limb motor impairment, in unilateral UE tasks after SCI. Using the GRASSP assessment as a

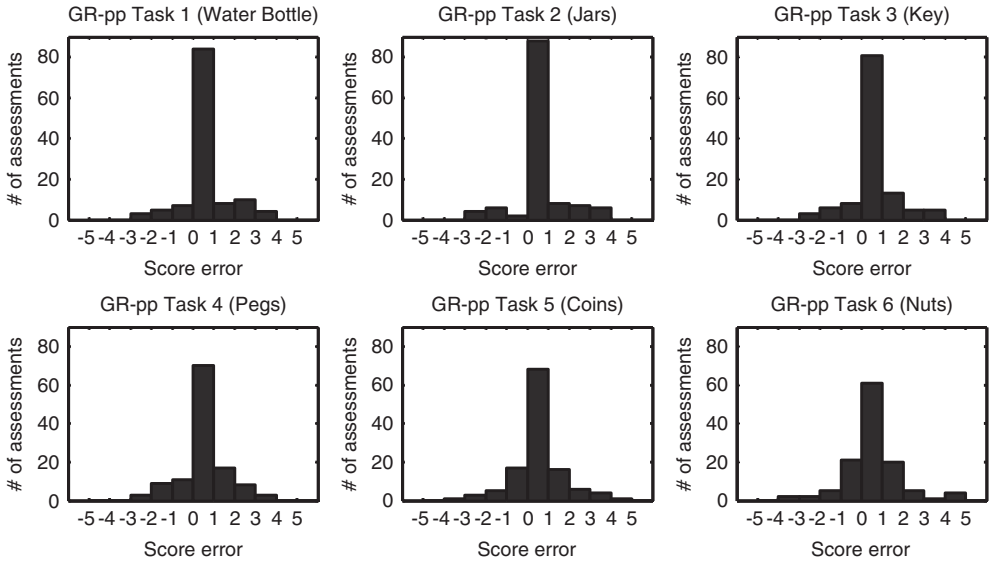

Figure 4 Histograms of errors between the actual and predicted sub-scores, for each of the six classifiers. The results in this figure correspond to the *post hoc* analysis. A full color version of this figure is available at the *Spinal Cord* journal online.

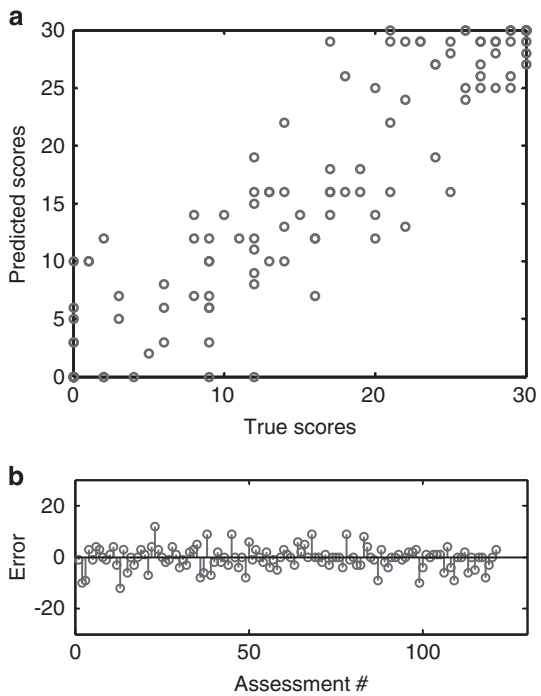

Figure 5 (**a**) Predicted vs actual total GR-pp scores. (**b**) Error between the predicted and actual total score for each of the assessments in the data set. The results in this figure correspond to the *post hoc* analysis. A full color version of this figure is available at the *Spinal Cord* journal online.

framework for this investigation, we found a very strong correlation between the predicted and true total GR-pp scores, as indicated by a Spearman's ρ of 0.84. We further investigated whether predictions based on both motor and sensory evaluations would be more successful than predictions based on motor evaluations alone but did not find this to be the case. We additionally demonstrated that a fine-grained description of impairment is beneficial for predicting task performance, by showing that predictions based only on neurological or motor levels were substantially poorer.

Implications for automated assessments of UE function

Our results have important implications in two areas. The first of these is the development of sensor-based automated assessments of UE function after SCI. In particular, our findings can inform the choice of variables that automated systems need to measure. If predicting functional task performance from impairment was not feasible, automated assessments would require highly sophisticated or specialized systems able to interpret complex movements and evaluate their quality. However, our results suggest that these complex approaches may be unnecessary: because task performance can be predicted from impairment in this context, automated systems need only to measure impairment. Therefore, combinations of existing and flexible technologies, such as dynamometers and low-cost motion capture systems, could form the basis for effective automated assessments of UE function after SCI, provided that strategies to manage confounding factors could be put into place.

Our findings regarding the need for sensory information are also encouraging. Although sensory function is fundamental to a complete neurological assessment, as well as to effective prehension,[28,29] we focused here on a more restricted question. We sought to establish whether sensory information added substantially to motor information for predicting functional task performance within the resolution of the GRASSP. We found that it did not. If sensory function was in fact required for these predictions, it would be detrimental to the development of automated assessments, as sensory evaluation requires a degree of conversation and interaction with the individual under evaluation, which would not be feasible in a reliable manner with an automated system. Our results on this point therefore further support the feasibility of automated assessments. In our analysis, the inclusion of sensory information actually resulted in a very slight decrease in performance compared with using the motor information alone; this phenomenon could be the result of the increased dimensionality of the classifier inputs, leading to some possible model overfitting.

When comparing the prediction performance for each of the six tasks in the GR-pp module (Figures 2 and 4), no strong trends were noted. Slightly higher errors were observed for coin handling and putting nuts on screws (tasks 5 and 6), whereas the best performance was found for the water bottle and jar lid tasks (tasks 1 and 2),

indicating that performance on tasks requiring fine manipulation may be harder to predict from the data available here.

The number of studies that have investigated automated UE assessments in SCI is so far relatively small. Our group previously found that range-of-motion and grip strength measures from the ArmeoSpring rehabilitation device could be used to predict concurrent scores from the GRASSP, SCIM and the Action Research Arm Test (ARAT), with adjusted R^2 values ranging from 0.54 to 0.78.[18] Those results are consistent with the findings reported here. Keller *et al.*[20] have used the ARMin robot to quantify the psychometric properties of data collected by the device, including range-of-motion, reachable workspace, quality of movement, joint torques and joint stiffness. They concluded that measures from the ARMin would be suitable for inclusion in automated assessments of UE function after SCI. Prochazka and Kowalczewski[19] proposed an automated test, based on the ReJoyce rehabilitation robot, and found a strong relationship with the ARAT ($R^2 = 0.88$). Their approach is a good example of an automated system designed to measure task performance directly: the robot includes a specialized manipulandum replicating several common grip tasks (cylindrical grasp, key grip and so on). Trincado-Alonso and colleagues[21] have correlated UE kinematic measurements from inertial measurement units with clinical scales, whereas Popp *et al.*[22] used inertial measurement units to monitor active propulsion during wheelchair use after SCI. The number of related studies conducted in clinical populations other than SCI is higher (a review is available in the study by de los Reyes-Guzman *et al.*,[30] with the majority of studies focusing on stroke), but the feasibility of generalizing automated approaches across populations has previously been put in doubt.[18]

Implications for the interpretation of manual assessments of UE function

The second implication of our results is in the interpretation and use of manual assessments of UE function, such as the GRASSP. The finding that impairment can accurately predict functional task performance can enrich the interpretation of motor testing scores and help better understand their implications. Likewise, the ability to predict how motor recovery in specific muscles is likely to affect functional performance has important implications for the planning of care and interventions. This ability may also lead to the development of reduced or adaptive versions of UE assessments, in which motor testing can inform which aspects of function need to be tested in more detail. Finally, characterizing the relationships between impairment and functional task performance is needed to guide the development of more effective interventions and is therefore of broad interest to both scientists and clinicians in the fields of SCI and rehabilitation.

Sources of variability in the data

We used a data set containing assessments from individuals with a wide range of characteristics, including varied injury levels, AIS grades, times since injury and hand dominance. The size of the retrospective data set is much larger than the sample sizes used in the automated assessment studies listed above. The use of such a large and heterogeneous population is beneficial in allowing the classifiers to identify robust relationships that will remain valid for a wide range of individuals.

The improved results of the *post hoc* analysis compared with the initial attempt suggest that confounding factors may cause an assessment to significantly deviate from the relationships captured in the classifiers. After removing a small number of outliers, corresponding to only 6.2% of the data set, performance improved noticeably, with the Spearman's ρ for the total scores increasing from 0.84 to 0.92. The outliers likely correspond to evaluations that should have been excluded but were not identified because of insufficient comments in the records. The outlier scores are listed in Supplementary Table S1 and support the notion that confounding factors were present, as many of the scores are counter-intuitive (for example, lower than expected GR-pp scores given the GR-str scores or *vice versa*). Factors that could have contributed to these outliers in the data include additional support provided to the right arm during manipulation tasks, movement restrictions resulting from pain or other injuries, or irregularities in scoring. Although the use of a *post hoc* analysis with outliers removed may raise questions about the generalizability of the results, the initial analysis with the outliers included still revealed an excellent correlation between the predicted and actual total GR-pp scores. As a result, none of our conclusions are contingent on the *post hoc* analysis. We have included it rather in an effort to quantify how our results may have been impacted by the use of retrospective data, as the data collection procedures were not tailored specifically to the question investigated here.

Even assuming that all potential confounding factors have been removed from the data set, substantial variability is still expected to remain in the relationship between impairment and functional task performance. Compensatory movement strategies are expected to have a role, as is the fact that the GRASSP can only capture partial information: not all muscles in the UE are measured, and GR-str, GR-sens and GR-pp scores all rely on ordinal scales that by definition are limited in their resolution. Participant motivation and fatigue may also have a role. These factors can explain the instances where our predictions had large errors. Further work will be required to identify additional variables that can improve the predictive performance and reduce the number of cases where large errors are observed. At the same time, our results demonstrated that prediction of functional task performance from impairment measures is possible to a significant degree even with classifiers that do not incorporate any of the factors just listed. This finding is novel and of great relevance to the development of automated assessments.

Limitations

This study was conducted using a retrospective data set, which limited our ability to control exactly how the GRASSP assessments were performed. For example, in participants with proximal weakness but some preserved distal function (such as a central cord injury), the arm was sometimes supported so that the hand could be tested. Although this was judged acceptable for the purposes of validating the GRASSP, it could skew the results of our classifiers, which might believe based on the motor scores that the individual cannot reach forward and possibly conclude that low scores should be ascribed for all of the tasks. Including such an assessment in our analysis would weaken the ability of the machine learning to identify meaningful relationships. The inconsistent presence of notes in the data set made it impossible to define precise *a priori* rules for the inclusion or exclusion of entries. The *post hoc* analysis was intended to mitigate these factors, but cannot guarantee that all data points were excluded that should have been, or that some outliers were not simply owing to natural variability.

We further emphasize that the *post hoc* analysis was conducted solely to compensate for the use of a retrospective data set containing entries skewed by confounding factors. In a prospective study, these data points would have been eliminated before analysis through exclusion criteria. The removal of outliers is not intended to imply that the proposed prediction methods are applicable only to spinal cord injuries with certain neurological characteristics but not others.

The use of the GRASSP limits our investigation to types of functional performance that are included in that assessment. For instance, bi-manual task performance is not within the scope of the

present analysis. However, our objective was to determine the extent to which fundamental relationships between impairment and functional performance could be captured through machine learning. Now that this has been established, the prediction of performance on specific types of functions or tasks will require the development of automated assessments tailored to the context of interest.

CONCLUSION

The ability to perform unilateral functional UE tasks after SCI can be predicted from motor testing scores (Spearman's ρ of 0.84 between predicted and actual total GR-pp scores or 0.92 in the *post hoc* analysis with outliers removed). This finding provides insight into the relationships between impairment and functional performance after SCI. Indeed, these results have important implications for the development of automated assessments of UE function after SCI, because the findings suggest that it is not necessary for such systems to address the difficult task of measuring functional task performance directly. Rather, systems that rely on measures of impairment (strength, range of motion) obtainable using existing technology could be feasible. Automated UE assessments will have benefits in optimizing rehabilitation in the sub-acute phase of injury, supporting telerehabilitation interventions and designing more efficient clinical trials.

DATA ARCHIVING

There were no data to deposit.

CONFLICT OF INTEREST

Dr S Kalsi-Ryan, Prof MC Verrier, Dr A Curt and Dr MG Fehlings are part of the GRASSP development team, which receives royalties for sales and licensing of the GRASSP. Dr S Kalsi-Ryan is Director of Neural Outcomes Consulting, Inc., which manufactures the GRASSP. Royalties and sales from the GRASSP are modest and partially cover the development costs of this outcomes tool. Dr J Zariffa declares no conflict of interest.

ACKNOWLEDGEMENTS

The studies during which the GRASSP data sets were created were generously supported by the Dana and Christopher Reeve Foundation, the Ontario Neurotrauma Foundation, the Rick Hansen Institute, the Craig H Neilsen Foundation, the Canadian Institutes of Health Research and the Physiotherapy Foundation of Canada. Dr MG Fehlings acknowledges support from the Halbert Chair in Neural Repair and Regeneration and the Dezwirek Foundation.

1 Consortium for Spinal Cord Medicine. Outcomes Following Traumatic Spinal cord Injury: Clinical Practice Guidelines For Health-care Professionals. *J Spinal Cord Med* 2000; **23**: 289–316.
2 Popovic MR, Kapadia N, Zivanovic V, Furlan JC, Craven BC, McGillivray C. Functional electrical stimulation therapy of voluntary grasping versus only conventional rehabilitation for patients with subacute incomplete tetraplegia: A randomized clinical trial. *Neurorehabil Neural Repair* 2011; **25**: 433–442.
3 Grossman RG, Fehlings M, Frankowski R, Burau KD, Chow DS, Tator C *et al.* A prospective multicenter phase 1 matched comparison group trial of safety, pharmacokinetics, and preliminary efficacy of riluzole in patients with traumatic spinal cord injury. *J Neurotrauma* 2014; **31**: 239–255.
4 Guzman R, Schubert M, Keller-Lang D, Huhn SL, Curt A. 196 human neural stem cell transplantation in chronic SCI: Interim results of a phase I/II trial. *Neurosurgery* 2013; **60**(Suppl 1): 185.
5 Zariffa J, Kapadia N, Kramer JL, Taylor P, Alizadeh-Meghrazi M, Zivanovic V *et al.* Feasibility and efficacy of upper limb robotic rehabilitation in a subacute cervical spinal cord injury population. *Spinal Cord* 2012; **50**: 220–226.
6 Mackinnon SE, Yee A, Ray WZ. Nerve transfers for the restoration of hand function after spinal cord injury. *J Neurosurg* 2012; **117**: 176–185.
7 World Health Organization. *International Classification of Functioning, Disability and Health (ICF)*. World Health Organization, 2001.
8 Kirshblum SC, Burns SP, Biering-Sorensen F, Donovan W, Graves DE, Jha A *et al.* International standards for neurological classification of spinal cord injury (revised 2011). *J Spinal Cord Med* 2011; **34**: 535–546.
9 Kalsi-Ryan S, Beaton D, Curt A, Duff S, Popovic MR, Rudhe C *et al.* The graded redefined assessment of strength sensibility and prehension: reliability and validity. *J Neurotrauma* 2012; **29**: 905–914.
10 Marino RJ, Kern SB, Leiby B, Schmidt-Read M, Mulcahey MJ. Reliability and validity of the capabilities of upper extremity test (CUE-T) in subjects with chronic spinal cord injury. *J Spinal Cord Med* 2014; **38**: 498–504.
11 Kapadia N, Zivanovic V, Verrier M, Popovic MR. Toronto rehabilitation Institute–Hand function test: Assessment of gross motor function in individuals with spinal cord injury. *Top Spinal Cord Inj Rehabil* 2012; **18**: 167–186.
12 Catz A, Itzkovich M, Tesio L, Biering-Sorensen F, Weeks C, Laramee MT *et al.* A multicenter international study on the spinal cord independence measure, version III: Rasch psychometric validation. *Spinal Cord* 2007; **45**: 275–291.
13 Hitzig SL, Noreau L, Balioussis C, Routhier F, Kairy D, Craven BC. The development of the spinal cord injury participation and quality of life (PAR-QoL) tool-kit. *Disabil Rehabil* 2013; **35**: 1408–1414.
14 Oleson CV, Marino RJ. Responsiveness and concurrent validity of the revised capabilities of upper extremity-questionnaire (CUE-Q) in patients with acute tetraplegia. *Spinal Cord* 2014; **52**: 625–628.
15 Kozlowski AJ, Heinemann AW. Using individual growth curve models to predict recovery and activities of daily living after spinal cord injury: An SCIRehab project study. *Arch Phys Med Rehabil* 2013; **94**(4 Suppl): 4.
16 Pretz CR, Kozlowski AJ, Charlifue S, Chen Y, Heinemann AW. Using rasch motor FIM individual growth curves to inform clinical decisions for persons with paraplegia. *Spinal Cord* 2014; **52**: 671–676.
17 Forsyth R, Thuy V, Salorio C, Christensen J, Holford N. Review: Efficient rehabilitation trial designs using disease progress modeling: A pediatric traumatic brain injury example. *Neurorehabil Neural Repair* 2010; **24**: 225–234.
18 Zariffa J, Kapadia N, Kramer JL, Taylor P, Alizadeh-Meghrazi M, Zivanovic V *et al.* Relationship between clinical assessments of function and measurements from an upper-limb robotic rehabilitation device in cervical spinal cord injury. *IEEE Trans Neural Syst Rehabil Eng* 2012; **20**: 341–350.
19 Prochazka A, Kowalczewski J. A fully automated, quantitative test of upper limb function. *J Mot Behav* 2015; **47**: 19–28.
20 Keller U, Schölch S, Albisser U, Rudhe C, Curt A, Riener R *et al.* Robot-assisted arm assessments in spinal cord injured patients: A consideration of concept study. *PloS ONE* 2015; **10**: e0126948.
21 Trincado-Alonso F, Dimbwadyo-Terrer I, de los Reyes-Guzman A, Lopez-Monteagudo P, Bernal-Sahun A, Gil-Agudo A. Kinematic metrics based on the virtual reality system toyra as an assessment of the upper limb rehabilitation in people with spinal cord injury. *Biomed Res Int* 2014; **2014**: 904985.
22 Popp WL, Brogioli M, Leuenberger K, Albisser U, Frotzler A, Curt A *et al.* A novel algorithm for detecting active propulsion in wheelchair users following spinal cord injury. *Med Eng Phys* 2016; **38**: 267–274.
23 Kalsi-Ryan S, Beaton D, Ahn H, Askes H, Drew B, Curt A *et al.* Responsiveness, sensitivity, and minimally detectable difference of the graded and redefined assessment of strength, sensibility, and prehension, version 1.0. *J Neurotrauma* 2015; **33**: 307–314.
24 Sollerman C, Ejeskar A. Sollerman hand function test. A standardised method and its use in tetraplegic patients. *Scand J Plast Reconstr Surg Hand Surg* 1995; **29**: 167–176.
25 Kalsi-Ryan S, Beaton D, Curt A, Popovic MR, Verrier MC, Fehlings MG. Outcome of the upper limb in cervical spinal cord injury: Profiles of recovery and insights for clinical studies. *J Spinal Cord Med* 2014; **37**: 503–510.
26 Kalsi-Ryan S, Beaton D, Ahn H, Askes H, Drew B, Curt A *et al.* Responsiveness, sensitivity and minimally detectable difference of the graded and redefined assessment of strength, sensibility, and prehension, version 1.0 (GRASSP V1). *J Neurotrauma* **33**: 307–314.
27 Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. Springer: New York. 2009. http://dx.doi.org/10.1007/978-0-387-84858-7.
28 Kalsi-Ryan S, Beaton D, Curt A, Duff S, Jiang D, Popovic MR *et al.* Defining the role of sensation, strength, and prehension for upper limb function in cervical spinal cord injury. *Neurorehabil Neural Repair* 2014; **28**: 66–74.
29 Johansson RS, Flanagan JR. Coding and use of tactile signals from the fingertips in object manipulation tasks. *Nat Rev Neurosci* 2009; **10**: 345–359.
30 de los Reyes-Guzman A, Dimbwadyo-Terrer I, Trincado-Alonso F, Monasterio-Huelin F, Torricelli D, Gil-Agudo A. Quantitative assessment based on kinematic measures of functional impairments during upper extremity movements: A review. *Clin Biomech (Bristol, Avon)* 2014; **29**: 719–727.

Supplementary Information accompanies this paper on the Spinal Cord website (http://www.nature.com/sc)