This work demonstrated the feasibility of performing accurate and reliable quantitative assessments of individual motor impairments in the dominant arm through tasks performed at home without supervision by the researchers.

**Why this paper**: This, naturally, makes it difficult to perform such assessments at scale, longitudinally, or with populations that cannot easily access testing facilities. Further, lab-based assessments may not be representative of real-world performance [35, 54], thus limiting the value of such assessments for tailoring accessibility solutions. However, attempts to collect other measures using passive approaches have been less successful [65]. Beyond all of these approaches, other passive digital phenotyping approaches have not been adapted yet for individuals with substantial impairments [12, 17]. Active digital phenotyping in the wild requires explicit effort on the part of the participants but it has been shown to produce results that replicate those obtained in conventional laboratory settings for both the general population [19, 31, 50, 56] and for individuals with unusual abilities, such as children, the elderly, and people with impairments [28, 49]. These results demonstrate that data obtained using active digital phenotyping approaches in unsupervised settings are valid when aggregated over a large number of individuals. **To help improve our understanding of the value of unsupervised active digital phenotyping for making accurate individual assessments of motor impairments, we conducted a study to analyze the validity, test-retest reliability, and acceptability of at-home use of a current active digital phenotyping system called Hevelius.**

**Main point:** Taken together, our results demonstrate that although single measurements of motor impairments collected in unsupervised settings can have higher variance than those collected in conventional laboratory settings, the simple approach of aggregating (taking a median of) a small number of consecutive unsupervised sessions can be sufficient to produce results that are as accurate or better than those obtained in supervised settings and that have high test-retest reliability. **For that reason, we used the dominant arm component of BARS as the clinical ground truth against which we evaluated Hevelius.**

Hevelius: Because motor performance changes substantially throughout a person’s lifetime (see, e.g., [14]), using age-specific z-scores makes it possible to separate the effects of a medical condition from the effects of development and aging.

**Collection of normative data from a large number of diverse healthy participants.**

**Hevelius in clinic:** Patients used Hevelius on the same day as their scheduled visit with a neurologist; thus, same-day assessments of disease severity were available.

**Hevelius at Home:** For this project, we built on the version of Hevelius that had been used in a clinical setting and adapted it for use at home without supervision by the researchers.

**Conclusion:** As expected, data from a single unsupervised session matched the clinician-assigned scores less accurately than the data obtained during a single session supervised by a researcher. However, aggregating data from just two consecutive unsupervised sessions was sufficient to make the BARS estimates as accurate as those obtained in a supervised setting. Increasing the number of consecutive unsupervised sessions that were aggregated further improved the accuracy of the score estimates and the test-retest reliability of both BARS score estimates and of the individual measures.

We examined the possibility that the session-to-session variability in each participant’s performance could be explained at least in part by changes in participants’ mood and energy levels. However, neither caregiver reports of participant state nor participant self-reports (nor the combination of the two) were significantly associated with within-participant session-to-session differences in BARS score estimates. In future work, we will collect more qualitative data to try to identify possible causes of session-to-session variability.

To reduce the number and frequency of explicit measurements, one future direction would be to develop methodologies that combine active and passive phenotyping. As mentioned earlier, Hevelius exemplifies active digital phenotyping as it requires participants to perform carefully specified tasks while the measurements are being collected. Passive phenotyping techniques use mobile phones or specialized wearable devices to unobtrusively collect data while people go about their natural activities. A solution that combined both approaches could leverage active phenotyping for infrequent but accurate measurements that could be used to calibrate and interpret more frequent data from passive measurements.

These limitations notwithstanding, we consider it a strength of this work that it engaged with a rare disease and a pediatric population, both of which have relatively little representation in the literature. We also note that techniques for performing accurate quantitative assessments of motor behavior have the potential to be particularly valuable for supporting remote care, research, and clinical trials for patients with rare neurological disorders given that most of these patients live far away from specialists knowledgeable about their disease.

Today’s paper is kinda a concrete example of what to do in research.

Different sort of task designs which capture motor and movement vs cognitive. Intelligent task design will be focus of research.

Do you wanna have standardized quality data in supervised vs low quality data in unsupervised data?

Challenge with HCI is finding participants

In HCI, you do inter-rater reliability

How do you find pre-symptomatic patients? Familial, Genetic etc.