JMIR Neurotechnology

For research exploring how technologies (e.g. information technology, neural engineering, neural interfacing, clinical data science, robotics, eHealth/mHealth) can be applied in clinical neuroscience (e.g., neurology, neurosurgery, neuroradiology) to prevent, diagnose, and treat neurological disorders.
Volume 2 (2023) ISSN 2817-092X Editors-in-Chief: Pieter Kubben MD, PhD

Contents

Review

Original Papers

Review

Remote Testing Apps for Multiple Sclerosis Patients: Scoping Review of Published Articles and Systematic Search and Review of Public Smartphone Apps

Jacob B Michaud[1,2], MD; Cameron Penny[3], BSc; Olivia Cull[2], BSc; Eric Hervet[4], PhD; Ludivine Chamard-Witkowski[2,5], MD

[1]Department of Internal Medicine, Dalhousie University, Halifax, NS, Canada

[2]Centre de Formation Médicale du Nouveau-Brunswick, Moncton, NB, Canada

[3]Faculty of Medicine, Dalhousie University, Halifax, NS, Canada

[4]Department of Computer Science, Université de Moncton, Moncton, NB, Canada

[5]Department of Neurology, Dr.-Georges-L.-Dumont University Hospital Center, Moncton, NB, Canada

Corresponding Author:
Ludivine Chamard-Witkowski, MD
Department of Neurology
Dr.-Georges-L.-Dumont University Hospital Center
330 Université Ave
Moncton, NB, E1C 2Z3
Canada
Phone: 1 506 869 721
Email: Ludivine.Witkowski@vitalitenb.ca

Abstract

Background: Many apps have been designed to remotely assess clinical status and monitor symptom evolution in persons with multiple sclerosis (MS). These may one day serve as an adjunct for in-person assessment of persons with MS, providing valuable insight into the disease course that is not well captured by cross-sectional snapshots obtained from clinic visits.

Objective: This study sought to review the current literature surrounding apps used for remote monitoring of persons with MS.

Methods: A scoping review of published articles was conducted to identify and evaluate the literature published regarding the use of apps for monitoring of persons with MS. PubMed/Medline, EMBASE, CINAHL, and Cochrane databases were searched from inception to January 2022. Cohort studies, feasibility studies, and randomized controlled trials were included in this review. All pediatric studies, single case studies, poster presentations, opinion pieces, and commentaries were excluded. Studies were assessed for risk of bias using the Scottish Intercollegiate Guidelines Network, when applicable. Key findings were grouped in categories (convergence to neurological exam, feasibility of implementation, impact of weather, and practice effect), and trends are presented. In a parallel systematic search, the Canadian Apple App Store and Google Play Store were searched to identify relevant apps that are available but have yet to be formally studied and published in peer-reviewed publications.

Results: We included 18 articles and 18 apps. Although many MS-related apps exist, only 10 apps had published literature supporting their use. Convergence between app-based testing and the neurological exam was examined in 12 articles. Most app-based tests focused on physical disability and cognition, although other domains such as ambulation, balance, visual acuity, and fatigue were also evaluated. Overall, correlations between the app versions of standardized tests and their traditional counterparts were moderate to strong. Some novel app-based tests had a stronger correlation with clinician-derived outcomes than traditional testing. App-based testing correlated well with the Multiple Sclerosis Functional Composite but less so with the Expanded Disability Status Scale; the latter correlated to a greater extent with patient quality of life questionnaire scores.

Conclusions: Although limited by a small number of included studies and study heterogeneity, the findings of this study suggest that app-based testing demonstrates adequate convergence to traditional in-person assessment and may be used as an adjunct to and perhaps in lieu of specific neurological exam metrics documented at clinic visits, particularly if the latter is not readily accessible for persons with MS.

(JMIR Neurotech 2023;2:e37944) doi:10.2196/37944

KEYWORDS

multiple sclerosis; mobile application; mobile phone; app; mHealth; eHealth; digital health; telehealth

Introduction

Multiple sclerosis (MS) has a fluctuating clinical course punctuated by relapses, remissions, and progressive deterioration for many affected patients. As such, the neurologist requires an accurate representation of the symptomatology of the patient with MS in order to evaluate the efficacy of treatment [1].

Infrequent and intermittent monitoring as provided at office visits may not truly reflect the day-to-day functioning and quality of life of patients living with MS [2]. Persons with MS may also have recall bias when reporting symptoms to their neurologist [2]. Additionally, symptoms in MS can fluctuate depending on fatigue, mood, and weather; thus, the cross-sectional nature of the information obtained from an individual clinic visit may be of limited accuracy compared with trends in symptoms over time [3,4]. The need for at-home MS follow-up has been further emphasized by the current COVID-19 pandemic, in which many medical centers have implemented in-person patient visit limits to reduce the spread of the virus [5].

Remote evaluation of clinical status and symptoms in persons with MS could serve as a means of obtaining additional information that is not provided by the traditional office visit. Many apps for remote assessment of persons with MS exist, ranging from symptom logs, patient-reported outcome trackers, assessments of cognitive function and fine motor skills, as well as drug adherence and adverse drug event trackers [6-8]. The objective of this review was to identify and evaluate apps designed to enable remote assessment of persons with MS and whether the means of assessment utilized in these various apps are supported by scientific evidence.

Methods

Review Sources

A scoping review was performed to identify articles evaluating apps dedicated to the remote testing and follow-up of persons with MS. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines were followed for this portion of the review [9]. No protocol for this review was previously published.

A separate review of the Canadian Apple App Store and Google Play Store was conducted in parallel. This was done in order to identify apps available for public use, including some identified in the literature search as well as those that had not been formally studied prior to dissemination.

Eligibility Criteria

Scientific papers were included if they met the following criteria: The study evaluated the use of remote monitoring of persons with MS via smartphone or tablet app and was published in English, French, or Spanish prior to January 17, 2022. Cohort studies, feasibility studies, and randomized controlled trials were included in this review. Studies were included if the application was used to measure one or more of the following functional domains: physical disability, fatigue, visual symptoms, urinary symptoms, balance, mood symptoms, pain, cognition, or ambulation. Exclusion criteria included pediatric studies, single case studies, poster presentations, opinion pieces, and commentaries.

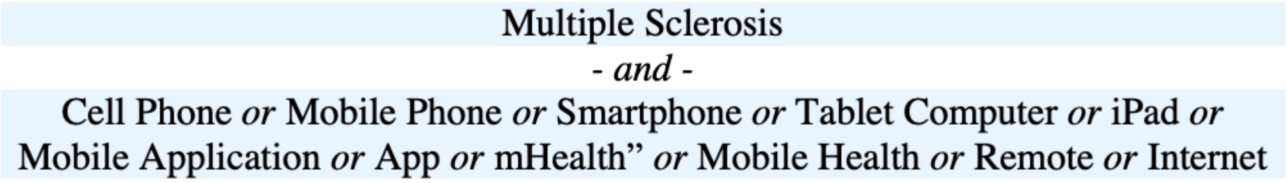
Publicly available apps that were intended for symptom tracking or app-based testing of persons with MS were included in the app review portion of this paper if they were able to measure one or more of the aforementioned metrics.

Search Strategy

PubMed/Medline, EMBASE, CINAHL, and Cochrane databases were searched from inception to January 17, 2022, to identify studies suitable for inclusion. The search strategy is detailed in Figure 1, and the detailed search strategy is presented in Multimedia Appendix 1.

As for the apps, the Canadian iOS Apple App Store and Android Google Play Store were searched using the term “Multiple Sclerosis” for publicly available apps.

Figure 1. Search strategy.



Multiple Sclerosis

- and -

Cell Phone *or* Mobile Phone *or* Smartphone *or* Tablet Computer *or* iPad *or*
Mobile Application *or* App *or* mHealth” *or* Mobile Health *or* Remote *or* Internet

Data Collection and Analysis

Two authors (JBM and CP) independently screened studies for the inclusion criteria based on title and abstract. The articles were then subject to an independent full-text review, and inclusion was determined by consensus. The references of included studies were screened to identify any additional articles suitable for inclusion that were not captured in the initial search strategy. The aforementioned authors collected data on application testing metrics as well as on convergence with standard neurological exam findings (Pearson correlation coefficients and linear mixed effects estimates), feasibility of implementation (qualitative assessment and adherence rates), weather analysis, and practice effect. Data collection also included participant age, diagnosis, baseline Expanded Disability Status Scale (EDSS), study design, study funding, and follow-up period. Authors JBM and OC assessed included articles for risk of bias using the Scottish Intercollegiate Guidelines Network (SIGN) checklist, when applicable [10]. Relevant articles were grouped in primary outcome categories, and data were presented qualitatively.

Authors JBM and CP independently screened the title and description of the apps, and inclusion was determined by consensus. The included apps were then reviewed, and the functional domains evaluated were documented.

Presentation

For the purpose of readability, this article considered correlation coefficients |*r*|≥0.75 to be strong, 0.75>|*r*|≥0.5 to be moderate, 0.5>|*r*|≥0.25 to be weak, and |*r*|<0.25 to not be correlated.

Results

Study and App Identification

A total of 2433 studies were identified using the search strategy defined in the Methods section. Following duplicate removal and title and abstract screening, 77 studies were selected for full-text review. Of these studies, 18 were in keeping with the predefined inclusion criteria (Figure 2). All 18 studies were found to be of acceptable or high quality using the SIGN checklist [10].

As for the app store review, the search yielded 79 apps in the Apple App Store and 339 apps in the Google Play Store. After removal of duplicates and title and description screening, 25 apps were selected for full app review. Of these apps, only 18 were deemed to fit the inclusion criteria (Figure 3). Of the 18 apps included, 2 had supporting literature that was identified in the scoping review portion of this paper [11-14].

Figure 2. Included articles.

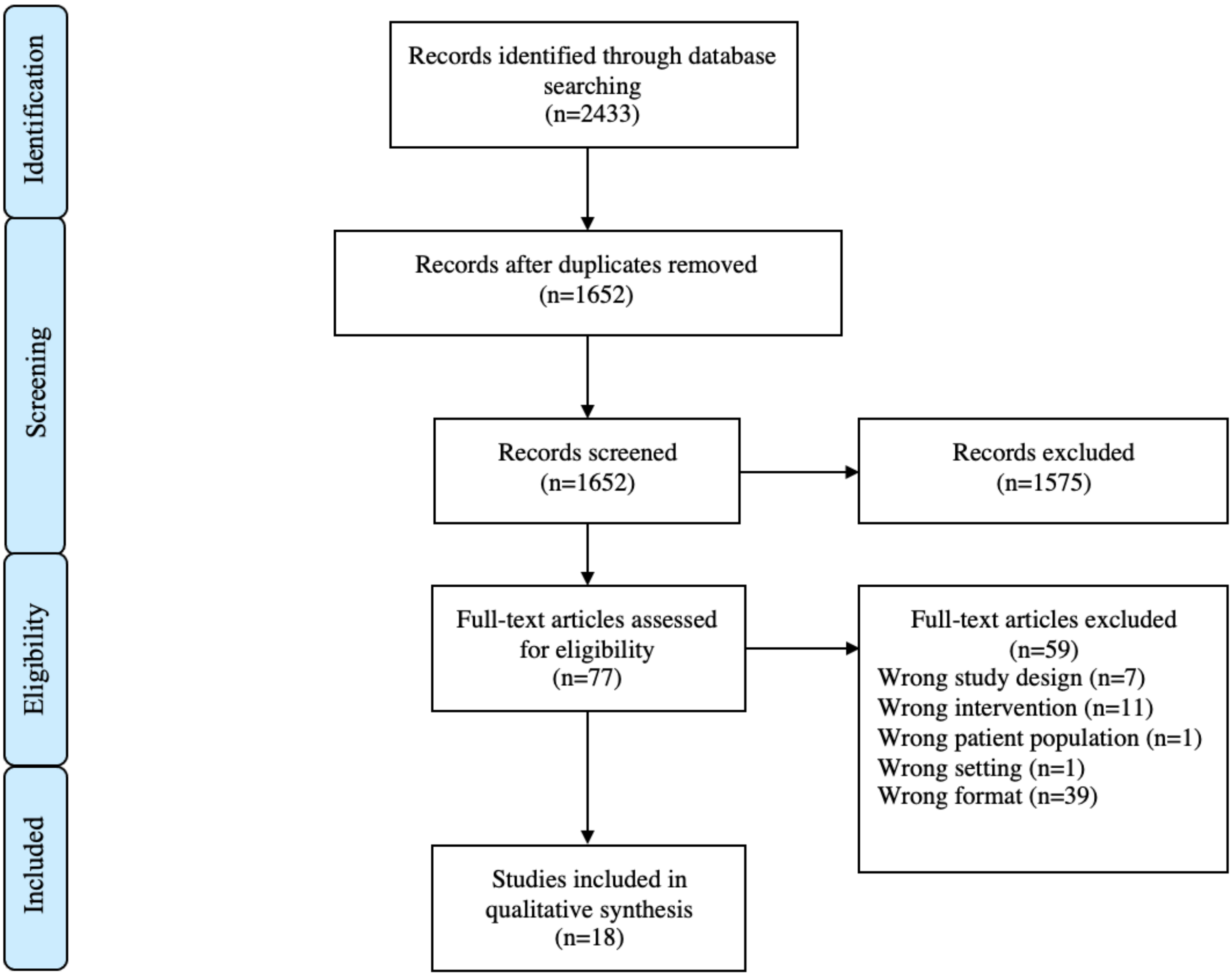
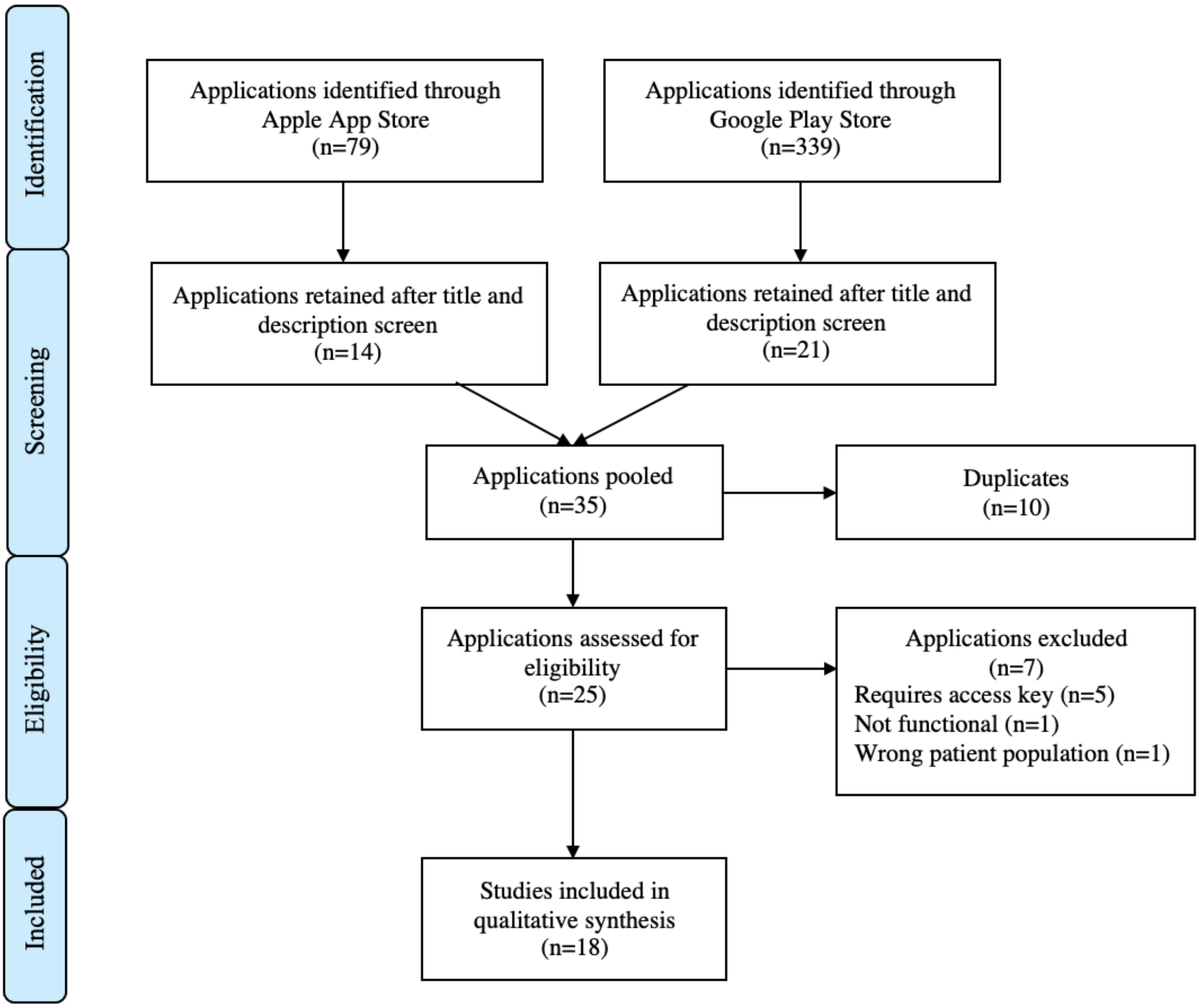

Identification
Screening
Eligibility
Included
Records identified through database searching (n=2433)
Records after duplicates removed (n=1652)
Records screened (n=1652)
Records excluded (n=1575)
Full-text articles assessed for eligibility (n=77)
Full-text articles excluded (n=59)
Wrong study design (n=7)
Wrong intervention (n=11)
Wrong patient population (n=1)
Wrong setting (n=1)
Wrong format (n=39)
Studies included in qualitative synthesis (n=18)

Figure 3. Included apps.


Identification
Screening
Eligibility
Included
Applications identified through Apple App Store (n=79)
Applications identified through Google Play Store (n=339)
Applications retained after title and description screen (n=14)
Applications retained after title and description screen (n=21)
Applications pooled (n=35)
Duplicates (n=10)
Applications assessed for eligibility (n=25)
Applications excluded (n=7)
Requires access key (n=5)
Not functional (n=1)
Wrong patient population (n=1)
Studies included in qualitative synthesis (n=18)

Characteristics of Included Studies

Of the 18 articles included, 12 sought to compare apps with a neurologist exam, disability scale, or recognized standardized tests [11-13,15-23]. The feasibility of implementing an app designed for remote monitoring of persons with MS was evaluated in 3 studies [24-26], 2 articles compared quality of life questionnaires with app-based functional tests and clinician-reported outcomes [25,27], and 2 apps assessed the local weather’s impact on persons with MS-reported fatigue and app functional test results [25,28]. Finally, 1 article evaluated the practice effect of repeated at-home MS testing [14] (Table 1).

Table 1. Characteristics of included studies.

Author(s), year	Countries	Study design	Type of multiple sclerosis
Hsu et al, 2021 [22]	United States	Prospective cohort	RR[a], PP[b], SP[c], CIS[d], unknown
Golan et al, 2021 [27]	Israel	Prospective cohort	RR, PP, SP
Pratap et al, 2020 [28]	United States	Prospective cohort	RR, PP, SP
Hsu et al, 2021 [15]	United States	Cross-sectional	RR, PP, SP, CIS
Newland et al, 2019 [26]	United States	Descriptive study	RR, SP
Midaglia et al, 2019 [24]	Spain, United States	Prospective cohort	RR, PP, SP
Montalban et al, 2021 [12]	Spain, United States	Prospective cohort	RR, PP, SP
Woelfle et al, 2021 [14]	Switzerland	Prospective cohort	N/A[e]
Lam et al, 2021 [18]	Netherlands	Prospective cohort	RR, PP, SP
van Oirschot et al, 2020 [19]	Netherlands	Prospective cohort	RR
van Oirschot et al, 2021 [23]	Netherlands	Prospective cohort	RR
Boukhvalova et al, 2018 [16]	United States	Cross-sectional	RR, PP, SP
Boukhavalova et al, 2019 [17]	United States	Prospective cohort	RR, PP, SP
Maillart et al, 2019 [11]	France	Crossover study	RR, PP
Tanoh et al, 2021 [13]	France	Prospective cohort	RR, PP
Bove et al, 2015 [25]	United States	Prospective cohort	RR, PP, SP, CIS
Lam et al, 2021 [20]	Netherlands	Prospective cohort	RR, PP, SP
Lam et al, 2022 [21]	Netherlands	Prospective cohort	RR, PP, SP

[a] RR: relapsing remitting multiple sclerosis.
[b] PP: primary progressive multiple sclerosis.
[c] SP: secondary progressive multiple sclerosis.
[d] CIS: clinically isolated syndrome.
[e] N/A: not available.

Characteristics of Included Apps

Of the 18 apps included, 5 had objective symptom testing through mobile phone sensors. The other 13 did not have active tests but did allow for symptom logging. Of the apps included in this study, 2 had complimentary data that were identified during the scoping review portion of the current study.

Four apps were exclusively found on the Apple App Store, 8 apps were exclusively found in the Google Play Store, and 6 apps were found in both stores. All but 2 of the apps included were free of charge.

Scoping Review Outcomes

As aforementioned, the reviewed articles were categorized according to 4 main objectives: evaluating convergence with neurological exam, feasibility of implementation of an app for persons with MS, evaluating the practice effect of repeated at-home testing, and comparing app-based tests with quality of life questionnaires and local weather.

Convergence With the Neurological Exam

Of the 18 articles, 14 articles described 12 apps that measured physical disability and correlated these with findings on clinical exam. These measures of physical disability were done by tap tests [16], shape drawing tests [11,13], pinching tests [12], assessment of passively acquired keyboard metrics [18,20], or using a level test wherein one must balance their phone in order to keep a ball in a designated screen area [17]. Visual symptoms were measured in 2 apps using tests of steering around obstacles [15] or a mobile vision test [11]. Cognitive function was measured in 6 apps: 3 apps used an electronic version of the Symbol Digit Modalities Test (SDMT) [11-13,18,19]; 1 used a go-no go test coupled with multitasking and visuomotor steering [15]; 1 used a battery of attention, working memory, and goal management evaluations [22]; and 1 measured keystroke dynamics including keystroke latency, emoji use, and word length [20,21]. Ambulation was measured in 3 apps using an app-based timed 25-foot walk test (T25FW) [11,13], 2-minute walk test (2MWT) [23], U-turn test [12], or maximum distance walked test [11,13]. The main tests and functional domains can be found in Table 2.

One study compared the MS Suite app balloon popping test to the 9-Hole Peg Test (9HPT) and found that the app slightly outperformed the 9HPT in its ability to correlate with clinician-derived outcomes [16]. The number of balloons popped correlated strongly with cerebellar function and moderately with upper extremity strength and motor exam. The study also included 4 patients who could no longer perform the 9HPT due to severe disease but were able to perform the balloon popping test.

Table 2. App tests from scientific articles and comparators for convergence with neurological exam or patient questionnaires.

App and functional domains	App test	Comparator
Adaptive Cognitive Evaluation [22]		
Cognition	Boxed task, sustained attention task, spatial span	SDMT[a]
ElevateMS [28]		
PD[b]	Finger tapping, finger to nose	PDDS[c], Neuro-QoL[d]
Ambulation, balance	Walk and balance test	PDDS, Neuro-QoL
Cognition	Voice-controlled DSST[e]	PDDS, Neuro-QoL
Evo Monitor [15]		
PD	Go/no go, tilt to steer, and combination of both tasks	MSFC-4[f], EDSS[g]
Cognition	Go/no go, tilt to steer, and combination of both tasks	BICAMS[h]
Floodlight [12,14]		
PD	Draw a shape, pinching test	9HPT[i], EDSS
Balance	Static balance test	BBS[j]
Cognition	sSDMT[k]	SDMT
Ambulation	2MWT[l], U-turn test	T25FW[m], EDSS
MSCopilot [11,13]		
PD	Spiral test	9HPT
Visual	Vision test	SLCLAT[n]
Cognition	Cognition test (sSDMT)	SDMT, PASAT[o]
Ambulation	Walking test	T25FW, EDSS
MS Sherpa [18,19,23]		
Cognition	sSDMT	SDMT
Ambulation	e-2MWT[p]	2MWT
MS Suite [16,17]		
PD	Balloon popping, tap test, tilt test	NeurEx[q], EDSS
Cognition	Tilt test	SDMT
NeuroKeys [20,21]		
PD	Press-press latency, release-release latency, hold time, flight time, precorrection slowing, correction duration, post correction slowing, after punctuation pause, emoji sentiment score [11]	EDSS, 9HPT
Cognition	Press-press latency, release-release latency, hold time, flight time, precorrection slowing, correction duration, post correction slowing, after punctuation pause, emoji sentiment score [11]	SDMT
Fatigue	Press-press latency, release-release latency, hold time, flight time, precorrection slowing, correction duration, post correction slowing, after punctuation pause, emoji sentiment score [11]	CIS-F[r]

[a] SDMT: Symbol Digit Modalities Test.

[b] PD: physical disability.

[c] PDSS: Patient-Determined Disease Steps.

[d] Neuro-QoL: Quality of Life in Neurological Disorders.

[e] DSST: Digit Symbol Substitution Test.

[f] MSFC-4: Multiple Sclerosis Functional Composite 4.

[g] EDSS: Expanded Disability Status Scale.

[h] BICAMS: Brief International Cognitive Assessment for Multiple Sclerosis.

[i] 9HPT: 9-Hole Peg Test.

[j] BBS: Berg Balance Scale.

[k] sSDMT: smartphone SDMT.

[l] 2MWT: 2-minute walk test.

[m] T25FW: timed 25-foot walking test.

[n] SLCLAT: Sloan Low Contrast Letter Acuity Test.

[o] PASAT: Paced Auditory Serial Addition Test.

[p] e-2MWT: electronic 2MWT.

[q] NeurEx: digitalized neurological examination.

[r] CIS-F: Checklist Individual Strength Fatigue subscale.

Keystroke dynamics were found to have weak correlation with the EDSS and moderate correlation with the SDMT in 1 study [20]. Another found that the use of emojis with more neutral sentiment as well as decreased word length were responsive to changes in the EDSS in a manner that was statistically significant [21].

One study evaluating the correlation of the smartphone SDMT (sSDMT) with the traditional SDMT found a moderate correlation for tests done in the morning and in the evening for the MS Sherpa app [18]. In 2 other studies comparing MS Sherpa's sSDMT as well as Floodlight's sSDMT to the traditional SDMT, strong correlations were found between these tests [12,19].

Two studies compared their app-based tests with the SDMT. The first compared the Evo Monitor multitasking test with SDMT and found a moderate correlation [15]. The second compared the SDMT and MS Suite level test, in which the time a virtual ball stayed in the center of the screen was found to correlate moderately with the SDMT [17]. These same 2 studies compared the multitasking test and level test with the EDSS. Both correlated weakly with the EDSS [15,17].

The MS Copilot app included several tests: spiral drawing test, maximum distance walked without stopping, verbal SDMT, and low contrast vision test. The *z* score of participants' test batteries correlated strongly with the Multiple Sclerosis Functional Composite (MSFC) *z* score [11]. Another MS Copilot battery comprising of maximum walking distance, shape drawing, and SDMT correlated moderately with the EDSS [13].

In 1 study, the Floodlight app's pinching test correlated moderately with the 9-HPT. It also found that Floodlight's U-turn test correlated moderately with the T25FW. Of the Floodlight tests, the U-turn test had the strongest correlation with the EDSS despite the weak correlation (r=–0.45; P<.001) [12]. Individual test scores were not compounded in this study as they were in the MS Copilot study [13].

Finally, MS Sherpa's smartphone 2MWT measurements were found to be approximately 8.43 meters greater than those measured traditionally. In this same study, there was no statistically significant correlation identified between the app-based 2MWT and EDSS [23].

Feasibility of Implementation

The feasibility of implementing an app to monitor symptoms in persons with MS was assessed in 3 studies. Adherence rates were 51% for an app requiring 12 months of daily data collection (n=38) [25]; 70% for an app requiring daily, weekly, fortnightly, or on-demand activities (n=76) [24]; and 87% for an app requiring 7 consecutive days of testing and a repeat test 4 weeks later (n=32) [26].

Quality of Life Questionnaires

App-based quality of life questionnaires were evaluated in 2 studies: 1 compared app-derived neurological quality of life (Neuro-QoL) questionnaires to in-app functional tests. Using a linear mixed effects model, the study found that the following Neuro-QoL domains correlated significantly with app tests: Upper extremity function was correlated with finger tapping test, lower extremity function was correlated with walk and balance tests, and cognitive function was correlated with the voice-based Digit Symbol Substitution Test (DSST) [28].

Another study assessed the e-Diary app, in which an app-based questionnaire was used to derive a Bodily Function Summary Score that was then compared to clinician-reported outcomes. This study found a strong correlation between the Bodily Function Summary Score and EDSS scores [27].

Weather

Whether increasing local temperature had a negative impact on in-app tests was evaluated in 2 studies [25,28]. The first included 495 persons with MS and found that increasing temperature had a significant negative impact on finger tapping, DSST, and finger to nose [28]. However, the second study, following 22 persons with MS, found no statistically significant association between the Modified Fatigue Inventory Scale and temperature or daylight hours [25].

Practice Effect

The development of a practice effect with repeated at-home app-based MS testing was assessed in 1 study. Data included in this study were derived from the Floodlight app. Domains assessed included daily repetition of finger pinching, shape drawing, 2MWT, U-turn test, static balance test, and weekly repetition of virtual SDMT. The study found improvement in test scores ranging from 11% to 54.2% on daily repetition of tests with the exception of the 2MWT. For the sSDMT, an average improvement of 40.8% was observed after 5 weeks of weekly testing [14].

The key findings of each included article are presented in Table 3.

Table 3. Key findings of included studies.

App and author, year	All assessed functional domains	Number of people with MS[a]	Key findings
Adaptive Cognitive Evaluation			
Hsu et al, 2021 [22]	Cognition	53	Boxed reaction time of persons with MS correlated most strongly with SDMT[b] (*r*=–0.50; *P*<.001), including when covariates were accounted for (*r*=–0.43; *P*=.002). Sustained attention span and spatial span were not significantly correlated with SDMT.
e-Diary			
Golan et al, 2021 [27]	PD[c], visual, urinary, mood, pain, cognition	97	e-diary–derived PROs[d] were significantly correlated with corresponding functional system scores (0.38<*r*<0.8; *P*<.001). The sum of bodily functions showed a correlation coefficient of *r*=0.77 (*P*<.001) with clinician EDSS[e].
ElevateMS			
Pratap et al, 2020 [28]	PD, balance, cognition, weather	495	Neuro-QoL[f] categories correlated significantly with finger tapping (β[g]=0.4; *P*<.001), walk and balance (β=–99.18; *P*=.02), and DSST[h] (β=1.60; *P*=.03). Baseline PDDS was associated with finger to nose (β=.01; *P*=.01). Increasing temperature had a significant impact on finger tapping, DSST (β=–.06; *P*=.009), and finger to nose.
Evo Monitor			
Hsu et al, 2021 [15]	PD, cognition	100	Evo Monitor multitasking test was associated with SDMT (*r*=0.52; *P*<.001), EDSS (*r*=–0.35; *P*<.01), and T25FW[i] (*r*=–0.41; *P*<.001). Go/no go and tilt to steer tests were associated with SDMT (*r*=–0.31; *P*=.001 and *r*=0.40; *P*<.001, respectively).
Fatigue			
Newland et al, 2019 [26]	PD, pain, cognition	32	Most participants (87%) completed all of the surveys required (7 consecutive days and repeat 4 weeks later).
Floodlight			
Midaglia et al, 2019 [24]	PD, fatigue, balance, mood, pain, cognition, ambulation	76	70% of participants were adherent to all active tests. 75% of participants were adherent to all tests except 2MWT[j]. Mean satisfaction with the app at week 12 was 74.1% and at week 24 was 73.7%.
Montalban et al, 2021 [12]	PD, balance, cognition	76	Strongest correlation was found between sSDMT[k] and SDMT (*r*=0.82, *P*<001). Pinching test correlated with 9HPT[l] (*r*=0.64, *P*<.001). U-turn test correlated with T25FW (*r*=–0.52, *P*<.001). Strongest correlation with EDSS was with U-turn test (*r*=–0.45, *P*<.001). Static balance test was not significantly associated with Berg Balance Scale.
Woelfle et al, 2021 [14]	PD, balance, cognition, ambulation	171-262	sSDMT, when repeated at 7-day intervals, had an average improvement of 40.8%. The practice effect was reached after 11 repetitions for one-half and after 35 repetitions for 90%. Finger pinching, draw a shape, U-turn, and static balance had average improvements of 54.2%, 23.9%, 11.0%, and 28.6%, respectively. 2MWT was not significantly associated with improvement.
MS Copilot			
Maillart et al, 2019 [11]	PD, visual, cognition, ambulation	141	App combined task *z* score correlated with the MSFC[m] *z* score (*r*=0.81; *P*<.001).
Tanoh et al, 2021 [13]	PD, visual, cognition, ambulation	116	Summed scores of maximum walking distance, draw a shape, and mobile SDMT correlated with EDSS (*r*=–0.65; *P*<.001).
MS Sherpa			
Lam et al, 2021 [18]	Cognition	102	sSDMT and SDMT correlation coefficients were *r*=0.687 (*P*<.001) in the morning and *r*=0.622 (*P*<.001) in the evening, with a regression coefficient of 0.87.
van Oirschot et al, 2020 [19]	Cognition	25	The interclass correlation coefficient between SDMT and sSDMT results was 0.784, and the Pearson correlation coefficient was *r*=0.85 (*P*<.001).
van Oirschot et al, 2021 [23]	Cognition, ambulation	25	Distance walked on e-2MWT was, on average, 8.43 meters greater than that with traditional 2MWT. There was no significant correlation between EDSS and e-2MWT.

App and author, year	All assessed functional domains	Number of people with MS[a]	Key findings
MS Suite			
Boukhvalova et al, 2018 [16]	PD, cognition	76	Balloon popping had correlation coefficients of *r*=0.62, *r*=0.75, and *r*=0.62 (*P*<.0001) with upper extremity strength, cerebellar function, and upper extremity motor exam, respectively. These values were *r*=0.59, *r*=0.57, and *r*=0.61 for the traditional 9HPT. Tap test was associated with 9HPT (*r*=0.66; *P*<.0001)
Boukhvalova et al, 2019 [17]	PD, cognition	112	Level test time spent in center of the level test correlated with SDMT (*r*=0.57; *P*<.0001) and, to a lesser degree, with EDSS (*r*=–0.35; *P*<.01).
N/A[n]			
Bove et al, 2015 [25]	PD, balance, cognition, weather	38	Adherence rate for the app was 51% at 12 months. Of those who completed the 1-year study (n=22), no significant association between MFIS[o] and temperature (*P*=.18) nor daylight hours (*P*=.091) was noted.
Neuro keys			
Lam et al, 2021 [20]	PD, cognition, fatigue	85	EDSS was most correlated with latency between key release (*r*=0.407, *P*<.001). Overall, the release-release latency keystroke metric correlated the most with SDMT (*r*=–0.553 *P*<.01).
Lam et al, 2022 [21]	PD, cognition	94	The keystroke features most responsive to changes in EDSS were emoji sentiment neutrality and word length, with AUCs[p] of 0.79 and 0.72, respectively.

[a]MS: multiple sclerosis.

[b]SDMT: Symbol Digit Modalities Test.

[c]PD: physical disability.

[d]PROs: patient-reported outcomes.

[e]EDSS: Expanded Disability Status Scale.

[g]Neuro-QoL: quality of life in neurological disorders.

[g]Linear mixed effects estimate.

[h]DSST: Digit Symbol Substitution Test.

[i]T25FW: timed 25-foot walk.

[j]2MWT: 2-minute walk test.

[k]sSDMT: smartphone SDMT.

[l]9HPT: 9-Hole Peg Test.

[m]MSFC: Multiple Sclerosis Functional Composite.

[n]N/A: not available.

[o]MFIS: Modified Fatigue Impact Scale.

[p]AUCs: areas under the curves.

App Review

Of the 18 identified apps, 5 had a remote testing function. Of the 5 apps with remote testing abilities, all tested for physical disability and fine motor skills. Assessment of motor skills was done through tapping tests as in BeCare and MS Care Connect; drawing a shape or following a path as in Floodlight, MS Care, and MS Copilot; or a 9HPT equivalent as in Neurons. With regard to disability, 1 app, BeCare, measured arm raises, while Floodlight measured pinch and thumb strength.

Visual symptoms were evaluated by 3 of the apps. This was done by contrast sensitivity tests and measured optokinetic nystagmus as in BeCare, color vision tests as in MS Care Connect, or low-contrast vision tests as in MS Copilot.

Cognitive testing was performed in all 5 apps: 4 apps (BeCare, Floodlight, MS Care Connect, and MS Copilot) used the SDMT; 2 apps used modified versions of recognized MS tests like the Paced Auditory and Visual Serial Addition Test as in Neurons and the Stroop test as in BeCare; and some apps used other tests like stacking donuts in ascending size on pegs, memorizing words and matching them to categories, and tapping blocks in a memorized sequence as in MS Care Connect or memorizing animals as in BeCare.

All 5 apps had measures of ambulation: 3 apps (BeCare, Neurons, and MS Care Connect) had the T25FW, and 2 apps had time-limited walk tests such as BeCare's 6-minute walk test or Floodlight's 2MWT. BeCare also measured the Timed Up and Go test. Floodlight implemented passive monitoring of daily ambulation, while MS Copilot measured maximum distance walked.

Only 1 app, Floodlight, had a dedicated static balance test. Another app, MS Care Connect, measured reaction time. The BeCare app measured the ability to discriminate between mobile device vibration frequency. That same app also had an audio transcription test.

Symptom logging functions were found in 13 other apps, either through free-text entry or selecting within a list of suggested neurological symptoms. These are included in Table 4.

Table 4. Characteristics of included apps.

App name	Platform	Developer	Brief description
Aby	Both	Biogen Inc	Log MS[a] symptoms
Bearable - Symptom and Mood Tracker	GPS[b]	Bearable	Log MS symptoms
BeCare MS Link	Both	BeCare Link LLC	Testing for PD[c], visual, cognitive, ambulation, mood
Emilyn: My MS Companion	Both	BreakthroughX Health GmbH	Log MS symptoms
Floodlight[d]	Both	Roche SAS	Log MS symptoms; testing for PD, cognitive, balance, ambulation
Healthstories MS	AAS[e]	Jacob Wachsman	Log MS symptoms
icompanion	Both	Icometrix Inc	Log MS symptoms, may perform prEDSS[f] or Neuro-QoL[g]
Innov SEP	GPS	Mallouki Adil	Log MS symptoms.
MSAA-My MS Manager	AAS	At Point of Care, LLC	Log MS symptoms, generate MFIS[h] score
MS Care Connect	GPS	InterPro Bioscience Inc	Log MS symptoms; testing for PD, cognitive, ambulation
MSCopilot[d]	GPS	Ad Scientiam	Testing for PD, visual, cognitive, ambulation
MS Corner	GPS	Progentec Diagnostics	Log MS symptoms
MS Notes Journal	GPS	Roger Hartley	Log MS symptoms
MS Relapse Tool	Both	Darin Okuda	Log MS symptoms
MS Relapse Tracker/MS Attack	AAS	Flavia Chapa	Log MS symptoms, relapse probability assessment
Multiple Sclerosis Manager	GPS[f]	KingFishApps	Log MS symptoms.
Multiple Sclerosis Messenger	GPS	KingFishApps	Log MS symptoms and may send to MS nurse
Neurons	AAS	shazino	Testing for PD, cognitive, ambulation

[a] MS: multiple sclerosis.

[b] GPS: Google Play Store.

[c] PD: physical disability.

[d] App found to have supporting literature in the scoping review of scientific evidence.

[e] AAS: Apple App Store.

[f] prEDSS: patient-reported Expanded Disability Status Scale.

[g] Neuro-QoL: quality of life in neurological disorders.

[h] MFIS: Modified Fatigue Impact Score.

Discussion

This review sought to evaluate and summarize available literature and apps assessing remote testing for persons with MS. Though well-designed studies evaluating concordance between app testing and the neurological exams do exist, many apps operate outside the realm of currently available scientific evidence.

Comparison With Prior Work

To the authors' knowledge, this is the first scoping review with a specific focus on the use of apps for symptom monitoring and tracking clinical course in persons with MS. Previous reviews on this topic have employed a wider scope, examining all clinical trials with data pertaining to apps used in MS [6,7], while others narrowed the scope to apps used for self-assessment and rehabilitation [29] or to gait and postural control [30]. Of the 2 reviews with wider scopes, one was published in 2018 and predates all but one of the included articles [6], and the other included only 3 studies that focused on apps employing dexterity tests, accelerometers, or other sensing technologies [7].

Principal Findings

Many of the included studies demonstrated concordance between mobile testing for MS and various aspects of the neurological exam [11-13,15-23]. For example, the Adaptive Cognitive Evaluation, Elevate MS, EVO monitoring, Floodlight, MS Copilot, MS Suite, and NeuroKeys have all shown statistically significant correlations between the app and the physician's exam. The strongest correlation coefficients with standardized scales were seen with MS Copilot, when test results were pooled and compared with the MSFC [11]. However, pooled results did not have the same correlation strength with the EDSS. This may reflect the stronger similarities in the MS Copilot battery and the tests administered during the MSFC.

Although the EDSS remains an important aspect of the evaluation of persons with MS both in clinic and in the context of clinical trials, most apps seeking to correlate in-app testing with EDSS have obtained weak to moderate, albeit statistically significant, correlation coefficients [12,13,15,18,20]. The correlation coefficients were much greater with app-based e-diary scores [27]. This is notable, as the EDSS has previously been criticized for its poor assessments of upper limb and cognitive functions, which are 2 domains that are evaluated in most apps for which published data exists [31]. Additionally, the EDSS's nonlinearity may make it more difficult for testing-based apps to correctly obtain the EDSS score based on quantitative data derived from app-based testing [32].

One advantage to app-based evaluation of persons with MS is that virtual tests can be performed by persons with MS with more significant disability. One study found that some persons with MS were unable to perform the 9HPT yet were able to participate in app-based testing [16]. That said, app-based testing may be an obstacle to those with MS-related visual impairment who rely on tactile sensations to complete the required testing.

In terms of feasibility, adherence rates to the apps were lower for apps requiring daily participation for extended periods and higher for apps with less frequent testing [24-26]. This would suggest that adherence would be higher for apps that require less frequent active participation from persons with MS. Thus, striking the optimal balance between participant engagement and the adequacy of remote monitoring becomes important.

The increased frequency of app-based testing, when compared with infrequent office testing, may improve certain test results due to repeated practice. Woelfle et al [14] demonstrated improvement related to practice effect in most of the tests that comprise the Floodlight testing battery, an app that allows users to perform tests daily or weekly; however, this practice effect was not observed with the 2MWT, which evaluates walking, an activity generally performed daily by those who remain ambulatory. Similar practice effects have been described for the MSFC [33]. Clinicians who plan to use app-based testing as part of their evaluation of persons with MS should be wary of these effects when interpreting results, as they may mask deterioration or feign clinical improvement. Where applicable, a possible mitigation strategy would be to use alternating versions of tests. No studies have yet determined the optimal testing interval to avoid practice effect–related improvement.

Data on local temperature and its impact on app-based test performance have shown that increasing temperatures correlate negatively with test scores [28]. As such, apps that monitor local temperature may offer additional insight to the MS specialist who may not consider this factor when evaluating persons with MS.

Although many apps designed to track symptoms in persons with MS are publicly available on app stores, only 10 apps were identified in our scoping review as having published evidence supporting their use.

Limitations

This scoping review is limited first by the relatively small number of included articles as well as the heterogeneity of included articles. This renders drawing generalized conclusions difficult given the limited number of studies and the different comparators. As more data become available with the growth of mobile health (mHealth), future reviews may be able to compare different testing metrics with more certainty. The second limitation relates to the rapid evolution of mHealth publications and app development. This is supported by the fact that two-thirds of the included articles were published within the last 2 years. At the time of its publication, this review may not reflect the most recent data available.

Future Directions

Future app developers may wish to include both objective measures of clinical status as well as patient-reported outcomes in order to aid the neurologist in evaluating persons with MS, especially if the app is to assess the EDSS. The mobile version of the SDMT correlated well with the traditional SDMT and could be included as a measure of cognitive decline. Although current research does not describe the optimal testing interval, app testing should be used sparingly to encourage participation and reduce the practice effect. Developers may also wish to include local weather data at time of testing to allow for contextualization of at-home results.

Conclusion

The current review serves as a summary of the existing apps designed for monitoring of persons with MS and their supporting literature. Current evidence demonstrates adequate convergence of app-based testing to traditional in-person assessment. Although persons with MS will likely always require the human interaction of in-person follow-up, apps may be used as an adjunct to these visits for patients who are unable to see their neurologist on a regular basis. Although many apps with remote testing abilities are available to the public, a minority have published evidence supporting their use. Several apps had unique beneficial features; however, there was a significant amount of redundancy. Most app-based tests had a focus on physical disability and cognition. There remains a need for a comprehensive validated app that combines both patient-reported outcomes and multiple types of remote testing to better understand and care for persons with MS.

Acknowledgments

This study was supported by the Centre de formation médicale du Nouveau-Brunswick, Moncton, New Brunswick, Canada.

Data Availability

The data sets generated during or analyzed during the current study are available in Multimedia Appendix 2.

Authors' Contributions

JBM authored the original draft of this scoping review. JBM and CP independently screened studies for inclusion criteria. EH, OC, and LCW provided critical feedback and helped shape the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Detailed search strategy.
[DOCX File , 15 KB - neuro_v2i1e37944_app1.docx]

Multimedia Appendix 2
Data sets.
[XLSX File (Microsoft Excel File), 23 KB - neuro_v2i1e37944_app2.xlsx]

References

1. Dobson R, Giovannoni G. Multiple sclerosis - a review. Eur J Neurol 2019 Jan;26(1):27-40. [doi: 10.1111/ene.13819] [Medline: 30300457]
2. Block VJ, Bove R, Nourbakhsh B. The role of remote monitoring in evaluating fatigue in multiple sclerosis: a review. Front Neurol 2022 Jun 27;13:878313 [FREE Full text] [doi: 10.3389/fneur.2022.878313] [Medline: 35832181]
3. Leavitt VM, Sumowski JF, Chiaravalloti N, DeLuca J. Warmer outdoor temperature is associated with worse cognitive status in multiple sclerosis. Neurology 2012 Mar 07;78(13):964-968. [doi: 10.1212/wnl.0b013e31824d5834]
4. Tabrizi FM, Radfar M. Fatigue, sleep quality, and disability in relation to quality of life in multiple sclerosis. Int J MS Care 2015;17(6):268-274 [FREE Full text] [doi: 10.7224/1537-2073.2014-046] [Medline: 26664332]
5. Hollander JE, Carr BG. Virtually Perfect? Telemedicine for Covid-19. N Engl J Med 2020 Apr 30;382(18):1679-1681. [doi: 10.1056/NEJMp2003539] [Medline: 32160451]
6. Zayas-Garcia S, Cano-de-la-Cuerda R. [Mobile applications related to multiple sclerosis: a systematic review]. Rev Neurol 2018 Dec 16;67(12):473-483 [FREE Full text] [Medline: 30536361]
7. De Angelis M, Lavorgna L, Carotenuto A, Petruzzo M, Lanzillo R, Brescia Morra V, et al. Digital technology in clinical trials for multiple sclerosis: systematic review. J Clin Med 2021 May 26;10(11):2328 [FREE Full text] [doi: 10.3390/jcm10112328] [Medline: 34073464]
8. Stoll S, Litchman T, Wesley S, Litchman C. Multiple sclerosis apps: The dawn of a new era: A comprehensive review (P3.2-021). Neurology 2019 May;92(15 Supplement):1 [FREE Full text]
9. Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, PRISMA-P Group. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev 2015 Jan 01;4(1):1 [FREE Full text] [doi: 10.1186/2046-4053-4-1] [Medline: 25554246]
10. Methodology checklist 3: cohort studies. Scottish Intercollegiate Guidelines Network (SIGN): Healthcare Improvement Scotland. 2012. URL: http://www.sign.ac.uk/what-we-do/methodology/checklists/ [accessed 2023-01-16]
11. Maillart E, Labauge P, Cohen M, Maarouf A, Vukusic S, Donzé C, et al. MSCopilot, a new multiple sclerosis self-assessment digital solution: results of a comparative study versus standard tests. Eur J Neurol 2020 Mar;27(3):429-436 [FREE Full text] [doi: 10.1111/ene.14091] [Medline: 31538396]
12. Montalban X, Graves J, Midaglia L, Mulero P, Julian L, Baker M, et al. A smartphone sensor-based digital outcome assessment of multiple sclerosis. Mult Scler 2022 Apr 14;28(4):654-664 [FREE Full text] [doi: 10.1177/13524585211028561] [Medline: 34259588]
13. Tanoh I, Maillart E, Labauge P, Cohen M, Maarouf A, Vukusic S, et al. MSCopilot: New smartphone-based digital biomarkers correlate with Expanded Disability Status Scale scores in people with multiple sclerosis. Mult Scler Relat Disord 2021 Oct;55:103164. [doi: 10.1016/j.msard.2021.103164] [Medline: 34352512]
14. Woelfle T, Pless S, Wiencierz A, Kappos L, Naegelin Y, Lorscheider J. Practice effects of mobile tests of cognition, dexterity, and mobility on patients with multiple sclerosis: data analysis of a smartphone-based observational study. J Med Internet Res 2021 Nov 18;23(11):e30394 [FREE Full text] [doi: 10.2196/30394] [Medline: 34792480]
15. Hsu W, Rowles W, Anguera JA, Zhao C, Anderson A, Alexander A, et al. Correction: application of an adaptive, digital, game-based approach for cognitive assessment in multiple sclerosis: observational study. J Med Internet Res 2021 Jan 27;23(1):e27440 [FREE Full text] [doi: 10.2196/27440] [Medline: 33502997]

16. Boukhvalova AK, Kowalczyk E, Harris T, Kosa P, Wichman A, Sandford MA, et al. Identifying and quantifying neurological disability via smartphone. Front Neurol 2018 Sep 4;9:740 [FREE Full text] [doi: 10.3389/fneur.2018.00740] [Medline: 30233487]
17. Boukhvalova AK, Fan O, Weideman AM, Harris T, Kowalczyk E, Pham L, et al. Smartphone level test measures disability in several neurological domains for patients with multiple sclerosis. Front Neurol 2019 May 28;10:358 [FREE Full text] [doi: 10.3389/fneur.2019.00358] [Medline: 31191424]
18. Lam K, van Oirschot P, den Teuling B, Hulst H, de Jong B, Uitdehaag B, et al. Reliability, construct and concurrent validity of a smartphone-based cognition test in multiple sclerosis. Mult Scler 2022 Feb 26;28(2):300-308 [FREE Full text] [doi: 10.1177/13524585211018103] [Medline: 34037472]
19. van Oirschot P, Heerings M, Wendrich K, den Teuling B, Martens MB, Jongen PJ. Symbol Digit Modalities Test variant in a smartphone app for persons with multiple sclerosis: validation study. JMIR Mhealth Uhealth 2020 Oct 05;8(10):e18160 [FREE Full text] [doi: 10.2196/18160] [Medline: 33016886]
20. Lam K, Meijer K, Loonstra F, Coerver E, Twose J, Redeman E, et al. Real-world keystroke dynamics are a potentially valid biomarker for clinical disability in multiple sclerosis. Mult Scler 2021 Aug 05;27(9):1421-1431 [FREE Full text] [doi: 10.1177/1352458520968797] [Medline: 33150823]
21. Lam K, Twose J, McConchie H, Licitra G, Meijer K, de Ruiter L, et al. Smartphone-derived keystroke dynamics are sensitive to relevant changes in multiple sclerosis. Eur J Neurol 2022 Feb 14;29(2):522-534 [FREE Full text] [doi: 10.1111/ene.15162] [Medline: 34719076]
22. Hsu W, Rowles W, Anguera JA, Anderson A, Younger JW, Friedman S, et al. Assessing cognitive function in multiple sclerosis with digital tools: observational study. J Med Internet Res 2021 Dec 30;23(12):e25748 [FREE Full text] [doi: 10.2196/25748] [Medline: 34967751]
23. van Oirschot P, Heerings M, Wendrich K, den Teuling B, Dorssers F, van Ee R, et al. A two-minute walking test with a smartphone app for persons with multiple sclerosis: validation study. JMIR Form Res 2021 Nov 17;5(11):e29128 [FREE Full text] [doi: 10.2196/29128] [Medline: 34787581]
24. Midaglia L, Mulero P, Montalban X, Graves J, Hauser SL, Julian L, et al. Adherence and satisfaction of smartphone- and smartwatch-based remote active testing and passive monitoring in people with multiple sclerosis: nonrandomized interventional feasibility study. J Med Internet Res 2019 Aug 30;21(8):e14863 [FREE Full text] [doi: 10.2196/14863] [Medline: 31471961]
25. Bove R, White CC, Giovannoni G, Glanz B, Golubchikov V, Hujol J, et al. Evaluating more naturalistic outcome measures. Neurol Neuroimmunol Neuroinflamm 2015 Oct 15;2(6):e162. [doi: 10.1212/nxi.0000000000000162]
26. Newland P, Oliver B, Newland JM, Thomas FP. Testing feasibility of a mobile application to monitor fatigue in people with multiple sclerosis. J Neurosci Nurs 2019 Dec;51(6):331-334. [doi: 10.1097/JNN.0000000000000479] [Medline: 31688283]
27. Golan D, Sagiv S, Glass-Marmor L, Miller A. Mobile-phone-based e-diary derived patient reported outcomes: Association with clinical disease activity, psychological status and quality of life of patients with multiple sclerosis. PLoS One 2021 May 5;16(5):e0250647 [FREE Full text] [doi: 10.1371/journal.pone.0250647] [Medline: 33951061]
28. Pratap A, Grant D, Vegesna A, Tummalacherla M, Cohan S, Deshpande C, et al. Evaluating the utility of smartphone-based sensor assessments in persons with multiple sclerosis in the real-world using an app (elevateMS): observational, prospective pilot digital health study. JMIR Mhealth Uhealth 2020 Oct 27;8(10):e22108 [FREE Full text] [doi: 10.2196/22108] [Medline: 33107827]
29. Bonnechère B, Rintala A, Spooren A, Lamers I, Feys P. Is mHealth a useful tool for self-assessment and rehabilitation of people with multiple sclerosis? A systematic review. Brain Sci 2021 Sep 09;11(9):1187 [FREE Full text] [doi: 10.3390/brainsci11091187] [Medline: 34573208]
30. Abou L, Wong E, Peters J, Dossou MS, Sosnoff JJ, Rice LA. Smartphone applications to assess gait and postural control in people with multiple sclerosis: A systematic review. Mult Scler Relat Disord 2021 Jun;51:102943. [doi: 10.1016/j.msard.2021.102943] [Medline: 33873026]
31. Lamers I, Kelchtermans S, Baert I, Feys P. Upper limb assessment in multiple sclerosis: a systematic review of outcome measures and their psychometric properties. Arch Phys Med Rehabil 2014 Jun;95(6):1184-1200. [doi: 10.1016/j.apmr.2014.02.023] [Medline: 24631802]
32. Kesselring J. [Prognosis in multiple sclerosis]. Schweiz Med Wochenschr 1997 Mar 22;127(12):500-505. [Medline: 9148400]
33. Solari A, Radice D, Manneschi L, Motti L, Montanari E. The multiple sclerosis functional composite: different practice effects in the three test components. J Neurol Sci 2005 Jan 15;228(1):71-74. [doi: 10.1016/j.jns.2004.09.033] [Medline: 15607213]

Abbreviations

2MWT: 2-minute walk test
9HPT: 9-Hole Peg Test

DSST: Digit Symbol Substitution Test
EDSS: Expanded Disability Status Scale
mHealth: mobile health
MS: multiple sclerosis
MSFC: Multiple Sclerosis Functional Composite
Neuro-QoL: neurology quality of life
SDMT: Symbol Digit Modalities Test
SIGN: Scottish Intercollegiate Guidelines Network
sSDMT: smartphone SDMT
T25FW: timed 25-foot walk test

Edited by P Kubben; submitted 12.03.22; peer-reviewed by S El kefi, H Gandhi, H Mehdizadeh, M Tummalacherla, S Holm; comments to author 13.07.22; revised version received 06.09.22; accepted 30.12.22; published 06.02.23.

Please cite as:
Michaud JB, Penny C, Cull O, Hervet E, Chamard-Witkowski L
Remote Testing Apps for Multiple Sclerosis Patients: Scoping Review of Published Articles and Systematic Search and Review of Public Smartphone Apps
JMIR Neurotech 2023;2:e37944
URL: https://neuro.jmir.org/2023/1/e37944
doi:10.2196/37944
PMID:

Original Paper

The Easy and Versatile Neural Recording Platform (T-REX): Design and Development Study

Joaquín Amigó-Vega[1*], MSc; Maarten C Ottenhoff[2*], MSc; Maxime Verwoert[2], MSc; Pieter Kubben[2], MD, PhD; Christian Herff[2], PhD

[1]Computer Science Department, Gran Sasso Science Institute, L'Aquila, Italy

[2]Neurosurgery, School for Mental Health and Neuroscience, Maastricht University, Maastricht, Netherlands

[*]these authors contributed equally

Corresponding Author:
Joaquín Amigó-Vega, MSc
Computer Science Department
Gran Sasso Science Institute
Viale Francesco Crispi, 7
L'Aquila, 67100
Italy
Phone: 39 0862 4280 001
Fax: 39 0862 4280 001
Email: joaquin.amigo@gssi.it

Abstract

Background: Recording time in invasive neuroscientific research is limited and must be used as efficiently as possible. Time is often lost due to a long setup time and errors by the researcher, driven by the number of manually performed steps. Currently, recording solutions that automate experimental overhead are either custom-made by researchers or provided as a submodule in comprehensive neuroscientific toolboxes, and there are no platforms focused explicitly on recording.

Objective: Minimizing the number of manual actions may reduce error rates and experimental overhead. However, automation should avoid reducing the flexibility of the system. Therefore, we developed a software package named T-REX (Standalone Recorder of Experiments) that specifically simplifies the recording of experiments while focusing on retaining flexibility.

Methods: The proposed solution is a standalone webpage that the researcher can provide without an active internet connection. It is built using Bootstrap5 for the frontend and the Python package Flask for the backend. Only Python 3.7+ and a few dependencies are required to start the different experiments. Data synchronization is implemented using Lab Streaming Layer, an open-source networked synchronization ecosystem, enabling all major programming languages and toolboxes to be used for developing and executing the experiments. Additionally, T-REX runs on Windows, Linux, and macOS.

Results: The system reduces experimental overhead during recordings to a minimum. Multiple experiments are centralized in a simple local web interface that reduces an experiment's setup, start, and stop to a single button press. In principle, any type of experiment, regardless of the scientific field (eg, behavioral or cognitive sciences, and electrophysiology), can be executed with the platform. T-REX includes an easy-to-use interface that can be adjusted to specific recording modalities, amplifiers, and participants. Because of the automated setup, easy recording, and easy-to-use interface, participants may even start and stop experiments by themselves, thus potentially providing data without the researcher's presence.

Conclusions: We developed a new recording platform that is operating system independent, user friendly, and robust. We provide researchers with a solution that can greatly increase the time spent on recording instead of setting up (with its possible errors).

(JMIR Neurotech 2023;2:e47881) doi:10.2196/47881

KEYWORDS

recording; platform; flexible; data recording; neurotechnology; experiments

NeuroTech Dialogue

We propose a software package called T-REX (Standalone Recorder of Experiments) that is specifically designed for recording experiments. T-REX automates multiple manual actions, reducing the experimental overhead and error rate during recordings. With our system, researchers can centralize all their experiments into a simple local web interface, and set up, start, and stop experiments with a single button press. The user friendly interface can be used with different recording modalities, amplifiers, and participants, making it highly flexible. The software is executable on mainstream operating systems (Windows, Linux, and macOS) and does not require the use of a specific programming language for creating the experiments. It includes functionality to automatically record experimental data using a protocol frequently used in the community called Lab Streaming Layer. With T-REX, we simplify and streamline the recording of experiments for researchers while providing maximum flexibility in using different recording modalities, programming languages, operating systems, and amplifiers.

Introduction

Recording high-quality electrophysiological human brain activity is notoriously difficult. The best quality signal has both high spatial and temporal resolution and is recorded with invasive electrodes [1,2]. However, since implanting electrodes in humans for research purposes is a lengthy and challenging process with many safety and ethical concerns, scientists tend to use the clinical treatment of patients who receive implants for clinical purposes [3,4] as a research vehicle. Some examples are patients with medication-resistant epilepsy undergoing presurgical monitoring for resection surgery [5] or patients qualified for deep brain stimulation [6].

Because recordings should not interfere with clinical treatment, the time to record data for neuroscientific experiments in these patient groups is severely limited. For implanted epilepsy patients, the recording windows are usually a few days to 2 weeks. In contrast, for patients with deep brain stimulation, the recording windows are during surgery using microelectrode recordings, and between surgery and when the stimulator is turned on. During these recording windows, patients need time to recover and have sufficient general well-being to participate. Moreover, time spent on clinical treatment and other assessments that require recording time can further reduce the already limited recording time.

Therefore, the brief remaining time window should be used as efficiently as possible. In practice, this means that the time spent on recording should be maximized, while the time spent on setting up and solving errors should be minimized. Both the set-up time and error rate can be significantly reduced by automating as many manual actions as possible (eg, connecting to recording devices; starting experiments; selecting data streams; and starting, stopping, and synchronizing the recording). However, as experiments or recording setups change over time, it is often not worthwhile for research groups to invest in developing a more sophisticated system. It takes human resources, technical knowledge, and substantial time investment to move beyond custom-made systems, which are often only used internally and unavailable to the public. Aside from custom-made setups, there exist multiple measurement platforms, including BCI2000 [7], OpenVIBE [8], FieldTrip [9], NFBlab [10], and MEDUSA [11]. These systems can record data from many different amplifiers and include modules to design, analyze, and provide feedback during or after the experiments. While all these platforms also include good recording capabilities, they are more broadly focused on experimental design and analysis.

Additionally, these solutions limit the experiments that can be executed by the researcher in some way, either by targeting a specific type of experimental design or by imposing some hardware or software tool sets, such as programming language, input/output devices, or operating systems (OSs). Furthermore, not all platforms are open-source, which is not in the spirit of open science and impedes collective quality control and replicability. For example, FieldTrip requires the researcher to use the proprietary platform MATLAB, and BCI2000 and OpenVIBE impose the use of their tools and application programming interfaces. Additionally, the researcher must install a complete software package on the system, even when only the recording functionality is needed. Ashmaig et al [12] developed and described a system exclusively focused on continuous data recording for neurosurgical patients. The system provides a good use case for naturalistic long-term recordings but has an extensive list of hardware requirements and limits the researcher to Linux. Furthermore, not all research groups have the opportunity to perform long-term recordings.

While all these platforms provide good solutions for their use case and cover a significant part of the neural recording space, we observed that none of these platforms are specifically tailored to the setup and recording of experiments. Here, we describe the T-REX (Standalone Recorder of Experiments) platform that is specifically targeted to improve the recording of experiments. By automating the setup, start, and stop of experimental recordings, T-REX reduces the error rate and time spent between recordings. T-REX minimizes restrictions on hardware and software, is available on all major OSs, and is publicly available as an open-source project. This work presents T-REX's system design, functionality, usage, and potential implications for the field.

Methods

Requirements

We determined 3 criteria that the system should meet to make T-REX applicable to as many labs as possible. First, T-REX should be as independent as possible of tools, paradigms, OSs, and programming languages. Each lab has its preferred tool set, and ensuring independence means that researchers do not need to port their existing experiments to fit T-REX. Its only requirement is for the experiments to use Lab Streaming Layer (LSL) to stream data [13]. The backend of T-REX uses LSL to synchronize data across sources (see the section Details of LSL). Second, T-REX should be user friendly to both the researcher and the participant. Increasing simplicity will reduce error rates and the time spent on setting up, which can be achieved by automating multiple manual actions. Lastly, the system should be robust. This means that an experiment should only run when all requirements to run are met, and in case of technical problems, the experiment should retain the data up to that point and return to the *Home* screen.

System Outline

In brief, T-REX acts as the middleman handling the experimental overhead for the researcher (Figure 1). When using T-REX, the researcher can select an experiment by pressing a button on the main menu screen (Figure 2). T-REX will then check the availability of all required data streams and connect to the streams. Examples of data streams include a hand-tracking device sending coordinates of a person's hands and an amplifier recording the participant's neural activity. T-REX will then start the experiment user interface (UI) that instructs the participant on what task to perform. Upon successful start of the experiment UI, T-REX starts recording all data streams and saves them to a folder specified by the researcher. All data are saved by LSL into a single .xdf file. After the experiment is completed, the UI prompts the participant on how the experiment went and returns to the *Home* screen. During the full experiment loop, the actions that the researcher needs to perform are to start the required device data streams and select the experiment in the *Home* screen.

Figure 1. A schematic overview of the experiment loop of T-REX (Standalone Recorder of Experiments). (A) Data from the participants (eg, EEG, movement, and audio) are recorded by a variety of device inputs. Each input device should create a Lab Streaming Layer StreamOutlet to make the data available to record. (B) T-REX then provides a user interface for experiment selection. The backend finds the required data streams and records them. The rounded box shows the different software components (web interface, controller, and user configuration). (C) Example outputs of the experiment. These components interact with the participant (experiment user interface and stimuli), or the recorded data are saved. EEG: electroencephalography.

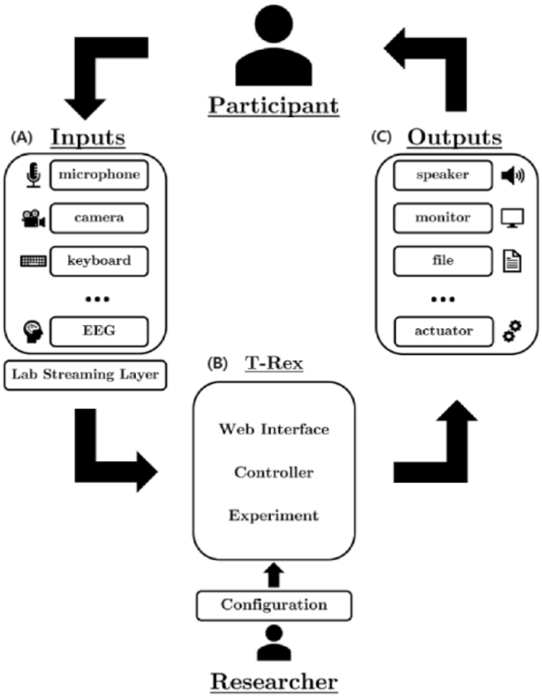
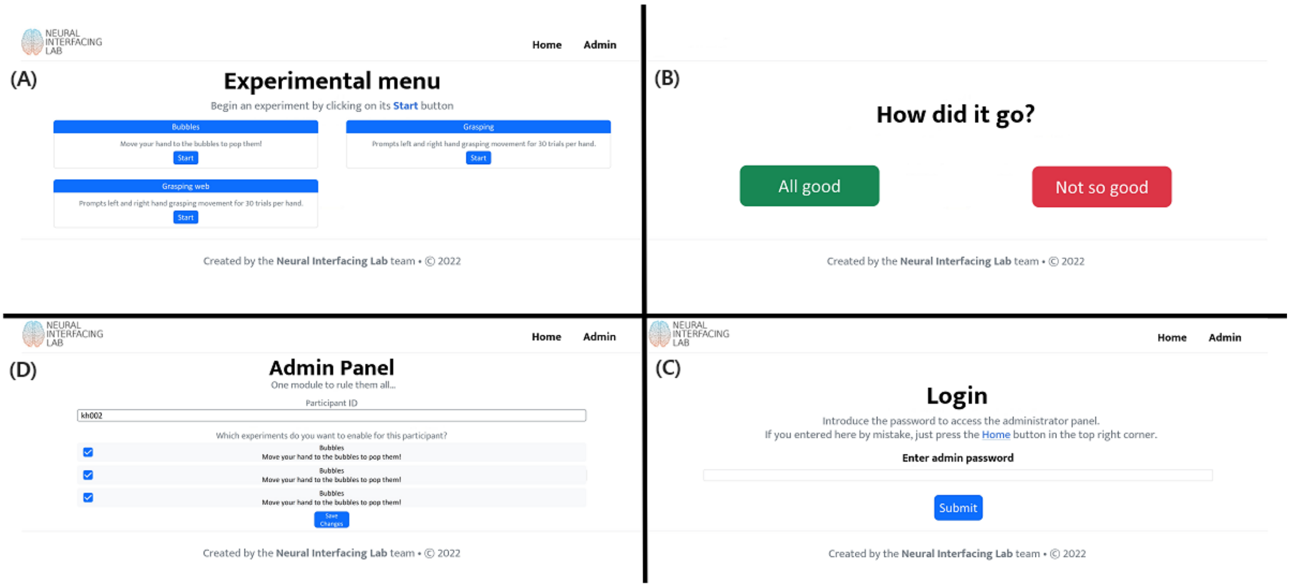


Figure 2. Representation of the main 4 windows of the web interface. (A) The Home window contains all the experiments accessible to the researcher, represented on a grid configuration. (B) The Experiment Feedback window allows obtaining feedback from the participants about their experience with the experiment. It is achieved through the green (“All good”) and red (“Not so good”) buttons. Participants can only continue after pressing one of these buttons. (C) The Admin Login window allows access to the administration panel by entering the password. (D) The Admin Configuration window allows the administrator to create new participants and modify their access to experiments.


NEURAL INTERFACING LAB
Home
Admin
(A)
Experimental menu
Begin an experiment by clicking on its Start button
Bubbles
Move your hand to the bubbles to pop them!
Start
Grasping
Prompts left and right hand grasping movement for 30 trials per hand.
Start
Grasping web
Prompts left and right hand grasping movement for 30 trials per hand.
Start
Created by the Neural Interfacing Lab team • © 2022
(B)
How did it go?
All good
Not so good
Created by the Neural Interfacing Lab team • © 2022
NEURAL INTERFACING LAB
Home
Admin
(D)
Admin Panel
One module to rule them all...
Participant ID
kh002
Which experiments do you want to enable for this participant?
☑ Bubbles
Move your hand to the bubbles to pop them!
☑ Bubbles
Move your hand to the bubbles to pop them!
☑ Bubbles
Move your hand to the bubbles to pop them!
Save Changes
Created by the Neural Interfacing Lab team • © 2022
NEURAL INTERFACING LAB
Home
Admin
(C)
Login
Introduce the password to access the administrator panel.
If you entered here by mistake, just press the Home button in the top right corner.
Enter admin password
Submit
Created by the Neural Interfacing Lab team • © 2022

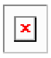
Materials, Software, and Technologies

T-REX has multiple components, including a local web interface, a recording backend, and a controller interface connecting these 2 components. The web interface (Figure 2A-D) is built using Bootstrap5 [14] for the frontend and the Python package Flask [15] for the backend. The recording backend uses LSL and handles data stream synchronization and recording itself (information is provided in the section Details of LSL). Lastly, the controller interface (information is provided in the section Controller) is implemented in Python 3.7+ and a few dependencies found in requirements.txt. T-REX is compatible with Windows, Linux, and macOS.

Details of LSL

T-REX uses LSL to synchronize the data streams from different devices, such as a variety of electroencephalography (EEG) amplifiers, audio streams, movement trackers, and cameras. The service handles “networking, time-synchronization, (near) real-time access, and optionally the centralized collection and recording of data” [13]. It is lightweight and has multilanguage and multiplatform support, including Unity and Android. LSL allows the researcher to send data via a data stream to a local network server, which can be recorded.

Basic usage involves defining a StreamOutlet that makes a time series data stream available on the network. The data are pushed per sample or per chunk into the outlet. By creating an outlet, the stream is made available to the local network of computers. The most basic usage (in Python) is represented in the following code block:



This code creates a StreamOutlet object with a name (“my_marker_stream”), type (“markers”), channel count (1), irregular sample rate (defined as 0.0), data type (“str”), and source ID (“my_unique_id”). Lastly, a sample containing “Experiment_start” is pushed to the outlet.

Inversely, to receive data, one can instantiate a StreamInlet and use inlet.pull_sample(). For a comprehensive overview, see the official documentation [13]. For T-REX to be able to record all data, the devices and the experiments themselves must all create a StreamOutlet (like the example above). If no StreamOutlet is created, T-REX will not be able to find and record the device and start the experiment. By using LSL, T-REX is able to connect to many popular experiment platforms, such as Psychopy [16], OpenSesame [17], and Presentation [18]. In case a stream is listed in the requirements provided by the config in an experiment but is not available, T-REX will throw an error and return to the *Home* screen. Thus, no experiment can start while missing a data stream.

Trigger

In some recording setups, a trigger marks the start and end of an experiment. In these setups, participants’ clinical data are recorded continuously and stored on a server. During an experiment, the data cannot be streamed directly and need to be retrieved afterward by the responsible data steward. The data steward can locate the requested data files by identifying the trigger pattern sent by the experimenter. Depending on the manufacturer, a trigger can be delivered via the amplifier or with a separate device. If it can be delivered internally, the experimenter can directly send triggers from within the experiment, and the trigger functionality of T-REX does not need to be used. T-REX provides some basic functionality to send a trigger code if an external device is required. In short, T-REX searches for a USB device with a name set in the main configuration file. It connects to this device and sets up an LSL stream. Then, if an experiment is started and the trigger flag in the main configuration file is set to True, the trigger class sends a user-defined code. When the experiment is finished, the trigger will be sent again, flagging the start and end of the complete experiment. The data steward can then retrieve the correct data with these trigger codes. At the same time as sending a trigger, the code also sends a marker to LSL, allowing for synchronization across data streams.

Software Components

The software consists of 2 main components: the *web interface* that handles the UI and the *controller* that sets up, starts, and stops all experiments (Figure 1B).

Web Interface

The web interface includes 4 windows: *Home*, *Experiment Feedback*, *Admin Login*, and *Admin Configuration* (Figure 2).

The *Home* window (Figure 2A) displays all the experiments in a grid. Experiment cards are shown on that grid with a title, description, and start button. When the button is pressed, the controller executes a command that starts the selected experiment. The command is defined by the researcher and specified on the configuration of the experiment (more details are provided in the section User Configuration). During the experiment, the web interface is on standby awaiting the completion of the experiment.

After completion, the participant is redirected to the *Experiment Feedback* window, where the question "How did the experiment go?" is prompted (Figure 2). The participant or researcher is required to select a feedback option to continue. This allows the researcher to save a brief experiment evaluation to assess data quality in later analysis. In potential future applications, the participants might perform the experiments by themselves. Then, this feedback is useful to flag the researcher to be aware of potential poor data quality. The feedback is stored under the file name feedback.txt in the same folder as the most recent .xdf file (that contains the data recorded from the experiment).

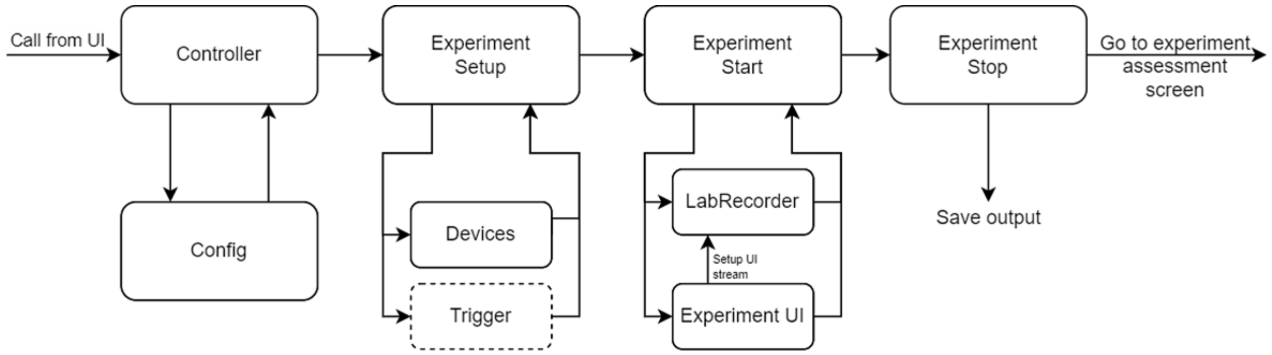
The *Admin Configuration* provides the researcher with a closed environment where the participant identifier can be selected and a selection of all available experiments is available. To access the *Admin Configuration*, the researcher must first log in using the password that is configured in the main configuration file (Figure 2C; details are provided in the section User Configuration). When logged in, the researcher can see the configuration of the active experimental session, composed of an alphanumeric participant identifier and their access to experiments. A list of all the experiments included in the platform is visible from this window, but only those with checked marks are visible to the participant. The changes in this window are only applied after pressing the "Save" button at the end of the page.

The web UI has been tested with Firefox (version 105.0.1), Chrome (version 106), Safari (version 16), and Edge (version 106), although it should be compatible with higher versions and other mainstream browsers.

Controller

The controller handles everything related to running an experiment and has 3 main parts: setup, start, and stop (Figure 3). The related code can be found in the ./libs directory.

Figure 3. Backend flow of running an experiment. When an experiment is started by pressing the start button on the card, the controller is called, loading the main configuration file and extracting the information received from the user interface (UI) about which experiment to run. Then, an experiment instance is created, loading the experiment-specific information and completing the setup in 3 steps. First, it checks for all devices and their Lab Streaming Layer streams. Second, it initializes a recorder instance and adds all streams to the list of streams it should record. Lastly, if a trigger is required for the selected experiment, it will set up a trigger class that searches and connects to the trigger. Once the subprocess call is returned, experiment sends the final trigger and stops the recorder. The data are saved in the ./output/ folder, and the researcher or participant is redirected to the experiment assessment screen (Figure 2B).



Setup

When an experiment is started by pressing the start button on the card, the controller class in *Controller.py* (Figure 3) is called, and it loads the main configuration file and extracts the information received from the UI about which experiment to run. With this information, an *experiment* instance is created, and its loading function is called.

Experiment loads the experiment-specific information and completes the setup in 3 steps. First, it checks for all devices and their LSL streams as defined by the researcher in the experiment configuration under device_inputs.

Subsequently, *experiment* initializes a *recorder* instance and adds all streams to the list of streams it should record. For a movement experiment [19-23], the streams recorded could be the neural amplifier and experimental triggers. Additionally, a movement tracker [24-26] or a force sensor [27] could be added. For speech perception [28-30] or auditory perception [31,32], the audio stream, experiment triggers, and neural data need to be recorded. For speech production [33-36], the streams could be neural data, microphone, and triggers. In the Results section, we provide some example experiments.

The last step is to check if a trigger is required for the selected experiment. If so, it will set up a trigger class that searches and connects to the trigger.

All devices must be connected and available to LSL before the *experiment* instance is called. As all requested devices are essential for successful recording, T-REX will raise an error and return to the UI if not all input devices are connected successfully.

Start

A user-defined command is called using Python's subprocess library to start the experiment UI. The command should be callable from the command line interface and can be set in the experiment-specific configuration. Because the experiment UI likely contains a stream that sends out experiment-related markers, *experiment* will start a loop on a user-defined timeout to search for the marker stream. Once found, usually almost instantly, the *recorder* will start recording all streams. Implementing the system this way does not restrict the research aside from using LSL. However, owing to the timeout, the experiment may start before the recording starts. This can only happen if the time between the setup of the experiment StreamOutlet and sending the first marker is shorter than the time that the *recorder* can find the stream and start the recording. Usually, finding the StreamOutlet and starting the recording is in the order of milliseconds. However, to entirely prevent the possibility of this happening, we recommend including a waiting screen in the experiment UI (eg, "Press button to start") or ensuring sufficient time (longer than the timeout set in the experiment configuration) between the setup of a StreamOutlet and the start of the experiment. Once connected to the experiment StreamOutlet, the experiment UI should start, and the *experiment* instance will wait until the called command is terminated and returned, which usually happens when the experiment UI window is closed.

Stop

Once the subprocess call is returned, *experiment* sends the final trigger and stops the *recorder*. The data are saved in the ./output/ folder, defined in the main configuration file (information is provided in the section User Configuration). An example of the created directory tree is provided in Multimedia Appendix 1.

Device Inputs

Each experiment can have multiple input devices, such as an amplifier measuring the neural data, a hand-tracking device, and a microphone. Any device can be included if it generates a StreamOutlet. Each device should send the data from the device to LSL, allowing it to be accessed by the other system components and to be recorded. The name, type, or source_id supplied to the StreamOutlet will be the values that T-REX will search for during experiment setup (information is provided in the section Controller). In practice, this means that either the name, type, or source_id needs to be supplied under device_inputs in the experiment configuration file (information is provided in the section Experiment Configuration). Since devices can be used for multiple experiments, we included a separate destination for all device input files (./exp_module/inputs), although input devices can be stored anywhere as long as they generate a StreamOutlet.

User Configuration

There are 2 types of configuration files that the researcher can set: main configuration and experiment-specific configuration. All configuration files are formatted in Yet Another Markup Language (YAML).

Main Configuration

The file config.yaml in the root folder contains the system-wide configuration. This configuration file contains information on general settings. Multimedia Appendix 2 provides a description of the different available options, and Multimedia Appendix 3 provides an example of the main configuration file. The main option under *path* is the path that all relative paths will be anchored to and should be set to the root folder. Most parameters are preset, but out and trigger configurations may vary between different recording setups and might need to be redefined.

Experiment Configuration

Each experiment included in T-REX requires a separate folder in ./exp_module/experiments/ and must include at least 2 files: config.yaml and the file to start the experiment. A full description of all the fields and different options in config.yaml can be found in Multimedia Appendix 4. The *name* and *description* define the text shown in the UI; *command* sets the command line interface command made by the controller class to start the experiment; and *exp_outlet* sets the name, type, or source_id that the experiment class will search for. For example, if the experiment UI is a Python script that will create a StreamOutlet named markers, the *command* to execute would be python .\exp_module\experiments\your_experiment_file.py and *exp_outlet*='markers'.

Results

Overview

We have included 3 different example experiments to provide a practical view of how to use T-REX. The examples can also serve as a quick start for researchers to create new experiments or adapt the ones included. A step-by-step explanation of adding a new experiment is described in the section Adding New Experiments to the Platform.

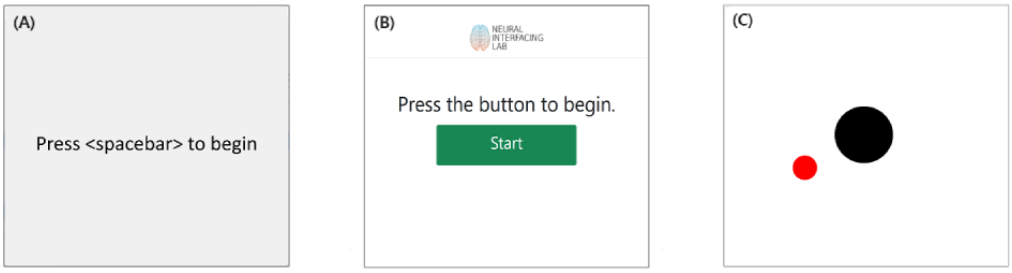
Case 1: Simple Experiment in Python

This experiment is a simple text-based instruction for a grasping task (Figure 4A). The participant is prompted by text in a Python Tkinter [37] window to continuously open and close either the left or right hand, as used previously [38]. The experiment requires neural data as the input device and generates a StreamOutlet to send markers that inform about the start and end of the experiment and of the trials. The neural data are acquired from a stream with *name=Micromed*, *type=EEG*, and *source_id=micm*01. These values are all set by the researcher. As T-REX will search for all 3 options (name, type, and source_id), only 1 must be provided. Therefore, the option under device_inputs in grasping\config.yaml is set to eeg (case insensitive). Next, the *marker* StreamOutlet that will be generated by the experiment has *source_id=emuidw*22. When the *experiment* class runs the experiment command (command field in grasping\config.yaml), it will search for these streams. Therefore, the exp_outlet field is set to 'emuidw22'. Finally, since the grasping experiment is Python-based, the command should use Python to call the script with the command: python .\exp_module\experiments\grasping\grasping.py. The configuration file used has been presented in Multimedia Appendix 5.

When these options are set, the experiment is ready to go and can be started by pressing the start button on the *Home* window. The Tkinter window opens and waits for the spacebar to be pressed. Once pressed, the experiment starts and is locked as the top viewed window until completion. When the experiment is finished and closed (ie, the command call ends and returns to the *experiment* class), the *experiment* instance stops the recording and saves the data. In-depth details on how experiments are started and stopped are described in the section Controller.

Figure 5 shows a random selection of 15 channels of neural data recorded with T-REX during the grasping experiment. Two streams were used in this experiment. First a *marker* StreamOutlet that sends all experiment-related markers, such as the start and end of the experiments and the start and end of each trial, with the accompanying label (move or rest). Second, an EEG StreamOutlet that streams the data from our Micromed Amplifier to LSL. With T-REX, these streams were automatically identified and recorded. The start and end of the colored columns (identifying move and rest trials) were determined by the recorded markers sent through the marker StreamOutlet. The synchronization by LSL ensures that the EEG and *marker* stream timestamps are the same.

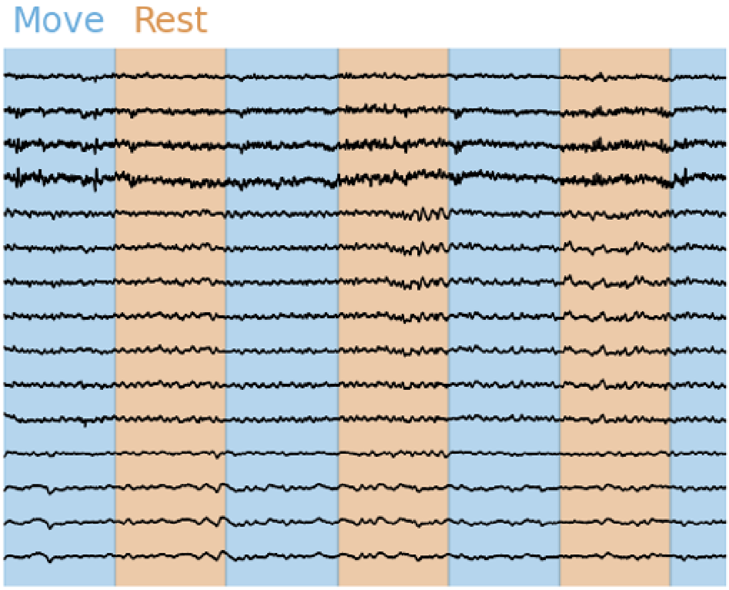
Figure 4. User interfaces for the 3 use case experiments included. (A) Grasping: simple text-based experiment built using the Python package Tkinter. (B) Grasping web experiment: reimplementation of the grasping experiment as a single page application (SPA) to allow its execution on any device with access to a web browser. (C) 3D hand-tracking experiment: the hand-tracking is performed using the LeapMotion controller, and the experiment is implemented in Python using the package Tkinter.



Press <spacebar> to begin

Press the button to begin.

Figure 5. Neural data were recorded with the grasping experiment using T-REX (Standalone Recorder of Experiments). Two streams were recorded during this experiment: an EEG stream and a marker stream. The data from the EEG stream are shown by the black lines, indicating the voltage over time in a selection of 15 neural electrodes. The marker stream sends the start and end of the experiment and the individual trials. These markers were used to determine the colored areas (blue and orange) shown. EEG: electroencephalography.



Case 2: Simple Experiment in a Web UI

We included the same grasping experiment as in Case 1 but implemented it in a web interface (Figure 4B). It uses a single page application (SPA) locally and thus can be created on any device with access to a web browser, like a laptop, tablet, and smartphone. The grasping web experiment also illustrates options other than a Tkinter window for experimenting. No internet connection is required, relieving some security concerns that could render execution on the web unsafe.

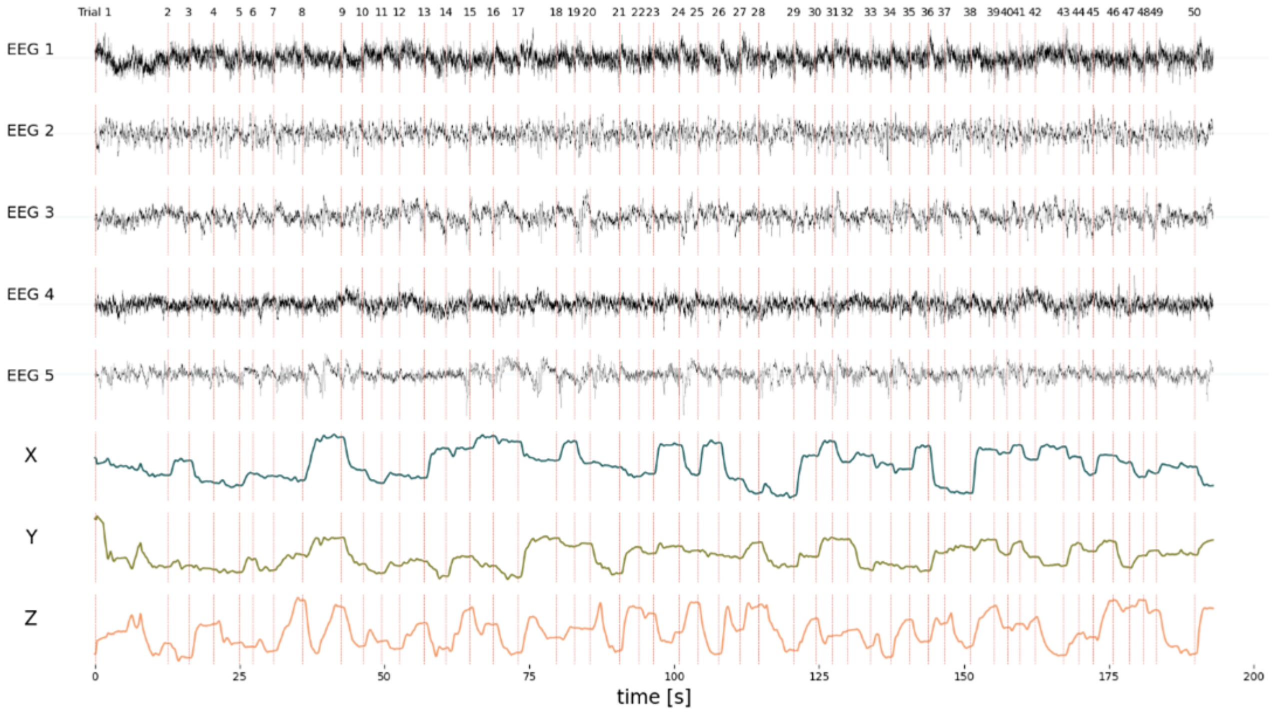
We constructed the experiment using HTML, CSS (Bootstrap5 for responsiveness and other visual aspects), and JavaScript for behavior. The device input is the same as in the Tkinter implementation of the experiment and the StreamOutlet containing the markers; thus, the device_inputs and exp_outlet are the same. The difference is in the command executed to start the experiment. In this case, start .\exp_module\experiments\graspingWeb\index.html is used. The configuration file used has been presented in Multimedia Appendix 6.

Once the experiment is started on the *Home* window, the *experiment* instance opens another tab on the browser displaying the “grasping_web” experiment. The experiment starts when the participant presses the green “Start” button. When the experiment is finished, the participant or researcher is prompted to press a red button to close the experiment. The GraspingWeb command call is finished at the button press and returns to the *experiment* instance, stopping the recording and saving the data.

Case 3: Multiple Devices

Lastly, we included a 3D hand-tracking experiment, where the goal is to hold a cursor (a black circle) on a target (a red circle). The cursor can be moved in 3 dimensions, where the third dimension controls the size of the circle (Figure 4C). In this case, the hand tracking is performed by the LeapMotion controller [39], but any other device can be used. We have provided a .exe file that reads the data from the tracker and sends it to an LSL StreamOutlet with *name*=*LeapLSL*, *type*=*Coordinates*, and *source_id*=*LEAPLSL*01. In addition to the hand-tracking information, we also need neural activity, for which we use the same StreamOutlet as described in Case 2. Lastly, the experiment is implemented in a Python Tkinter window and generates a marker stream similar to the stream described in the previous use case with *Source_id*=*BUBBLE*01. Thus, to set up the configuration for this experiment, we set the command to python .\exp_module\experiments\Bubbles\bubbles.py, exp_outlet to BUBBLE01, and device_inputs to LEAPLSL01 (the tracking information stream) and eeg (the neural data stream). To run the experiment, the researcher should start the device stream before the experiment is started in the *Home* screen (ie, run the .exe first). The configuration file used has been provided in Multimedia Appendix 7. An example of data recorded with T-REX for this experiment can be appreciated in Figure 6.

Figure 6. The combined data recorded from 3 different streams: an EEG stream, a marker stream, and a LeapMotion controller. The EEG channels are 5 channels randomly selected from 87 available channels. X, Y, and Z are the 3D coordinates of the palm of the hand, provided by a LeapMotion controller. The marker stream provides the shown trials (numbers on top with vertical dashed lines). To start and record this experiment, the LeapLSL stream has to be started, along with the EEG stream. Then, only the experiment needs to be started in T-REX (Standalone Recorder of Experiments). T-REX records all 3 streams (synchronized by Lab Streaming Layer), ultimately allowing to combine the 3 streams into this image. EEG: electroencephalography.



Mix and Match

We have presented only 3 examples showing different possibilities. Different devices can be included by adding a StreamOutlet name, type, or source_id to the list of device_outputs. The only requirement to add a device is that the data from the device can be sent to a LabStreamingLayer StreamOutlet. This code is either supplied by the manufacturer or written by the researcher. If this requirement is met, any medical device or technology can be included, as T-REX does not impose any further restrictions on technologies or types of experiments, including, but not limited to, speech production, audio or speech perception, movement, decision-making, and simple or naturalistic tasks [40,41]. For example, new experiments can also be built in Unity [42] or PyGame [43] to provide better graphical experiences.

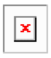
Adding New Experiments to the Platform

The following steps describe how to add a new experiment from scratch to T-REX:

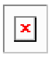
1. Create the experiment folder inside the directory ./exp_module/experiments/. An example of the directory tree for different example experiments can be found in Multimedia Appendix 8.
2. Create the experiment configuration file (config.yaml) inside the new folder. Information in Multimedia Appendix 9 can be used as the base example for creating this file, and the section Experiment Configuration contains a detailed description of each parameter.
3. Adjust the fields to the specific experiment.

After completing these initial steps, the experiment should be visible from the *Admin Configuration* panel. The researcher can set the experiment as “visible” from the admin panel by selecting its corresponding check mark. If configured as “visible,” it should appear on the *Home* window, and it can be executed by clicking on its respective button.

It is worth mentioning that when porting an already configured version of T-REX to a different OS, some parameters might need to be revised. For example, regarding the parameter command, when used on Windows to start a Python experiment, the definition is as follows:



However, when used on Unix or Unix-like systems, the definition changes to the following:



The difference comes because “/” is the path separator on Unix and Unix-like systems, and Microsoft uses “\”.

There might be other scenarios where the parameter command might differ between OSs; thus, we recommend revising each experiment configuration file when porting the platform to a different OS.

Practical Experience

At the time of writing, we entirely switched to recording with T-REX for our experiments at different recording sites. So far, we have recorded multiple experiments, involving speech, motor, and decision-making tasks. Furthermore, at one of the recording sites, we recorded using the trigger functionality included in T-REX. We see no indications of different data quality in our neural decoding endeavors. We can decode speech [44,45] and movement trajectories [46] with performance equal to that using our previous setup.

Discussion

We presented T-REX, an independent, user friendly, and robust system that minimizes the setup time and error rate. T-REX provides a simple UI and reduces the experimental setup to the press of a button. The software merges the LSL recording backend with a simple UI, automating experimental overhead for the researcher. T-REX reduces the setup time and error rate, resulting in more time spent recording neural data.

The simplicity of T-REX reduces the number of actions that the researcher must perform to only 2: starting the required devices and starting the experiment. The fewer manual actions the researcher needs to perform, the lower the chance that an error is made. It improves reliability and increases total data volume and time spent on recording. The LSL software package fully handles synchronization and recording. We decided on LSL as it is lightweight, is easy to use, has submillisecond timekeeping, and has a proven track record [47]. The flexibility of T-REX makes the system applicable in fields other than the neuroscientific context described here.

T-REX provides benefits for both the researcher and participant. A streamlined process may have multiple benefits from the perspective of the participant. It leaves more time to interact with the participant, making it more comforting and engaging. T-REX may be particularly beneficial for participants who are anxious or nervous about participating. Furthermore, a streamlined process conveys more professionalism and may improve participation satisfaction, ultimately increasing the willingness to participate in future research. Moreover, if the start and recording of experiments are simplified enough, participants may be able to run experiments themselves. The introduction of engaging and fun experiments that enable participants to run them as they like provides the participants with an opportunity to alleviate boredom and do something meaningful by contributing to scientific research. Together, both the researcher (more data) and the participant (more engagement) are benefitted. While T-REX has been developed with independent recording in mind, it is currently not being tested for that purpose.

In comparison with other available software platforms, T-REX is the only solution specifically focused on recording experiments, allowing it to remain lightweight. Platforms like BCI2000 [7], OpenViBE [8], and MEDUSA [11] offer comprehensive functionalities spanning the 3 stages of a BCI system: signal acquisition, signal processing, and feedback presentation. However, they require complete software installation even if only the recording module is needed. T-REX enhances the researcher experience by offering flexibility in the choice of programming language and technology for creating the experiments, unlike BCI2000 and OpenViBE, which mandate the use of C++; MEDUSA, which requires the use of Python; and NFBlab [10], which requires the use of its graphical UI. Regarding compatibility, T-REX holds a distinct advantage, supporting all major OSs, including Windows, Linux, and macOS. This is in contrast with BCI2000’s limited functionality outside Windows and MEDUSA’s exclusive Windows availability, as well as the system presented by Ashmaig et al

[12], which is Linux-bound. Each of these platforms has its strengths and excels in its intended function. T-REX provides a tailored solution for a specific part of neuroscientific research that allows it to remain simple and lightweight.

T-REX aims for simplicity, and setting up experiments in T-REX requires basic knowledge of command line interface usage. Moreover, experiments and devices must use LSL to make data available. Although LSL is available for all mainstream OSs and programming languages, experiments already used by researchers may require adjustments to the experiment code structure for inclusion in T-REX. Therefore, technical knowledge and usage of LSL may limit the applicability for some labs. Furthermore, T-REX is available for all mainstream OSs but may not apply to all different versions. Specifically, the command line interface version of LabRecorder, including the script that records and stores the multiple data streams, had to be built for different chipsets (M1 and M2) for macOS. These are currently included, but other architectures likely require a different build of LabRecorder. As T-REX matures, we expect more versions to become applicable.

T-REX is in ongoing development, and we have identified several potential future updates targeting an improved user experience. Device streams currently need to be started manually, and this may be performed automatically at the start of an experiment. This is also a requirement to enable participants to start recordings themselves, which is a main future improvement. Aside from ensuring that there are no manual actions except starting the experiments, allowing T-REX for independent use may require improved internal logging and error handling. Combined, these updates would reduce even more actions for both the researcher and participant, and increase the robustness of T-REX.

In conclusion, T-REX offers a flexible solution to record neuroscientific experiments. It streamlines setup and recording, and reduces error rates that increase the time spent on recordings. We envision T-REX to help standardize and simplify recording experiments and eventually allow recordings by participants independently. This may improve the overall satisfaction of participation and increase the amount of data collected. The open-source nature of T-REX is in the spirit of open science and increases its value through an increase in community knowledge.

Acknowledgments

CH acknowledges funding from the Dutch Research Council (Nederlandse Organisatie voor Wetenschappelijk Onderzoek) through the research project “Decoding Speech In SEEG (DESIS)” with project number VI.Veni.194.021.

Data Availability

The source code, installation guide, and example experiments can be found on GitHub [48]. T-REX is available under the permissive MIT License. As T-REX will be in ongoing development, we kindly invite researchers to provide feedback or contribute to this open-source project.

Authors' Contributions

JAV, MCO, MV, PK, and CH conceptualized the study. JAV, MCO, and PK performed the investigation. JAV, MCO, PK, and CH participated in the methodology. JAV and MCO contributed to project administration. PK and CH managed the resources. JAV, MCO, and PK contributed to the software. PK and CH supervised the study. JAV, MCO, and MV contributed to validation. JAV and MCO contributed to visualization. JAV and MCO wrote the original draft. JAV, MCO, MV, PK, and CH reviewed and edited the manuscript.

Conflicts of Interest

The author PK is the Editor-in-Chief of JMIR Neurotechnology. PK was not involved in any decisions made regarding this manuscript. All other authors declare no conflicts of interest.

Multimedia Appendix 1

The directory tree illustrates the content of the ./output/ folder when saving the experimental data gathered with one experiment. The output.xdf file is created upon experiment completion. It contains the recorded data from the preconfigured Lab Streaming Layer streams. The feedback.txt file contains the feedback the participant inputted on the Experiment Feedback window, and it is saved in the same folder as the most recent .xdf file.
[PNG File , 36 KB - neuro_v2i1e47881_app1.png]

Multimedia Appendix 2

The system-wide configuration file that must be placed inside the root folder of the project, which allows the researcher to configure the execution of T-REX.
[PNG File , 154 KB - neuro_v2i1e47881_app2.png]

Multimedia Appendix 3
Example of the main configuration file. Note that all paths are relative to the main parameter.
[PNG File , 63 KB - neuro_v2i1e47881_app3.png]

Multimedia Appendix 4
The different options for the experiment configuration file. Each experiment must include this file. The parameter command might need to be modified when porting the platform to a different operating system (from Windows to Linux or macOS, for example). It is up to the researcher to perform the redefinition.
[PNG File , 211 KB - neuro_v2i1e47881_app4.png]

Multimedia Appendix 5
Experiment configuration file used for the grasping experiment. This experiment presents simple instructions to the participant indicating continuous opening and closing of either the left or right hand. The visual interface was built using the Python Tkinter library.
[PNG File , 79 KB - neuro_v2i1e47881_app5.png]

Multimedia Appendix 6
Experiment configuration file used for the grasping web experiment. This experiment presents simple instructions to the participant indicating continuous opening and closing of either the left or right hand. The visual interface was built using HTML, CSS (Bootstrap5 for responsiveness and other visual aspects), and JavaScript for behavior.
[PNG File , 89 KB - neuro_v2i1e47881_app6.png]

Multimedia Appendix 7
Experiment configuration file used for the 3D hand-tracking experiment. The goal of the experiment is to hold the cursor on the target. The cursor can be moved in 3 dimensions, where the third dimension controls the size of the circle. In this case, the hand tracking is done by the LeapMotion controller.
[PNG File , 72 KB - neuro_v2i1e47881_app7.png]

Multimedia Appendix 8
The directory tree illustrates a system with 3 different folders, each for a different experiment (~/EXPERIMENT_1/, ~/EXPERIMENT_2/, and ~/EXPERIMENT_3/). Each experiment contains its own configuration file (config.yaml). The researcher can add any additional files to each folder.
[PNG File , 54 KB - neuro_v2i1e47881_app8.png]

Multimedia Appendix 9
Template that can be used for creating an experiment configuration file.
[PNG File , 138 KB - neuro_v2i1e47881_app9.png]

References

1. Herff C, Krusienski D, Kubben P. The potential of stereotactic-EEG for brain-computer interfaces: Current progress and future directions. Front Neurosci 2020;14:123 [FREE Full text] [doi: 10.3389/fnins.2020.00123] [Medline: 32174810]
2. Jacobs J, Kahana MJ. Direct brain recordings fuel advances in cognitive electrophysiology. Trends Cogn Sci 2010 Apr;14(4):162-171 [FREE Full text] [doi: 10.1016/j.tics.2010.01.005] [Medline: 20189441]
3. Feinsinger A, Pouratian N, Ebadi H, Adolphs R, Andersen R, Beauchamp MS, NIH Research Opportunities in Humans Consortium. Ethical commitments, principles, and practices guiding intracranial neuroscientific research in humans. Neuron 2022 Jan 19;110(2):188-194 [FREE Full text] [doi: 10.1016/j.neuron.2021.11.011] [Medline: 35051364]
4. Mercier MR, Dubarry A, Tadel F, Avanzini P, Axmacher N, Cellier D, et al. Advances in human intracranial electroencephalography research, guidelines and good practices. Neuroimage 2022 Oct 15;260:119438 [FREE Full text] [doi: 10.1016/j.neuroimage.2022.119438] [Medline: 35792291]
5. Chauvel P, Gonzalez-Martinez J, Bulacio J. Chapter 3 - Presurgical intracranial investigations in epilepsy surgery. In: Levin KH, Chauvel P, editors. Handbook of Clinical Neurology. Amsterdam, Netherlands: Elsevier; 2019:45-71.
6. Lozano AM, Lipsman N, Bergman H, Brown P, Chabardes S, Chang JW, et al. Deep brain stimulation: current challenges and future directions. Nat Rev Neurol 2019 Mar 25;15(3):148-160 [FREE Full text] [doi: 10.1038/s41582-018-0128-2] [Medline: 30683913]
7. Schalk G, McFarland D, Hinterberger T, Birbaumer N, Wolpaw J. BCI2000: A general-purpose brain-computer interface (BCI) system. IEEE Trans. Biomed. Eng 2004 Jun;51(6):1034-1043. [doi: 10.1109/tbme.2004.827072]

8. Renard Y, Lotte F, Gibert G, Congedo M, Maby E, Delannoy V, et al. OpenViBE: An open-source software platform to design, test, and use brain–computer interfaces in real and virtual environments. Presence: Teleoperators and Virtual Environments 2010 Feb 01;19(1):35-53. [doi: 10.1162/pres.19.1.35]
9. Oostenveld R, Fries P, Maris E, Schoffelen J. FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput Intell Neurosci 2011;2011:156869 [FREE Full text] [doi: 10.1155/2011/156869] [Medline: 21253357]
10. Smetanin N, Volkova K, Zabodaev S, Lebedev M, Ossadtchi A. NFBLab-A versatile software for neurofeedback and brain-computer interface research. Front Neuroinform 2018;12:100 [FREE Full text] [doi: 10.3389/fninf.2018.00100] [Medline: 30618704]
11. Santamaría-Vázquez E, Martínez-Cagigal V, Marcos-Martínez D, Rodríguez-González V, Pérez-Velasco S, Moreno-Calderón S, et al. MEDUSA©: A novel Python-based software ecosystem to accelerate brain-computer interface and cognitive neuroscience research. Comput Methods Programs Biomed 2023 Mar;230:107357 [FREE Full text] [doi: 10.1016/j.cmpb.2023.107357] [Medline: 36693292]
12. Ashmaig O, Hamilton L, Modur P, Buchanan R, Preston A, Watrous A. A platform for cognitive monitoring of neurosurgical patients during hospitalization. Front Hum Neurosci 2021;15:726998 [FREE Full text] [doi: 10.3389/fnhum.2021.726998] [Medline: 34880738]
13. Swartz Center for Computational Neuroscience: Lab Streaming Layer. GitHub, Inc. URL: https://github.com/sccn/labstreaminglayer [accessed 2023-09-21]
14. Getting started. Bootstrap. URL: https://getbootstrap.com/docs/5.1/getting-started/introduction/ [accessed 2023-09-21]
15. Flask. Pallets. URL: https://flask.palletsprojects.com/en/2.1.x [accessed 2023-09-21]
16. PsychoPy. URL: https://www.psychopy.org/ [accessed 2023-09-21]
17. OpenSesame. URL: https://osdoc.cogsci.nl [accessed 2022-09-26]
18. Neurobehavioral Systems. URL: https://www.neurobs.com/ [accessed 2022-10-18]
19. Ottenhoff M, Goulis S, Wagner L, Tousseyn S, Colon A, Kubben P. Continuously Decoding Grasping Movements using Stereotactic Depth Electrodes. 2021 Presented at: 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); November 01-05, 2021; Mexico. [doi: 10.1109/EMBC46164.2021.9629639]
20. Ottenhoff M, Verwoert M, Goulis S, Colon A, Wagner L, Tousseyn S, et al. Executed and imagined grasping movements can be decoded from lower dimensional representation of distributed non-motor brain areas. BioRxiv. URL: https://www.biorxiv.org/content/10.1101/2022.07.04.498676v1 [accessed 2023-09-21]
21. Li G, Jiang S, Meng J, Chai G, Wu Z, Fan Z, et al. Assessing differential representation of hand movements in multiple domains using stereo-electroencephalographic recordings. Neuroimage 2022 Apr 15;250:118969 [FREE Full text] [doi: 10.1016/j.neuroimage.2022.118969] [Medline: 35124225]
22. Li G, Jiang S, Paraskevopoulou SE, Chai G, Wei Z, Liu S, et al. Detection of human white matter activation and evaluation of its function in movement decoding using stereo-electroencephalography (SEEG). J Neural Eng 2021 Aug 12;18(4):0460c6. [doi: 10.1088/1741-2552/ac160e] [Medline: 34284361]
23. Merk T, Peterson V, Lipski W, Blankertz B, Turner R, Li N, et al. Electrocorticography is superior to subthalamic local field potentials for movement decoding in Parkinson's disease. Elife 2022 May 27;11:11 [FREE Full text] [doi: 10.7554/eLife.75126] [Medline: 35621994]
24. Mondini V, Kobler RJ, Sburlea AI, Müller-Putz G. Continuous low-frequency EEG decoding of arm movement for closed-loop, natural control of a robotic arm. J Neural Eng 2020 Aug 11;17(4):046031. [doi: 10.1088/1741-2552/aba6f7] [Medline: 32679573]
25. Coste CA, William L, Fonseca L, Hiairrassary A, Andreu D, Geffrier A, et al. Activating effective functional hand movements in individuals with complete tetraplegia through neural stimulation. Sci Rep 2022 Oct 06;12(1):16189 [FREE Full text] [doi: 10.1038/s41598-022-19906-x] [Medline: 36202865]
26. Hosseini S, Shalchyan V. Continuous decoding of hand movement from EEG signals using phase-based connectivity features. Front Hum Neurosci 2022;16:901285 [FREE Full text] [doi: 10.3389/fnhum.2022.901285] [Medline: 35845243]
27. Shah S, Tan H, Brown P. Continuous force decoding from deep brain local field potentials for Brain Computer Interfacing. 2017 Presented at: 8th International IEEE/EMBS Conference on Neural Engineering (NER); May 25-28, 2017; Shanghai, China. [doi: 10.1109/NER.2017.8008367]
28. Patel P, van der Heijden K, Bickel S, Herrero JL, Mehta AD, Mesgarani N. Interaction of bottom-up and top-down neural mechanisms in spatial multi-talker speech perception. Curr Biol 2022 Sep 26;32(18):3971-3986.e4. [doi: 10.1016/j.cub.2022.07.047] [Medline: 35973430]
29. Prinsloo K, Lalor E. General auditory and speech-specific contributions to cortical envelope tracking revealed using auditory chimeras. J Neurosci 2022 Oct 12;42(41):7782-7798 [FREE Full text] [doi: 10.1523/JNEUROSCI.2735-20.2022] [Medline: 36041853]
30. Biau E, Schultz BG, Gunter TC, Kotz SA. Left motor δ oscillations reflect asynchrony detection in multisensory speech perception. J. Neurosci 2022 Jan 27;42(11):2313-2326. [doi: 10.1523/jneurosci.2965-20.2022]

31. Hausfeld L, Disbergen NR, Valente G, Zatorre RJ, Formisano E. Modulating cortical instrument representations during auditory stream segregation and integration with polyphonic music. Front Neurosci 2021 Sep 24;15:635937 [FREE Full text] [doi: 10.3389/fnins.2021.635937] [Medline: 34630007]
32. Hausfeld L, Shiell M, Formisano E, Riecke L. Cortical processing of distracting speech in noisy auditory scenes depends on perceptual demand. Neuroimage 2021 Mar;228:117670 [FREE Full text] [doi: 10.1016/j.neuroimage.2020.117670] [Medline: 33359352]
33. Angrick M, Herff C, Mugler E, Tate MC, Slutzky MW, Krusienski DJ, et al. Speech synthesis from ECoG using densely connected 3D convolutional neural networks. J Neural Eng 2019 Jun 16;16(3):036019 [FREE Full text] [doi: 10.1088/1741-2552/ab0c59] [Medline: 30831567]
34. Herff C, Diener L, Angrick M, Mugler E, Tate MC, Goldrick MA, et al. Generating natural, intelligible speech from brain activity in motor, premotor, and inferior frontal cortices. Front Neurosci 2019 Nov 22;13:1267 [FREE Full text] [doi: 10.3389/fnins.2019.01267] [Medline: 31824257]
35. Angrick M, Ottenhoff MC, Diener L, Ivucic D, Ivucic G, Goulis S, et al. Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity. Commun Biol 2021 Sep 23;4(1):1055 [FREE Full text] [doi: 10.1038/s42003-021-02578-0] [Medline: 34556793]
36. Moses DA, Metzger SL, Liu JR, Anumanchipalli GK, Makin JG, Sun PF, et al. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. N Engl J Med 2021 Jul 15;385(3):217-227 [FREE Full text] [doi: 10.1056/NEJMoa2027540] [Medline: 34260835]
37. tkinter. Python. URL: https://docs.python.org/3/library/tkinter.html [accessed 2023-09-21]
38. Ottenhoff M, Verwoert M, Goulis S, Wagner L, van Dijk J, Kubben P, et al. Global motor dynamics - Invariant neural representations of motor behavior in distributed brain-wide recordings. bioRxiv. URL: https://www.biorxiv.org/content/10.1101/2023.07.07.548122v1 [accessed 2023-09-21]
39. Leap Motion Controller. Ultraleap. URL: https://www.ultraleap.com/product/leap-motion-controller [accessed 2023-09-21]
40. Sonkusare S, Breakspear M, Guo C. Naturalistic stimuli in neuroscience: Critically acclaimed. Trends Cogn Sci 2019 Aug;23(8):699-714. [doi: 10.1016/j.tics.2019.05.004] [Medline: 31257145]
41. Hamilton LS, Huth AG. The revolution will not be controlled: natural stimuli in speech neuroscience. Lang Cogn Neurosci 2020 Jul 22;35(5):573-582 [FREE Full text] [doi: 10.1080/23273798.2018.1499946] [Medline: 32656294]
42. Unity. URL: https://unity.com [accessed 2023-09-21]
43. Pygame. URL: https://www.pygame.org/wiki/about [accessed 2023-09-21]
44. Amigó-Vega J, Verwoert M, Ottenhoff M, Kubben P, Herff C. Decoding articulatory trajectories during speech production from intracranial EEG. In: Proceedings of the 10th International Brain-Computer Interface Meeting. 2023 Presented at: 10th International Brain-Computer Interface Meeting; June 6-9, 2023; Brussels, Belgium p. Article ID: 144441.
45. Verwoert M, Ottenhoff M, Amigó-Vega J, Goulis S, Wagner L, Kubben P, et al. Evaluating implant locations for a minimally invasive speech BCI. In: Proceedings of the 10th International Brain-Computer Interface Meeting. 2023 Presented at: 10th International Brain-Computer Interface Meeting; June 6-9, 2023; Brussels, Belgium p. Article ID: 144185.
46. Ottenhoff M, Verwoert M, Goulis S, Colon A, Kubben P, Shanechi M, et al. Decoding hand kinematics from brain-wide distributed neural recordings. In: Proceedings of the 10th International Brain-Computer Interface Meeting. 2023 Presented at: 10th International Brain-Computer Interface Meeting; June 6-9, 2023; Brussels, Belgium.
47. Wang Q, Zhang Q, Sun W, Boulay C, Kim K, Barmaki RL. A scoping review of the use of lab streaming layer framework in virtual and augmented reality research. Virtual Reality 2023 May 02;27(3):2195-2210. [doi: 10.1007/S10055-023-00799-8]
48. T-Rex source code and documentation. GitHub, Inc. URL: https://github.com/neuralinterfacinglab/t-rex [accessed 2023-09-21]

Abbreviations

EEG: electroencephalography
LSL: Lab Streaming Layer
OS: operating system
T-REX: Standalone Recorder of Experiments
UI: user interface
YAML: Yet Another Markup Language

Edited by T de Azevedo Cardoso; submitted 05.04.23; peer-reviewed by S Zhao, J Geng, S Hacking; comments to author 13.07.23; revised version received 17.08.23; accepted 07.09.23; published 24.10.23.

Please cite as:
Amigó-Vega J, Ottenhoff MC, Verwoert M, Kubben P, Herff C
The Easy and Versatile Neural Recording Platform (T-REX): Design and Development Study
JMIR Neurotech 2023;2:e47881
URL: https://neuro.jmir.org/2023/1/e47881
doi:10.2196/47881
PMID:

Original Paper

Clinical Perspectives on Using Remote Measurement Technology in Assessing Epilepsy, Multiple Sclerosis, and Depression: Delphi Study

Jacob A Andrews[1,2], BA, MA, PhD; Michael P Craven[1,3], PhD; Boliang Guo[4], PhD; Janice Weyer[5], BA, BSc; Simon Lees[5]; Spyridon I Zormpas[5], MSc; Sarah E Thorpe[5], BSc; Julie Devonshire[5]; Victoria San Antonio-Arce[6], MD, PhD; William P Whitehouse[7], BSc, MBBS; Jessica Julie[8], BSc; Sam Malins[9], PhD; Alexander Hammers[10], MD, PhD; Andreas Reif[11], MD; Henricus G Ruhe[12,13], MD, PhD; Federico Durbano[14], MD; Stefano Barlati[15], MD; Arjune Sen[16], BM BCh, MA, PhD; Jette L Frederiksen[17], MD; Alessandra Martinelli[18], MD, PhD; Antonio Callen[19], MD; Joan Torras-Borrell[20], MD; Nuria Berrocal-Izquierdo[21], MD, PhD; Ana Zabalza[22,23], MD, PhD; Richard Morriss[1,2], MD, PhD; Chris Hollis[1,2], BSc, MBBS, DCH, PhD; The RADAR-CNS Consortium[24]

[1]National Institute for Health and Care Research MindTech MedTech Co-operative, University of Nottingham, Nottingham, United Kingdom

[2]Academic Unit of Mental Health and Clinical Neurosciences, School of Medicine, University of Nottingham, Nottingham, United Kingdom

[3]Human Factors Research Group, Faculty of Engineering, University of Nottingham, Nottingham, United Kingdom

[4]National Institute for Health and Care Research Applied Research Collaboration East Midlands, School of Medicine, University of Nottingham, Nottingham, United Kingdom

[5]Patient Advisory Board, RADAR-CNS, Kings College London, London, United Kingdom

[6]Freiburg Epilepsy Center, Medical Center, Faculty of Medicine, University of Freiburg, Freiburg, Germany

[7]Division of Child Health, Obstetrics and Gynaecology, School of Medicine, University of Nottingham, Nottingham, United Kingdom

[8]National Institute for Health and Care Research Biomedical Research Centre for Mental Health, Kings College London, London, United Kingdom

[9]Nottinghamshire Healthcare National Health Service Foundation Trust, Nottingham, United Kingdom

[10]School of Biomedical Engineering and Imaging Sciences, King’s College London, London, United Kingdom

[11]Department of Psychiatry, Psychosomatic Medicine and Psychotherapy, Goethe University, Frankfurt am Main, Germany

[12]Department of Psychiatry, Radboud University Medical Center, Nijmegen, Netherlands

[13]Donders Institute for Brain, Cognition and Behavior, Radboud University, Nijmegen, Netherlands

[14]Department of Mental Health and Addictions, Aziende Socio Sanitarie Territoriali Melegnano e della Martesana, Lombardy, Italy

[15]Department of Clinical and Experimental Sciences, University of Brescia, Brescia, Italy

[16]Oxford Epilepsy Research Group, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, United Kingdom

[17]Department of Neurology, Rigshospitalet Glostrup, University of Copenhagen, Copenhagen, Denmark

[18]Istituto di Ricovero e Cura a Carattere Scientifico Istituto Centro San Giovanni di Dio Fatebenefratelli, Brescia, Italy

[19]Department of Neurology, Parc Sanitari Sant Joan de Déu, Barcelona, Spain

[20]Centre d'Atenció Primària Sant Llàtzer, Consorci Sanitari de Terrassa, Barcelona, Spain

[21]Parc Sanitari Sant Joan de Deu, Barcelona, Spain

[22]Servei de Neurologia-Neuroimmunologia, Centre d’Esclerosi Múltiple de Catalunya (Cemcat), Vall d’Hebron Hospital Universitari, Barcelona, Spain

[23]Vall d’Hebron Institut de Recerca, Barcelona, Spain

[24]Kings College London, London, United Kingdom

Corresponding Author:
Jacob A Andrews, BA, MA, PhD
National Institute for Health and Care Research MindTech MedTech Co-operative
University of Nottingham
Institute of Mental Health
Triumph Road
Nottingham, NG7 2TU
United Kingdom
Phone: 44 01157484 218
Email: jacob.andrews@nottingham.ac.uk

Abstract

Background: Multiple sclerosis (MS), epilepsy, and depression are chronic central nervous system conditions in which remote measurement technology (RMT) may offer benefits compared with usual assessment. We previously worked with clinicians, patients, and researchers to develop 13 use cases for RMT: 5 in epilepsy (seizure alert, seizure counting, risk scoring, triage support, and trend analysis), 3 in MS (detecting silent progression, detecting depression in MS, and donating data to a biobank), and 5 in depression (detecting trends, reviewing treatment, self-management, comorbid monitoring, and carer alert).

Objective: In this study, we aimed to evaluate the use cases and related implementation issues with an expert panel of clinicians external to our project consortium.

Methods: We used a Delphi exercise to validate the use cases and suggest a prioritization among them and to ascertain the importance of a variety of implementation issues related to RMT. The expert panel included clinicians from across Europe who were external to the project consortium. The study had 2 survey rounds (n=23 and n=17) and a follow-up interview round (n=9). Data were analyzed for consensus between participants and for stability between survey rounds. The interviews explored the reasons for answers given in the survey.

Results: The findings showed high stability between rounds on questions related to specific use cases but lower stability on questions relating to wider issues around the implementation of RMT. Overall, questions on wider issues also had less consensus. All 5 use cases for epilepsy (seizure alert, seizure counting, risk scoring, triage support, and trend analysis) were considered beneficial, with consensus among participants above the a priori threshold for most questions, although use case 3 (risk scoring) was considered less likely to facilitate or catalyze care. There was very little consensus on the benefits of the use cases in MS, although this may have resulted from a higher dropout rate of MS clinicians (50%). Participants agreed that there would be benefits for all 5 of the depression use cases, although fewer questions on use case 4 (triage support) reached consensus agreement than for depression use cases 1 (detecting trends), 2 (reviewing treatment), 3 (self-management), and 5 (carer alert). The qualitative analysis revealed further insights into each use case and generated 8 themes on practical issues related to implementation.

Conclusions: Overall, these findings inform the prioritization of use cases for RMT that could be developed in future work, which may include clinical trials, cost-effectiveness studies, and the commercial development of RMT products and services. Priorities for further development include the use of RMT to provide more accurate records of symptoms and treatment response than is currently possible and to provide data that could help inform patient triage and generate timely alerts for patients and carers.

(JMIR Neurotech 2023;2:e41439) doi:10.2196/41439

KEYWORDS

mobile health; mHealth; wearable; wearable technology; smartphone; use case; implementation; epilepsy; multiple sclerosis; MS; depression; depressive disorder; Delphi; remote measurement technology; RMT; central nervous system; nervous system disorder; neurology; neurological disorder; expert panel; expert opinion; perspective; mobile phone

Introduction

Background

Digital and mobile health technologies, including smartphone-based monitoring and wearable devices, have a wide range of applications in clinical practice [1-3]. A clinical “use case” describes how a technology can be implemented in a clinical context, including the expected benefit and expected beneficiary. Clinical use cases are essential for determining the outcomes to be used in trials evaluating effectiveness as well as for obtaining regulatory approvals and explaining the benefits of a health care technology to potential funders and patients. The adoption and scaling of novel technologies in health care are dependent on a well-defined use case with a clearly defined problem to be addressed [4]. The inclusion of clinicians in the development of such technologies is known to be important for successful implementation, as it ensures the appropriateness of technology for the specific requirements of patients and the health care system [5].

Remote Assessment of Disease and Relapse–Central Nervous System (RADAR-CNS) was a 6-year project to understand the feasibility and acceptability of using remote measurement technology (RMT) to collect health-relevant data from individuals living with epilepsy, multiple sclerosis (MS), or depression [6]. The project was a collaboration across 6 European countries (Denmark, Germany, Italy, the Netherlands, Spain, and the United Kingdom) and has involved the development of a bespoke, open-source platform RADAR-base. The platform collates data from commercially available Fitbit smart watches measuring activity, heart rate, and heart rate variability; the Empatica E4 wrist-worn epilepsy seizure detection device; Bittium Faros accelerometer and electrocardiogram Holter devices; and bespoke apps for passive sensing and active collection of user-entered data (THINC-it) [7]. We refer to the combination of the platform, the apps, and the commercial devices as the RADAR-CNS RMT system. Observational studies have been conducted to establish the feasibility and acceptability of collecting data from individuals living with MS, epilepsy, or depression using these sensors, apps and platform to develop new predictive algorithms based on the data set [8-10]. Patient involvement has been conducted throughout the program, and patient focus groups and other involvement studies have been conducted in multiple European countries to elicit patient views and inform the RMT under development [10-13].

The aim of this study was to specify priority use cases for RMT in 3 central nervous system disorders (epilepsy, MS, and depression). An initial set of 13 use cases were developed through discussion with health care professionals (HCPs) and researchers working in each of the 3 clinical work packages within the project. The development of these use cases considered the fit to the target population, the potential for a positive impact on the health and safety of patients, whether the use case would offer an improvement on current methods, and the existence of prior evidence to support the use case. These were also informed by our prior work, which included: a small-scale survey with patient advisers, HCPs, and researchers [14]; in-depth interviews with HCPs [15]; and a large-scale survey of 1006 clinicians on the current and potential use of RMT and apps in clinical practice [16] and the potential value of remote measurement data [17]. The Delphi study then sought to prioritize among the 13 use cases (5 in epilepsy, 3 in MS, and 5 in depression) to determine which of these would be most practicable and useful in the eyes of the expert clinician panel, who were outside of the consortium and so offered a more objective point of view. The number of use cases included in the study was considered to be manageable without overburdening participants.

The use cases were also presented to the RADAR-CNS Patient Advisory Board (PAB) to seek further input ahead of this study in a short consultation via Microsoft Teams, with diagrams and descriptions of use cases provided by email in advance. The RADAR-CNS PAB includes members living with each of the 3 conditions from multiple countries across Europe. Illustrations of the 13 use cases are included in Multimedia Appendix 1.

The final use cases for epilepsy are as follows:

1. Seizure alert: enabling real-time seizure warnings to patients and carers.
2. Seizure counting: improving detection of different types of seizures to enable more accurate overall seizure records.
3. Risk scoring: detecting cycles of seizure occurrence to reveal risk levels at different times.
4. Triage support: enhancing patient triage based on RMT data submitted wirelessly to patient record systems.
5. Trend analysis: reliably detecting a change in the number of seizures that a patient has over a specified period.

The final use cases for MS are as follows:

1. Detecting silent progression: making use of more granular measurements to detect otherwise invisible markers of progression, enabling patients to evidence changes they experience.
2. Detecting depression: identifying markers of depression in the first year after MS diagnosis.
3. Data donation: automatic collection and storage of patient data in biobanks or mega-databases.

The final use cases for depression are as follows:

1. Detecting trends: detailed symptom tracking and aggregation of multiple types of data.
2. Treatment review: measuring adherence to cognitive behavioral therapy or other treatment regimens and treatment response.
3. Self-management: monitoring and providing nudges to a patient to improve their condition.
4. Comorbid monitoring: detecting depression in patients with chronic physical health conditions.
5. Carer Alert: providing an alert to a carer or relative when a person with depression is in a period of very low activity.

Aims and Objectives

The aim of this work was to prioritize the use cases with the potential for the greatest benefit for further development according to the views of HCPs external to the project. We sought to establish the type of benefit that each might offer according to a medical device design framework [18]. We also aimed to explore further related issues:

1. Acceptability of the level of burden (clinical time) required to apply RMT in practice.
2. Acceptability of the amount of data that would be generated by RMT.
3. The extent of required technical support for clinicians to make the best use of RMT.
4. Preferred mode of training or technical support for clinicians.
5. The extent of required technical support for patients to make the best use of RMT.

We aimed to seek a consensus in these areas, where prior work has shown that there is a disagreement between HCPs. In addition, we aimed to determine which prior known concerns about the use of RMT in clinical practice would actually prevent or discourage HCPs from using RMT with patients. As we recruited an international sample, we were also interested in exploring how the potential implementation of RMT might differ between countries.

Methods

Overview

The Delphi methodology has been widely used in health care research to gather expert opinions [19-21]. Key characteristics of the Delphi method include consultation of experts, elements of iteration and feedback to participants to enable a form of communication between them, and statistical methods used to summarize group responses to ensure the robustness of analyses [22]. Delphi studies can be used without face-to-face contact while still enabling the gathering of group opinions, which is of benefit when those whose opinions are required have busy schedules (eg, in clinical settings) or may be located across multiple countries [23-25], as is the case in this study. There were also obvious benefits to this approach during the COVID-19 pandemic.

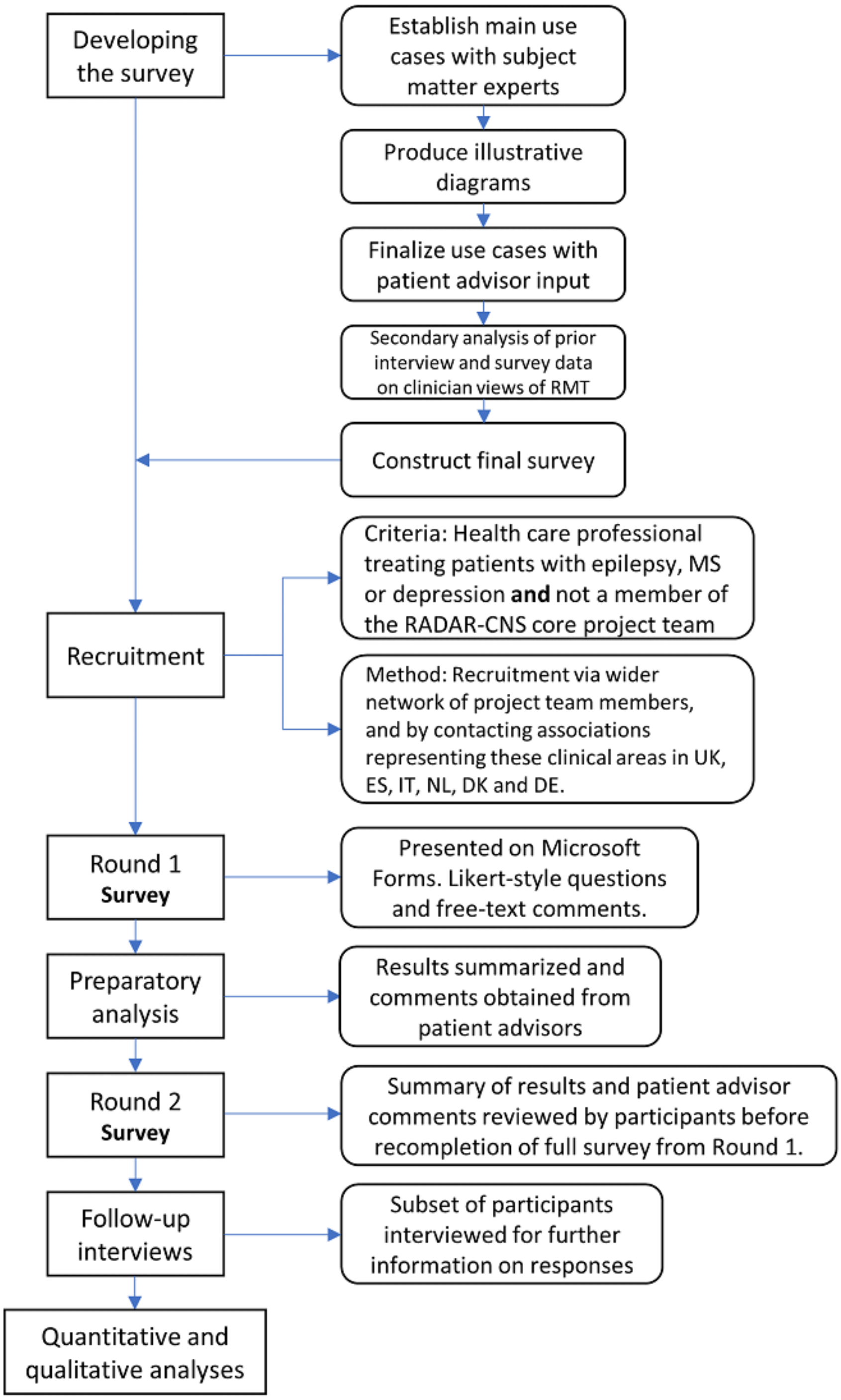
Delphi studies feature multiple survey rounds, with feedback given to the participants between each. For example, Murphy et al [21] used a 3-round model to gather views and opinions on the potential of digital tools for mental health in the United Kingdom. The first round of a Delphi study typically asks open-ended questions, and these are used to generate closed questions for a survey in the second round, often with Likert-style responses. In the third round (if there is one), participants review the summarized results from the prior round and are then able to change their responses if required [26]. This results in either greater consensus among the groups or sustained disagreement, both of which are of interest [25]. Other models may omit the first qualitative round [19] and may include follow-up interviews after the final survey round [27].

Procedure

Overview

This study adopted the Delphi methodology for the context of RADAR-CNS. The study procedure is summarized in Figure 1. As the project had already canvassed opinions from HCPs in surveys and interviews, we replaced the first qualitative round with a reanalysis of our existing data to generate the survey for use in this study. It is recommended that Delphi surveys be completed within 30 minutes [25]. Thus, we used diagrams of use cases (included in Multimedia Appendix 1) to aid rapid comprehension and to permit engagement with the ideas presented.

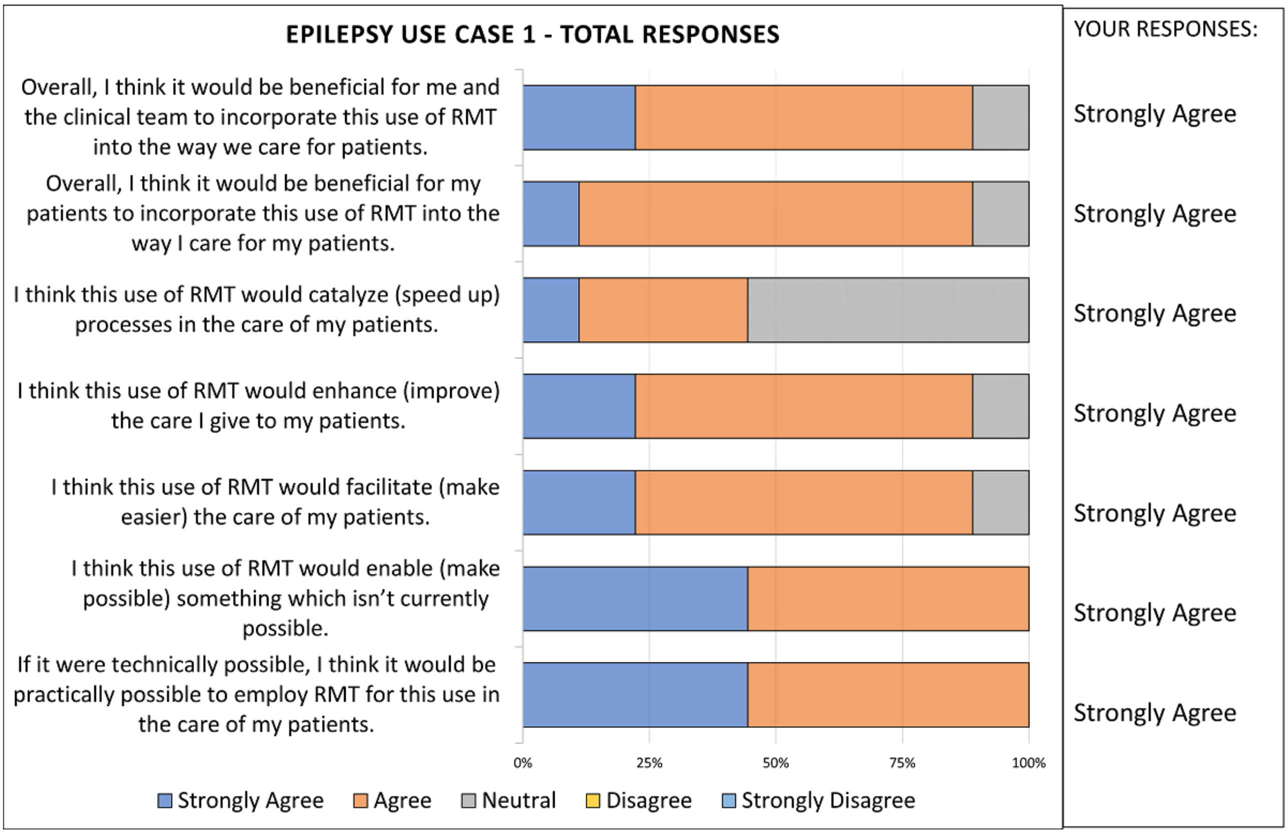
Figure 1. Flowchart showing study process. DE: Germany; DK: Denmark; ES: Spain; IT: Italy; MS: multiple sclerosis; NL: Netherlands; RADAR-CNS: Remote Assessment of Disease and Relapse–Central Nervous System; RMT: remote measurement technology; UK: United Kingdom.


Developing the survey
Establish main use cases with subject matter experts
Produce illustrative diagrams
Finalize use cases with patient advisor input
Secondary analysis of prior interview and survey data on clinician views of RMT
Construct final survey
Recruitment
Criteria: Health care professional treating patients with epilepsy, MS or depression and not a member of the RADAR-CNS core project team
Method: Recruitment via wider network of project team members, and by contacting associations representing these clinical areas in UK, ES, IT, NL, DK and DE.
Round 1 Survey
Presented on Microsoft Forms. Likert-style questions and free-text comments.
Preparatory analysis
Results summarized and comments obtained from patient advisors
Round 2 Survey
Summary of results and patient advisor comments reviewed by participants before recompletion of full survey from Round 1.
Follow-up interviews
Subset of participants interviewed for further information on responses
Quantitative and qualitative analyses

There was a gap of 3 months between the dissemination of the first survey round and the dissemination of the second survey round. After the first round, the research team gathered responses and produced graphs and tables to form a Summary of Results document. In a further adaptation to the traditional Delphi methodology, this Summary of Results was presented to the RADAR-CNS PAB to request commentary on clinicians' responses. Patients who reviewed the round 1 responses consisted of 2 people: 1 living with epilepsy (male) and 1 living with MS (female). Unfortunately, the members of the PAB living with depression did not respond to requests to provide comments. Patients were sent a summary of round 1 results in graphs, tables, and free-text comments, along with a short video explaining what was expected from them. They wrote their comments in a word document or email and returned them to the first author.

For round 2, HCP participants received the Summary of Results, which incorporated patient comments, together with a link to the round 2 survey for completion, and were instructed to review the Summary of Results before completing the second round. The Summary of Results was personalized to each participant, with their own responses indicated next to graphs showing summary responses (Figure 2). Graphs were used to provide an "at-a-glance" overview of the results for quick interpretation. Free-text comments from Delphi experts and from patients were provided in boxes below the graphs for participants to review.

Figure 2. An example showing the style of feedback provided to Delphi panel members after round 1 and showing how individual responses were combined with group responses in the Summary of Results. RMT: remote measurement technology.



Delphi studies have been criticized for their closed nature, which prevents discussion of the results [26]. This study sought to overcome this criticism by including follow-up interviews to discuss the results with individual participants. Therefore, we used a mixed method sequential explanatory approach to collect our data. Figure 1 shows a flowchart summarizing the process.

Recruitment

HCPs were recruited from multiple countries where RADAR is active via multiple routes. The main method of recruitment was by clinical academics within the RADAR-CNS consortium disseminating recruitment materials to their clinical academic colleagues. In addition, we contacted specialist associations representing clinicians treating each of the 3 conditions across the 6 European countries of the consortium (the United Kingdom, Germany, Italy, Spain, the Netherlands, and Denmark). We also contacted prior research participants external to the consortium and contacted clinicians who had previously expressed interest in the RADAR-CNS project.

Inclusion Criteria

We required participants to be experts in the area by virtue of their experience working in the clinical care of people with epilepsy, MS, or depression. We specified that participants should not be members of the RADAR-CNS consortium to gain an external view of the potential of the technology in clinical practice.

Survey Design

The survey was composed of 4 main sections: demographics, use case evaluation, questions related to the implementation of RMT, and rating of concerns. The full survey is included in Multimedia Appendix 2.

There were 3 separate use case evaluation sections, one for each clinical condition. The epilepsy section evaluated 5 use cases, the depression section evaluated 5 use cases, and the MS section evaluated 3 use cases. For each use case, there were 7 questions, which included general evaluations of practicality and benefit, plus 4 adapted from the framework of medical device design by Sharples et al [18]. Using this framework, we sought to identify the specific type of benefit offered by each use case: whether it "enabled" something new, "enhanced" existing practices, "facilitated" (made easier) existing processes, or "catalyzed" (sped up) existing processes. Respondents rated each item on a 5-point scale, ranging from 1 (strongly agree) to 5 (strongly disagree).

The "further questions" section covered clinical time, frequency of data collection, technical support requirements, usefulness or value of data, and payment and reimbursement. The "rating of concerns" section was intended to explore the extent to which different barriers to use discovered in our prior work would affect respondents' intention to use RMT in clinical practice. Each response option in this section was more detailed and required longer to read than those in the previous sections, so we kept the number of response options low to facilitate completion (the options were "Would not prevent me using RMT with my patients"; "Would prevent use in some situations"; "Would prevent use entirely"; or "Don't know").

The second round survey was identical to the first, except that demographics questions were omitted, and coauthorship of the resulting paper was offered via an opt-in tick box. We chose to request full recompletion of the survey rather than only requesting completion of questions where consensus had not been reached because we were equally interested in areas of disagreement as we were in gaining consensus.

Follow-up Interviews

A subset of participants who indicated interest in a follow-up interview were contacted to arrange a 30-minute slot for a web-based interview using Microsoft Teams. The interviews followed a semistructured format using an interview guide instrument (Multimedia Appendix 3). The aims of the interviews were to gain further insight into HCPs' views of the use cases for the RADAR-CNS RMT system and to understand country-specific contextual factors that might affect the implementation of RMT in each country. As such, we conducted interviews across a range of European countries and across the 3 conditions.

Ethics Approval

The methods were performed in accordance with relevant guidelines and regulations and were approved by the University of Nottingham Faculty of Medicine and Health Sciences Research Ethics Committee (ref: 315-0721; Multimedia Appendix 4).

Analysis

Quantitative Analysis

The quantitative data consisted of Likert-style responses from the 2 survey rounds. These were scored from 1 (strongly disagree) to 5 (strongly agree), except for the final question on barriers to RMT use, which was scored using a 3-point scale, plus an option for "do not know." Numbers reporting "do not know" were included in the denominator of the percentage calculations.

To evaluate the consensus among respondents in each round, we used a predetermined threshold percentage of similar responses on an item [25]. We determined that consensus had been reached if 70% of the responding participants scored an item within the same grouping (agree, neutral, or disagree), where scores of 4 or 5 were grouped as "agree," scores of 1 or 2 were grouped as "disagree," and scores of 3 were considered neutral. This effectively recreated a 3-point scale, which is considered preferable over the analysis of 5-point scales for survey data in clinical contexts [28].

To evaluate the stability between rounds, we used the Gwet agreement coefficient, which is found to have more stable performance than kappa scores [29]. The scores were weighted to account for ordinality in the variables. The coefficients were compared against benchmarks from Altman [30].

Analyses were conducted only on responses received—no imputation was judged to be necessary to account for missing data, given that the study focused on eliciting a small number of expert opinions, with only descriptive statistics used to analyze the data.

Qualitative Analysis

Survey comments and interview transcripts were analyzed using template analysis [31]. An a priori theme was included in the initial template for each use case, and for each implementation topic covered in the survey. Themes were iteratively added, deleted, renamed, and reorganized to create the final template of themes. We used the final template of themes to triangulate the qualitative and quantitative data.

Results

Participants

A total of 23 clinicians treating patients with epilepsy, MS, or depression were recruited, with participation from all 6 European countries where RADAR-CNS is active (the United Kingdom, Germany, Spain, Denmark, the Netherlands, and Italy) and with representation from clinicians treating each of the 3 clinical conditions (Table 1). We expected some dropouts between the first and second rounds but were able to retain 74% (17/23) of the round 1 participants in round 2. A total of 9 respondents completed the interviews.

Table 1. Participant demographics.

	Round 1 (n=23), n (%)	Round 2 (n=17), n (%)	Interviews (n=9), n (%)
Age group (years)			
30-39	6 (26)	6 (35)	1 (11)
40-49	5 (22)	4 (24)	1 (11)
50-59	9 (39)	5 (29)	5 (56)
60-70	3 (13)	2 (12)	2 (22)
Gender			
Woman	8 (35)	6 (35)	3 (33)
Man	15 (65)	11 (65)	6 (66)
Job role			
Consultant (medical)	16 (70)	11 (65)	8 (89)
Health care scientist or researcher	2 (9)	2 (12)	0 (0)
Clinical psychologist	1 (4)	1 (6)	0 (0)
General practitioner	1 (4)	1 (6)	0 (0)
Nurse	1 (4)	1 (6)	1 (11)
Psychological well-being practitioner	1 (4)	0 (0)	0 (0)
Other (unspecified)	1 (4)	1 (6)	0 (0)
Relevant condition treated			
Depression	10 (43)	8 (47)	3 (33)
Epilepsy	8 (35)	6 (35)	4 (44)
Multiple sclerosis	3 (13)	2 (12)	1 (11)
Multiple sclerosis, epilepsy, and depression	1 (4)	0 (0)	0 (0)
Epilepsy and multiple sclerosis	1 (4)	1 (6)	1 (11)
Country			
United Kingdom	8 (35)	5 (29)	4 (44)
Spain	5 (22)	4 (24)	1 (11)
Italy	4 (17)	3 (18)	0 (0)
Germany	3 (13)	2 (12)	3 (33)
Denmark	1 (4)	1 (6)	0 (0)
Australia	1 (4)	1 (6)	0 (0)
Netherlands	1 (4)	1 (6)	1 (11)

Quantitative Results: Consensus and Stability

The research team decided that a third survey round was not required: 97.4% (114/117) of question items had a high or very high level of stability of responses between rounds 1 and 2, indicating that a third round would have had limited benefit. Multimedia Appendix 5 [30] presents the results for consensus and stability for all questions in the survey.

Epilepsy Use Case Questions

The threshold for consensus was reached on 74% (26/35) of questions on the epilepsy use cases in the first round and 86% (30/35) of questions in the second round, demonstrating a move toward consensus. For all of these items, consensus was reached that respondents agreed or strongly agreed with the statements presented, rather than selecting “disagree,” “strongly disagree,” or “neutral.” Five of the questions moving to consensus in round 2 concerned the fifth epilepsy use case, using RMT for trend analysis, and indicated a change in views toward agreement that this use case would “enable” new possibilities, “facilitate” care (make care easier), “enhance” care, and benefit patients and clinical teams.

It is notable that in round 1, for epilepsy use cases 1 to 3, the question on “catalyzing” (speeding up) existing processes received fewer "agree" or above responses than all other statements, indicating less confidence that these use cases would speed up existing processes. The PAB identified this pattern, and their comments were fed back to the participants ahead of round 2. There was comparatively low stability for these questions between round 1 and round 2, suggesting that participants changed their minds about this question, perhaps in response to the PAB’s comments.

The point estimate for the Gwet agreement coefficient statistic fell in the “very good” strength of agreement range (0.80-1.00) for 94% (33/35) of items, indicating a very high overall stability of epilepsy responses between rounds 1 and 2.

MS Use Case Questions

The threshold for agreement was reached on 10% (2/21) of questions on the MS use cases in the first round and 0% (0/21) of questions in the second round, demonstrating a move away from consensus. Where consensus was reached, it was a consensus that respondents “agreed” or “strongly agreed” with the statements presented rather than selecting “disagree,” “strongly disagree,” or “neutral.”

Fewer participants completed the MS use case questions in the second round (n=3) compared with the number completing epilepsy questions (n=6) and depression (n=9). This meant that even when a majority of 67% (2/3) of participants expressed an opinion, this did not cross the threshold of 70%, requiring a unanimous vote for this to occur. This explains the comparatively lower number of questions showing consensus in the MS group.

The point estimate for the Gwet agreement coefficient fell in the “very good” strength of agreement range (0.80-1.00) for 81% (17/21) of items, indicating a high overall stability of responses between rounds 1 and 2.

Depression Use Case Questions

In the depression use cases, the threshold for agreement was reached on 83% (29/35) of questions in the first round and 89% (31/35) of questions in the second round. Where consensus was reached, it was a consensus that respondents “agreed” or “strongly agreed” with the statements presented, rather than selecting “disagree,” “strongly disagree,” or “neutral.”

The point estimate for the Gwet agreement coefficient fell in the “very good” strength of agreement range (0.80-1.00) for all 100% (35/35) of items, indicating a very high overall stability of responses between rounds 1 and 2.

Further Questions Section

There were 19 questions on further considerations of RMT (clinical time, frequency of data collection, technical support requirements, usefulness or value of data, and payment and reimbursement), which were rated by all 17 participants who completed both rounds. Participants reached a consensus of “agree” or “strongly agree” on 42% (8/19) of questions in round 1 and maintained this level of consensus in round 2 (Multimedia Appendix 5). A smaller proportion of questions in this section reached consensus than those in the section on use cases. For 11% (2/19) of questions in this section, participants moved from no consensus to a consensus that they “disagreed” or “strongly disagreed” with the statement. These questions were that “receiving data on a patient’s condition would be an added burden” and that “mood scores need to be collected from patients at risk of mental health conditions on a daily basis.” A total of 9 question items in this range did not reach agreement in the first or second rounds. The Gwet agreement coefficient showed “very good” stability between rounds (Altman benchmark 0.80-1.00) on only 63% (12/19) of questions in this range, indicating greater changeability between rounds for these questions compared with those relating to the use cases.

Concerns Questions

There were 7 questions on concerns about RMT, which could be rated as a serious concern (“Would prevent use entirely”), a medium concern (“Would prevent use in some situations”), or a lesser concern (“Would not prevent me using RMT with my patients”). There was no consensus for any question in this set in round 2.

The stability of responses between rounds 1 and 2 for these questions was lower than that for other parts of the survey. The change in responses was not uniform in one direction or the other, and neither was there a distinct movement toward or away from extreme responses (rating a concern as severe or lesser), indicating less certainty in relation to these questions compared with other sections of the survey.

Qualitative Findings: Final Template and Triangulation

Overview

The triangulation of the results is interwoven with the overall exposition of the qualitative results below. The final template consisted of 8 themes, each with multiple subthemes (Table 2).

Table 2. Final template of themes and subthemes.

Theme		Subthemes
1	Comments on specific use cases	• Depression (UCs[a] 1-5, general comments) • Epilepsy (UCs 1-5, general comments) • Multiple sclerosis (UCs 1-3, general comments)
2	Clinical time	• Implementing RMT[b] would be time costing • Implementing RMT would be time-saving • Other views on RMT and clinical time
3	Value of RMT data	• Disease-specific value • Moving beyond the subjective • Positive or negative views on value • Value is related to amount of data accessible
4	Frequency, amount, and type of data collection	• Collecting large amounts of data (over a year) • Daily reporting and recording • Desired frequency of data collection • Passive data collection vs active data collection • Technical support and its effect on clinical time
5	Payment and reimbursement	• Funding in clinical settings to support introduction of new technologies • Political drivers • Requirement for extra resource • Requirement to save costs or improve care
6	Country or context-specific factors relating to RMT implementation	• Germany • Netherlands • Spain • United Kingdom • Setting-specific factors
7	Inevitability of change and ongoing change in health care services	• Preference for at-distance care • Patients use RMT and bring data to clinic • Patients use RMT but don’t bring data to clinic • Coronavirus pandemic as stimulus for change
8	Barriers and concerns	• Comments on barriers listed in the survey • Clinician time • False alarms, false positives, false negatives • Interoperability • Patient anxiety • Reducing number of appointments • Other barriers not covered in the survey • Requirement for further research • Health care culture • Legal and regulatory • Patient behavior and situation

[a] UC: use case.

[b] RMT: remote measurement technology.

Comments on Specific Use Cases

The results on condition-specific use cases from the interviews inform the prioritization of use cases, as participants indicated which of the use cases they would find most useful and which least useful, with reasons to support these indications. Extracts from the interview transcripts for each use case are provided in Multimedia Appendix 6.

Epilepsy

All 5 epilepsy use cases were considered plausible, although participants stated that their utility depended on practicality and accuracy. Use cases 1, 2, and 4 (seizure alert, seizure counting, and triage support, respectively) were considered the most useful. This supported the quantitative data across both rounds.

Use case 1 (seizure alert) was considered helpful for motor seizures, which are highly associated with a sudden unexpected death in epilepsy. One participant working in Germany questioned the novelty of the solution (“we already have this for some devices" [Participant 2]), although it is understood that this is only for a patient at rest (not moving) and there is still a need for wearables that can detect motor seizures from active status. Participants indicated that an adequate level of sensitivity and specificity would be required, with 1 participant providing a detailed account of acceptable sensitivity and specificity (Multimedia Appendix 6).

Epilepsy use case 2 (seizure counting) was also considered useful, assuming appropriate levels of accuracy. Comments indicated that passive monitoring may be more accurate than patient diaries, for example, where a patient might forget to record some seizures. It was considered not to be feasible for clinicians to review data between clinic visits unless the system indicated the requirement for additional review based on particular thresholds and therefore performed some sort of triage.

Use case 3 (risk scoring) was thought to be less practical and more difficult to achieve. Interviewees thought there would be medicolegal risks and that they would not want to prevent patients from taking part in enjoyable activities where unnecessary. Use case 4 (triage support) was considered useful but less so than use cases 1 and 2. Concerns included lack of infrastructure, false positives, staffing resources, legal complications, and low availability of staff.

Use case 5 (trend analysis) was again considered potentially useful depending on evidence to support its effectiveness. Some responses to interview questions indicated that trend analysis could be one of the most useful applications of RMT in epilepsy, although quantitative findings from round 1 did not reach agreement.

Multiple Sclerosis

MS use case 1 (detecting silent progression) was considered useful for detecting progression early enough to slow down the condition. However, its benefits were considered to be restricted by the limited availability of medications to treat disease progression. Measuring gait was thought to be a useful mechanism for detecting silent progression ("to detect progression, the most useful would be all the tools that would be used to detect gait disorders" [Participant 3]).

There were mixed views on the usefulness of detecting depression in MS (use case 2). One interviewee indicated that the use of RMT in this way could be useful to open "a bit more conversation" with the patient (participant 19). Another interviewee stated, "it may be that detecting depression would show the development of the disease, but that would not help us so much" (participant 3). These contrasting views reflect the lack of consensus among experts in the quantitative survey results.

Use case 3 (biobanking MS data) was considered useful for future patients but not for current patients ("that it is very useful to collect this data, so I'll be interested, [...] but it will not necessarily have a direct impact to my patients" [Participant 19]), which explains the lack of consensus on the question about patient benefit. It was highlighted that there already exist biobanks for MS data and that RMT data could be added to these.

Depression

Interviewees indicated that use cases 1, 2, and 3 (detecting trends, reviewing treatment, and self-management, respectively) would be the most useful. Use case 1 (detecting trends) was thought to enable easier and better recording of patient-reported outcomes, which could save administrative time. Use case 2 (reviewing treatment) was considered useful if it could be implemented successfully within the treatment pathway. It was considered that use case 3 (self-management) would work well for some (but not all) patients.

It was considered that use case 4 (comorbid monitoring) might increase the rate of detecting depression ("we might encounter much more depression if we manage to do this" [Participant 12]), but this was considered the least viable use case, partly because it may not be possible to effectively treat comorbid depression if found, and partly because of concerns about confounding symptoms. These findings supported the quantitative data, where consensus was only reached on 4 of 7 questions about use case 4, compared with 6 of 7 or 7 of 7 for all other use cases. Use case 5 (carer alert) was also considered less useful, as participants thought that carers might not take on the required responsibility, being unwilling or unable to offer the right support and care.

General or Non–Condition-Specific Questions

Clinical Time

The responses showed that implementing RMT could be overall time saving if it reduced admissions, although the potential for RMT to identify otherwise unidentified symptoms may in fact require more clinical time to evaluate. RMT may reduce emergency department burden, where conditions are better managed. Comments suggested that time saving would depend on high accuracy. Several participants described practical ways in which RMT could be used to save time, eg, the use of thresholding, and having a patient manager specifically trained to manage RMT data. Some also suggested that the value of RMT may be in having a more detailed picture to improve care rather than saving time. Interviewees stated that having good quality, easily available technical support could save time for clinicians and encourage the continued use of RMT, although some were concerned about the cost of technical support.

Value of RMT Data

Interviewees suggested that RMT data would be useful in conditions outside the 3 covered in RADAR-CNS, eg, in the monitoring of bipolar disorder:

> *I wonder whether you also consider bipolar disorder if you talk about depression, I think you have to. Even if patients come from the unipolar depression side, they still might switch into mania.* [Participant 11]

To some extent, the value of the RMT data was correlated with the amount of data that could feasibly be collected. It was considered that "the more data you have, the less uncertainty there is" (participant 4) and that the data could give otherwise unavailable insights into patient condition:

> *Having the RMT background information in terms of their activities throughout the week, I think it will*

> *probably give us a little bit more information in terms of how they've been during the week rather than, you know, on that day, this is what they reported.* [Participant 5]

This linked with a wider subtheme on "moving beyond the subjective." Interviewees stated that RMT could provide objective data that would otherwise be represented only by subjective patient self-reports. It was also considered that RMT would enable clinicians to determine how much the patient's condition affected their everyday lives even when the patient said they were "fine."

Frequency, Amount, and Type of Data Collection

The required frequency of data collection would depend on the stage of the disease and treatment phase. In relation to mood, participants' desired frequency of mood report data ranged from daily to once fortnightly with various suggestions in between.

There was a suggestion that passive data collection was more valuable than active data collection, as compliance with active measures was expected to be low and because active measures can have the undesirable effect of inducing negative mood states:

> *From another trial, assisted active monitoring can even induce bad mood states because people then tend to ruminate, tend to think about the situation more than they probably should. So it should stress passive monitoring.* [Participant 11]

Payment and Reimbursement

Many participants expressed that the introduction of RMT in these use cases would require a large amount of extra resources, for the cost of devices, for staff who would monitor patient data, and for staff members who would help patients set up the technology. Interviewees also expressed that the essential requirement for the introduction of RMT would be that it could demonstrably save costs or provide strong evidence of an improvement in patient care:

> *Whatever you implement must not increase your workload, because otherwise especially doctors in their own practice won't use it because they don't get any extra money for that.*
>
> *So they must see a time saving benefit or a real quality improvement for patient care. That is what they expect, and here it's really important to stress that it's not putting extra work on doctors, but makes life easier actually.* [Participant 11]

There were mixed views on whether reducing the number of patient appointments, as a result of monitoring their condition remotely, would be useful. This reflected the quantitative survey results, where 9 of 17 reported that their service would lose money if appointments were reduced and 7 of 17 reported that their service would not. Payment regulations for a clinician treating epilepsy in Germany meant that a reduction in the number of appointments would cause a reduction in income for his service. He stated that a political change would be required to incentivize the use of new wearable solutions in epilepsy. Conversely, a clinician treating epilepsy in the United Kingdom stated that where appointments were saved for 1 patient, these would be filled by another, as the demand for the service was so great: "There's too much demand, so don't worry about dropping income because of dropping demands" (participant 4).

Country- or Context-Specific Factors Relating to RMT Implementation

Germany

Interviewees in Germany gave mixed reports on the potential for the reimbursement of RMT. One interviewee said that some wearable devices were already provided to patients with epilepsy, paid for by the health service, providing evidence of a precedent for the funding of RMT. They added that RMT may have limited cost-saving benefits because of the requirement for additional staff. Another interviewee said that the organization of remuneration for health care in Germany is old fashioned and limits the ability to introduce new technologies. Another said that where doctors run their own clinics, they are free to use any technologies they see fit as long as they can convince their budget manager of the benefit the new technology will offer. This interviewee also mentioned the German law introduced in recent years to incentivize the introduction of digital health technologies and described the requirement for these technologies to provide strong supporting evidence:

> *They need to demonstrate, well, evidence for helpfulness. It's not the level of a randomized control trial, but they need to have data that the app would be instrumental in reducing health care burden, and then the provider gets reimbursed. So it's like prescribing a medication or something like that.* [Participant 11]

The Netherlands

The single interviewee from the Netherlands explained that there is hope and enthusiasm that RMT may offer patient and health care service benefits in depression in the Netherlands, but that there is as yet little implementation. They contrasted the National Health Service (NHS) in the Netherlands with the situation in Germany, where it was perceived that individual German hospitals needed to attract patients and that RMT may offer a competitive advantage, whereas the Dutch NHS did not need to do so.

Spain

The interviewee treating patients with MS in Spain suggested that regulatory factors might complicate the introduction of RMT in Spain and that there was little money in the Spanish health service to introduce RMT. However, they mentioned that because of the COVID-19 pandemic, many patients with MS whose condition is stable now have remote visits via videoconferencing as a matter of course, which have laid the cultural groundwork for a change in patient monitoring and management.

The United Kingdom

Interviewees distinguished the United Kingdom from other countries by highlighting how health care practitioners treating MS in other countries may be paid per visit, and it is thus in their interest to have patients attend clinics. However, in the

United Kingdom, no such pay-per-appointment system is in place. Therefore, UK practitioners may be more keen than those in other countries to make the best use of RMT data to cancel appointments where unnecessary.

It was also highlighted that insurer-based health systems compete for patients, but the UK NHS does not, so there would be less motivation to introduce RMT as a competitive advantage in the NHS.

Although interviewees pointed out that there is a political push for increased implementation of digital solutions in the NHS, 1 interviewee suggested it would be politically unpopular for the NHS to offer consumer-grade electronic goods for health-related purposes free on the NHS:

> *Say with an Apple Watch retail price, probably three, 400 pounds, I don't know. You could see the social envy creeping up and saying, oh I'm not paying for epilepsy patients to get an Apple Watch which I can't afford myself so consumer electronics is one thing.* [Participant 9]

Similar to other countries, UK interviewees stated that funding for new medical technologies is focused on research trials rather than on implementation. They stated that implementation is assumed and is not highly regarded in terms of researcher or practitioner prestige.

Setting-Specific Factors

In relation to clinical alert-based systems (ie, those where crossing a threshold in patient RMT data may trigger an alert to a medical team), interviewees stated that these would have easier applications in acute hospital settings rather than in community-based settings. However, there was concern that larger centers or hospitals would be likely to see greater adoption of remote technologies because they have more funds available to cover excess costs and that this would contribute to inequality:

> *You end up having three hospitals that they are already providing a good care to provide a bit better care. So if anything, the inequality of care will widen.* [Participant 19]

Inevitability of Change in Health Services

Interviewees commented that changes were inevitable in health care services and that pathways and procedures are often evolving. In relation to RMT, some interviewees indicated that they were aware of patients already using wearables to monitor health, with some of these sharing data with clinicians (and some not). The coronavirus pandemic was discussed as a stimulus for lasting change that may lay the groundwork for the future implementation of different types of RMT. Respondents reported that some patients were fearful of face-to-face contact with health care practitioners in light of the threat of COVID-19.

Barriers and Concerns

Barriers Covered in the Survey

There were mixed views on whether false positives and false negatives from RMT would be problematic, reflecting the lack of consensus in this area in the quantitative results. One interviewee stated that false positives or negatives would not undermine the usefulness of RMT, as such results can be expected from any measure, whether digital or analog. Interviewees also mentioned interoperability and patient anxiety (covered in the survey), where if RMT made patients more anxious, this could drive the increased use of clinical resources, which could be problematic. Another interviewee mentioned that carers may be made more anxious by the introduction of RMT:

> *Sometimes parents can sometimes focus too much on stuff that's not relevant, and then that gets in the way of them focusing on the more important things.* [Participant 4]

Barriers Not Covered in the Survey

There was concern among interviewees that patients may not adhere well to monitoring regimes; that devices would be lost, stolen, or sold; or that patients may not have suitable internet or mobile data to enable the use of RMT. Other worries were that patients may buy cheap, less accurate, and unregulated devices if the appropriate devices are not provided for free and that the implementation of RMT would only be successful if patients believed that it would work, requiring adequate patient education.

Health care culture was identified as an important barrier. It was suggested that HCPs were often unaware of what technologies are readily available to support their patients. Interviewees stated that, in the United Kingdom, it is very difficult for an HCP to persuade more senior staff members of the necessity for any particular kind of technology that they were aware of:

> *They would say, oh, you gotta do a business case if you want to introduce new technology. The business case is quite difficult to do. There's very little admin support for it, unless it's a very high priority of the trust.* [Participant 4]

Legal and regulatory systems were also highlighted, with interviewees suggesting that these are not currently set up for technologies that are recurrently updated, eg, algorithms that update themselves. Data protection and ownership were also mentioned as key issues worthy of consideration when implementing new technologies.

Further research is necessary to determine the accuracy and reliability of off-the-shelf consumer technologies used within the system. CIs for their precision would be required to make use of these parameters successfully. The participants recommended trials of the specific use cases of the technologies under development to establish cost-effectiveness.

Discussion

Overview

The purpose of this mixed methods, sequential explanatory Delphi study was to prioritize among use cases for RMT in central nervous system disorders, which had been cocreated with clinicians and patients within the RADAR-CNS consortium. The results from the study have identified those likely to be of the most practical use and clinical benefit. The study has also contributed knowledge on country- and context-specific factors affecting implementation and revealed areas of consensus and disagreement among HCPs on practical aspects of RMT implementation.

Principal Findings

Epilepsy

Priority use cases for RADAR-CNS RMT in epilepsy from this study are: seizure alert, seizure counting, triage support, and trend analysis. All 5 use cases for epilepsy were considered "beneficial to patients"; however, use case 3 (risk scoring) was considered less likely to facilitate or catalyze care or be beneficial to clinical teams. Participants suggested that risk scoring would bring medicolegal risks and would be less practical and more difficult to achieve than other use cases.

Although other authors have commented on the potential of technology in these areas [32-34], our study has validated these as the most useful applications of RMT with clinicians treating epilepsy across Europe. Participants in the study were concerned about the possible medicolegal consequences of using devices to estimate the risk of seizures in epilepsy, which might be expected, given that prior work has highlighted the medicolegal responsibilities of clinicians treating patients with epilepsy in relation to driving and employment in the teaching profession [35].

Multiple Sclerosis

There was little consensus on the benefits of the use cases in MS. Despite repeated efforts to avoid dropout, the results in round 2 were only obtained from 3 participants, limiting the robustness of these results. There was greater consensus among the 6 participants completing round 1, particularly for use case 1 (detecting silent progression) and use case 2 (detecting depression in MS). The qualitative findings showed some dependencies, eg, that the system could hold value for the detection of silent progression of the disease if relevant treatments are available, whose delivery could be optimized by applying them at specific times relevant to the timely detection of changes by the system.

On reviewing our findings, patients commented that cognition is an important aspect of MS to measure the silent progression of the disease. They also mentioned that compliance with monitoring programs may be greater where they experienced a decline or instability in their condition. Related work has found that clinicians have concerns about collecting data with little clinical relevance [3,36], and our work illustrates this in the specific case of MS, where HCPs were skeptical about the benefits of detecting silent progression if no treatments were available to address it.

The UK Biobank now includes data on sleep and physical activity from wearable devices, and researchers have begun to analyze these data for various purposes [37]. Here, we provide evidence that some clinicians support such storage of data from patients with MS, although some are unclear about the benefits to the donating patient. Issues around the security and privacy of patient wearable data in biobanks were mentioned by the interviewees. The World Medical Association has adopted a declaration on ethical considerations regarding health databases and biobanks [38], and such issues will need to be duly considered for any future storage and sharing of wearable and smartphone sensor data in biobanks.

Depression

The use cases to be prioritized for RMT in depression include detecting trends, reviewing treatment, and self-management. Participants agreed that there would be benefits for all 5 of the depression use cases, although there was less consensus for use case 4 (comorbid monitoring), where qualitative results showed that participants thought it would be difficult to distinguish between symptoms of depression and symptoms of comorbid physical illnesses. Use case 5 (carer alert) was criticized in the interviews, as informal carers may not have the requisite skills or knowledge to adequately support patients with depression. Use cases 1 to 3 were seen as useful provided evidence could be generated to support their effectiveness.

Self-management was one of the most favored use cases for RMT in depression. There is some evidence supporting the effectiveness of smartphone apps for depression self-management [39], although qualitative evidence shows that users may download apps for short-term, inquisitive trials and may not adhere for longer term use [40]. Further work is required to establish what factors affect adherence to depression self-management apps and how the RADAR-CNS RMT system can be presented to patients to encourage continued use.

Other use cases supported by participants for depression were use case 1, detecting trends and use case 2, reviewing treatment, including monitoring of treatment response and side effects. Existing methods of detecting trends and monitoring treatment response rely predominantly on pen-and-paper mood diaries and outcome measures, such as the Patient Health Questionnaire-9. However, many such outcome measures have been converted to digital versions [41-43], and electronic mood diaries are also becoming available as smartphone or web applications [44,45], with some efforts to automatically detect symptoms and analyze trends from these user-entered data [2]. The multimodal, passive, and active combinations that could be offered using the RADAR-CNS RMT system are less commonly available, although some research has begun in this area [46]. This type of approach likely requires a higher level of regulatory approval than electronic mood diaries [47].

Further Questions

There was lower consensus and stability on these questions than those relating to the use cases, suggesting more differences of opinion and less fixed views on these issues. However, our findings clarify some points: It was expected that the implementation of RMT would require greater amounts of staff time and financial resources than the status quo. Evidence of cost-effectiveness was considered imperative. The RMT data were considered valuable for reducing uncertainty and moving beyond subjective measures. It has been suggested that RMT could offer benefits under conditions other than the 3 under consideration in RADAR-CNS, eg, bipolar disorder. There were mixed views on how frequently the data should be collected. Passive data were considered more useful than actively collected data because they required less input from the patient, who may forget to complete questionnaires and because passive data were considered to be less subjective.

Country-specific comments highlighted the difference between countries with NHSs (the United Kingdom, the Netherlands, and Spain) compared with countries with insurer-based health care systems (Germany). It was mentioned that RMT could be used to persuade customers to join 1 particular insurer or health service over another. Differences were also highlighted where countries may have a pay-per-patient-visit setup, wherein the use of RMT to check patients are stable and therefore reduce appointments would be financially problematic. The interviews covered only a limited range of countries, and there may be other barriers or facilitators to using RMT in health care systems in other countries. Barriers were raised regarding inequality between settings, patient behavior, health care culture, legal or regulatory issues, and use of off-the-shelf technologies.

Limitations

We would caution against overinterpretation of the consensus scores for MS, where only 3 participants responded in the second round. Unfortunately, we could not recruit more experts in MS during the time available for the study, despite extending the recruitment window and using multiple recruitment methods. This is a shortcoming of many studies seeking the views of MS clinicians, as there are few medics specializing in this condition. However, the combination of item ratings and interview findings relating to MS provides useful insights into how RMT could be used for patient benefits in this condition.

In addition, the aim of this study was to explore the applicability of the RADAR-CNS RMT system in 3 central nervous system disorders, limiting its relevance to other conditions or monitoring platforms. However, the methodology set out here will likely be of interest to others seeking approaches to evaluate the application of novel systems in health care, and the findings will be of interest to those developing a variety of digital interventions for the specific conditions discussed.

Interpretation and Implications for Research and Practice

Our work highlights the potential value of the implementation of RMT for 3 central nervous system disorders, including which applications of RMT would enable something new, enhance existing care, speed up existing processes, or facilitate or make easier the care of patients [14,18]. These results indicate that clinicians would consider RMT patient data to have sufficient value that it would be worth a financial outlay to implement RMT in clinical practice. However, health economic evaluation is required to determine the cost-effectiveness of applying RMT in each of these conditions [48], and the choice to implement is likely to be determined by whether cost-effectiveness is judged to meet local criteria that can vary by country [49].

The findings from this study and our prior work provide an indication of where costs may be incurred if RMT is implemented in a health care service [17]. The costs are likely to include introducing staff roles to manage patient data and provide technical support. Technical support staff could assist patients in setting up and maintaining RMT devices and support clinicians in making the best use of patient RMT data. The extra time for clinical staff to review patient data would be cost incurring, where fewer patients can be seen, although if this results in improvements in care and thereby patient condition, there may be an overall improvement in efficiency. The devices themselves and their maintenance also incur a cost for a health care service wishing to implement them. Although many of the technical devices incorporated within the RADAR-CNS system are consumer-grade technologies that may be owned by patients, our findings suggest that there is a need to provide each patient with devices meeting specific standards of accuracy, adding to costs.

Participants largely suggested that decisions about the implementation of new technologies were top-down and that commissioners and health service leaders would need to be convinced of the benefits of RMT for it to be implemented. In the United Kingdom, commissioners are often involved in redesigning services to incorporate new and beneficial technologies or products [50]. The exception to this was Germany, where doctors who run their own services at a local level are able to work with their budget holders to decide on the implementation of particular technologies. However, it is expected that these technologies should be demonstrably both clinically effective and cost-effective. In the United Kingdom, the National Institute of Health and Care Excellence provides guidance on the evidence required for approval of digital technologies [51]. Similarly, in the European Union, the European Medicines Agency works with groups to develop novel health care technologies to provide scientific advice [52]. Close working with these organizations would facilitate the further development and evaluation of the RADAR-CNS RMT system.

Conclusions

RMT offers new possibilities for the assessment of epilepsy, MS, and depression by enabling new ways of caring for patients, enhancing existing processes, facilitating care, and, in some areas, catalyzing or speeding up existing processes. Our study shows promise for the use of wearable technologies such as Fitbits, wrist-worn epilepsy seizure detection devices, and other wearable accelerometers, together with smartphones, in remote measurement and assessment systems. Priority use cases for the further development and evaluation of RMT in epilepsy according to this study are: more accurate seizure records, automatically analyzing trends, improving triage through review of RMT data, and alerting patients and carers to imminent seizures. In depression, priority use cases are using RMT to detect trends or changes in the condition, monitoring treatment response and using data to inform treatment decisions, and self-management through monitoring and behavioral nudges. Some clinicians recognize the benefits of RMT in the management of MS by enabling the detection of the silent progression of the disease, detecting depression, and enabling the donation of data to biobanks, although clear priorities among these cannot be distinguished from our results.

The implementation of RMT will have different implications in different health service models. Cost-effectiveness studies will be required to understand the economic value of implementing RMT in clinical practice in different regions. Future work could also usefully explore the potential of RMT in other clinical conditions, as well as seeking to understand factors affecting adherence to remote measurement regimes in real-world conditions. Clinicians participating in this study considered passive data to be more reliable than active data, and further work is required to understand whether digital biomarkers based on passive remote measurement data can be used as proxies or replacements for existing measures in these and other clinical conditions. Overall, the Delphi method has been useful for prioritizing use cases and deriving insights into the practical application of RMT in clinical practice.

Acknowledgments

The authors would like to acknowledge the support of all members of the Remote Assessment of Disease and Relapse–Central Nervous System (RADAR-CNS) consortium, who are named on the RADAR-CNS web page [6]. The RADAR-CNS project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant 115902 [53]. This Joint Undertaking receives support from the European Union’s Horizon 2020 research and innovation program and the European Federation of Pharmaceutical Industries and Associations. This communication reflects the views of the RADAR-CNS consortium, and neither the Innovative Medicines Initiative nor the European Union and European Federation of Pharmaceutical Industries and Associations are liable for any use that may be made of the information contained herein. The funding body has not been involved in study design, data collection or analysis, or data interpretation. This research was supported by the National Institute for Health and Care Research MindTech MedTech Co-operative, National Institute for Health and Care Research Nottingham Biomedical Research Centre, and National Institute for Health and Care Research Applied Research Collaboration East Midlands (United Kingdom). The views expressed are those of the authors and not necessarily those of the National Health Service, the National Institute for Health and Care Research, or the Department of Health and Social Care. We are grateful for the contributions of all members of the Patient Advisory Board and its coordinators at Kings College London.

Data Availability

The data from the study may be requested from the authors and will be made available on reasonable request, on approval by all members of the Remote Assessment of Disease and Relapse—Central Nervous System consortium, in accordance with the contractual agreements in place between institutions. A bilateral agreement must be entered to determine the terms and conditions of data sharing.

Authors' Contributions

The survey and interview guide were designed by JAA, MPC, RM, and CH, with input from JW, SL, SET, JD, and SIZ. Recruitment was conducted by JAA, with assistance from the Remote Assessment of Disease and Relapse—Central Nervous System consortium. The members of the expert panel included VSA-A, WPW, JJ, SM, AH, AR, HGR, FD, SB, AS, JLF, AM, AC, JT-B, NB-I, and AZ. The results were analyzed by JAA, MPC, and BG. The manuscript was drafted by JAA. Comments on the manuscript have been provided by all coauthors. All authors have read and approved the manuscript.

Conflicts of Interest

AZ has received travel expenses for scientific meetings from Biogen Idec, Merck Serono, and Novartis; speaking honoraria from Eisai; and a study grant from Novartis. RM has received research funding from Magstim plc, P1Vital Ltd and Electromedical Products Inc. RM has also received fees for serving on a Data Monitoring and Ethics Committee from Novartis.

Multimedia Appendix 1

Illustrations of final use cases.

[DOCX File , 2132 KB - neuro_v2i1e41439_app1.docx]

Multimedia Appendix 2

Delphi study survey instrument.

[DOCX File , 198 KB - neuro_v2i1e41439_app2.docx]

Multimedia Appendix 3

Interview guide.

[DOCX File , 15 KB - neuro_v2i1e41439_app3.docx]

Multimedia Appendix 4

Research Ethics Committee Letter of Approval for Delphi Study.

[DOCX File , 167 KB - neuro_v2i1e41439_app4.docx]

Multimedia Appendix 5
Results of consensus and stability analyses.
[DOCX File , 59 KB - neuro_v2i1e41439_app5.docx]

Multimedia Appendix 6
Illustrative extracts from interview transcripts relating to each use case.
[DOCX File , 18 KB - neuro_v2i1e41439_app6.docx]

References

1. Kang HS, Exworthy M. Wearing the future-wearables to empower users to take greater responsibility for their health and care: scoping review. JMIR Mhealth Uhealth 2022 Jul 13;10(7):e35684 [FREE Full text] [doi: 10.2196/35684] [Medline: 35830222]
2. Yim S, Lui L, Lee Y, Rosenblat J, Ragguett R, Park C, et al. The utility of smartphone-based, ecological momentary assessment for depressive symptoms. J Affect Disord 2020 Sep 01;274:602-609 [FREE Full text] [doi: 10.1016/j.jad.2020.05.116] [Medline: 32663993]
3. Davis MM, Freeman M, Kaye J, Vuckovic N, Buckley DI. A systematic review of clinician and staff views on the acceptability of incorporating remote monitoring technology into primary care. Telemed J E Health 2014 May;20(5):428-438 [FREE Full text] [doi: 10.1089/tmj.2013.0166] [Medline: 24731239]
4. Smuck M, Odonkor CA, Wilt JK, Schmidt N, Swiernik MA. The emerging clinical role of wearables: factors for successful implementation in healthcare. NPJ Digit Med 2021 Mar 10;4(1):45 [FREE Full text] [doi: 10.1038/s41746-021-00418-3] [Medline: 33692479]
5. Caulfield B, Reginatto B, Slevin P. Not all sensors are created equal: a framework for evaluating human performance measurement technologies. NPJ Digit Med 2019 Feb 14;2(1):7 [FREE Full text] [doi: 10.1038/s41746-019-0082-4] [Medline: 31304357]
6. RADAR-CNS: Remote Assessment of Disease and Relapse – Central Nervous System. URL: https://www.radar-cns.org/ [accessed 2023-03-31]
7. Ranjan Y, Rashid Z, Stewart C, Conde P, Begale M, Verbeeck D, Hyve, RADAR-CNS Consortium. RADAR-Base: open source mobile health platform for collecting, monitoring, and analyzing data using sensors, wearables, and mobile devices. JMIR Mhealth Uhealth 2019 Aug 01;7(8):e11734 [FREE Full text] [doi: 10.2196/11734] [Medline: 31373275]
8. Zhang Y, Folarin AA, Sun S, Cummins N, Ranjan Y, Rashid Z, et al. Predicting depressive symptom severity through individuals' nearby bluetooth device count data collected by mobile phones: preliminary longitudinal study. JMIR Mhealth Uhealth 2021 Jul 30;9(7):e29840 [FREE Full text] [doi: 10.2196/29840] [Medline: 34328441]
9. Böttcher S, Bruno E, Manyakov NV, Epitashvili N, Claes K, Glasstetter M, RADAR-CNS Consortium. Detecting tonic-clonic seizures in multimodal biosignal data from wearables: methodology design and validation. JMIR Mhealth Uhealth 2021 Nov 19;9(11):e27674 [FREE Full text] [doi: 10.2196/27674] [Medline: 34806993]
10. Simblett S, Evans J, Greer B, Curtis H, Matcham F, Radaelli M, RADAR-CNS consortium. Engaging across dimensions of diversity: a cross-national perspective on mHealth tools for managing relapsing remitting and progressive multiple sclerosis. Mult Scler Relat Disord 2019 Jul;32:123-132 [FREE Full text] [doi: 10.1016/j.msard.2019.04.020] [Medline: 31125754]
11. Simblett S, Matcham F, Siddi S, Bulgari V, Barattieri di San Pietro C, Hortas López J, RADAR-CNS Consortium. Barriers to and facilitators of engagement with mHealth technology for remote measurement and management of depression: qualitative analysis. JMIR Mhealth Uhealth 2019 Jan 30;7(1):e11325 [FREE Full text] [doi: 10.2196/11325] [Medline: 30698535]
12. Simblett SK, Biondi A, Bruno E, Ballard D, Stoneman A, Lees S, RADAR-CNS consortium. Patients' experience of wearing multimodal sensor devices intended to detect epileptic seizures: a qualitative analysis. Epilepsy Behav 2020 Jan;102:106717. [doi: 10.1016/j.yebeh.2019.106717] [Medline: 31785481]
13. Simblett SK, Bruno E, Siddi S, Matcham F, Giuliano L, López JH, RADAR-CNS Consortium. Patient perspectives on the acceptability of mHealth technology for remote measurement and management of epilepsy: a qualitative analysis. Epilepsy Behav 2019 Aug;97:123-129. [doi: 10.1016/j.yebeh.2019.05.035] [Medline: 31247523]
14. Craven MP, Andrews JA, Lang AR, Simblett SK, Bruce S, Thorpe S, RADAR-CNS Consortium. Informing the development of a digital health platform through universal points of care: qualitative survey study. JMIR Form Res 2020 Nov 26;4(11):e22756 [FREE Full text] [doi: 10.2196/22756] [Medline: 33242009]
15. Andrews JA, Craven MP, Jamnadas-Khoda J, Lang AR, Morriss R, Hollis C, RADAR-CNS Consortium. Health care professionals' views on using remote measurement technology in managing central nervous system disorders: qualitative interview study. J Med Internet Res 2020 Jul 24;22(7):e17414 [FREE Full text] [doi: 10.2196/17414] [Medline: 32706664]
16. Andrews JA, Craven MP, Lang AR, Guo B, Morriss R, Hollis C, RADAR-CNS Consortium. The impact of data from remote measurement technology on the clinical practice of healthcare professionals in depression, epilepsy and multiple

sclerosis: survey. BMC Med Inform Decis Mak 2021 Oct 13;21(1):282 [FREE Full text] [doi: 10.1186/s12911-021-01640-5] [Medline: 34645428]

17. Andrews JA, Craven MP, Lang AR, Guo B, Morriss R, Hollis C, RADAR-CNS Consortium. Making remote measurement technology work in multiple sclerosis, epilepsy and depression: survey of healthcare professionals. BMC Med Inform Decis Mak 2022 May 07;22(1):125 [FREE Full text] [doi: 10.1186/s12911-022-01856-z] [Medline: 35525933]
18. Sharples S, Martin J, Lang A, Craven M, O'Neill S, Barnett J. Medical device design in context: a model of user–device interaction and consequences. Displays 2012 Oct;33(4-5):221-232. [doi: 10.1016/j.displa.2011.12.001]
19. Abdi S, Witte LD, Hawley M. Exploring the potential of emerging technologies to meet the care and support needs of older people: a delphi survey. Geriatrics (Basel) 2021 Feb 13;6(1):19 [FREE Full text] [doi: 10.3390/geriatrics6010019] [Medline: 33668557]
20. Dewa LH, Murray K, Thibaut B, Ramtale SC, Adam S, Darzi A, et al. Identifying research priorities for patient safety in mental health: an international expert Delphi study. BMJ Open 2018 Mar 03;8(3):e021361 [FREE Full text] [doi: 10.1136/bmjopen-2017-021361] [Medline: 29502096]
21. Murphy C, Thorpe L, Trefusis H, Kousoulis A. Unlocking the potential for digital mental health technologies in the UK: a Delphi exercise. BJPsych Open 2020 Jan 28;6(1):e12 [FREE Full text] [doi: 10.1192/bjo.2019.95] [Medline: 31987060]
22. Trevelyan E, Robinson P. Delphi methodology in health research: how to do it? Eur J Integr Med 2015 Aug;7(4):423-428 [FREE Full text] [doi: 10.1016/j.eujim.2015.07.002]
23. Fink A, Kosecoff J, Chassin M, Brook RH. Consensus methods: characteristics and guidelines for use. Am J Public Health 1984 Sep;74(9):979-983. [doi: 10.2105/ajph.74.9.979] [Medline: 6380323]
24. Boulkedid R, Abdoul H, Loustau M, Sibony O, Alberti C. Using and reporting the Delphi method for selecting healthcare quality indicators: a systematic review. PLoS One 2011 Jun 9;6(6):e20476 [FREE Full text] [doi: 10.1371/journal.pone.0020476] [Medline: 21694759]
25. Belton I, MacDonald A, Wright G, Hamlin I. Improving the practical application of the Delphi method in group-based judgment: a six-step prescription for a well-founded and defensible process. Technol Forecast Soc Change 2019 Oct;147:72-82. [doi: 10.1016/j.techfore.2019.07.002]
26. Hasson F, Keeney S, McKenna H. Research guidelines for the Delphi survey technique. J Adv Nurs 2000 Oct;32(4):1008-1015. [Medline: 11095242]
27. Okoli C, Pawlowski S. The Delphi method as a research tool: an example, design considerations and applications. Inform Manage 2004 Dec;42(1):15-29 [FREE Full text] [doi: 10.1016/j.im.2003.11.002]
28. Lange T, Kopkow C, Lützner J, Günther KP, Gravius S, Scharf H, et al. Comparison of different rating scales for the use in Delphi studies: different scales lead to different consensus and show different test-retest reliability. BMC Med Res Methodol 2020 Feb 10;20(1):28 [FREE Full text] [doi: 10.1186/s12874-020-0912-8] [Medline: 32041541]
29. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol 2008 May;61(Pt 1):29-48. [doi: 10.1348/000711006X126600] [Medline: 18482474]
30. Altman D. Practical Statistics for Medical Research. Boca Raton, Florida, United States: CRC Press; 1990.
31. King N. Doing template analysis. In: Qualitative Organizational Research: Core Methods and Current Challenges. Thousand Oaks, California, United States: SAGE Publications; 2012.
32. Gope C. Use of a Smart Watch for seizure/abnormal motion activity monitoring and tracking. Epilepsy Behav 2015 May;46:52-53 [FREE Full text] [doi: 10.1016/j.yebeh.2015.02.049]
33. Gutierrez E, Crone N, Kang J, Carmenate Y, Krauss G. Strategies for non-EEG seizure detection and timing for alerting and interventions with tonic-clonic seizures. Epilepsia 2018 Jun;59 Suppl 1:36-41 [FREE Full text] [doi: 10.1111/epi.14046] [Medline: 29873833]
34. Lockman J, Fisher R, Olson D. Detection of seizure-like movements using a wrist accelerometer. Epilepsy Behav 2011 Apr;20(4):638-641 [FREE Full text] [doi: 10.1016/j.yebeh.2011.01.019] [Medline: 21450533]
35. Brown S, Bird J. Continuing professional development: medico-legal aspects of epilepsy. Seizure 2001 Jan;10(1):68-73; quiz 73 [FREE Full text] [doi: 10.1053/seiz.2001.0518] [Medline: 11181103]
36. Hardisty AR, Peirce SC, Preece A, Bolton CE, Conley EC, Gray WA, et al. Bridging two translation gaps: a new informatics research agenda for telemonitoring of chronic disease. Int J Med Inform 2011 Oct;80(10):734-744 [FREE Full text] [doi: 10.1016/j.ijmedinf.2011.07.002] [Medline: 21890403]
37. Willetts M, Hollowell S, Aslett L, Holmes C, Doherty A. Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96,220 UK Biobank participants. Sci Rep 2018 May 21;8(1):7961 [FREE Full text] [doi: 10.1038/s41598-018-26174-1] [Medline: 29784928]
38. WMA declaration of Taipei on ethical considerations regarding health databases and biobanks. World Medical Association. 2020 Jun 4. URL: https://www.wma.net/policies-post/wma-declaration-of-taipei-on-ethical-considerations-regarding-health-databases-and-biobanks/ [accessed 2022-05-30]
39. Ould Brahim L, Lambert SD, Feeley N, Coumoundouros C, Schaffler J, McCusker J, et al. The effects of self-management interventions on depressive symptoms in adults with chronic physical disease(s) experiencing depressive symptomatology: a systematic review and meta-analysis. BMC Psychiatry 2021 Nov 20;21(1):584 [FREE Full text] [doi: 10.1186/s12888-021-03504-8] [Medline: 34800995]

40. Pung A, Fletcher SL, Gunn JM. Mobile app use by primary care patients to manage their depressive symptoms: qualitative study. J Med Internet Res 2018 Sep 27;20(9):e10035 [FREE Full text] [doi: 10.2196/10035] [Medline: 30262449]
41. BinDhim NF, Shaman AM, Trevena L, Basyouni MH, Pont LG, Alhawassi TM. Depression screening via a smartphone app: cross-country user characteristics and feasibility. J Am Med Inform Assoc 2015 Jan 17;22(1):29-34 [FREE Full text] [doi: 10.1136/amiajnl-2014-002840] [Medline: 25326599]
42. Bush N, Skopp N, Smolenski D, Crumpton R, Fairall J. Behavioral screening measures delivered with a smartphone app: psychometric properties and user preference. J Nerv Ment Dis 2013 Nov;201(11):991-995. [doi: 10.1097/NMD.0000000000000039] [Medline: 24177488]
43. Pelletier J, Rowe M, François N, Bordeleau J, Lupien S. No personalization without participation: on the active contribution of psychiatric patients to the development of a mobile application for mental health. BMC Med Inform Decis Mak 2013 Jul 27;13(1):78 [FREE Full text] [doi: 10.1186/1472-6947-13-78] [Medline: 23890085]
44. Bakker D, Rickard N. Engagement in mobile phone app for self-monitoring of emotional wellbeing predicts changes in mental health: MoodPrism. J Affect Disord 2018 Feb;227:432-442 [FREE Full text] [doi: 10.1016/j.jad.2017.11.016] [Medline: 29154165]
45. Dubad M, Winsper C, Meyer C, Livanou M, Marwaha S. A systematic review of the psychometric properties, usability and clinical impacts of mobile mood-monitoring applications in young people. Psychol Med 2017 Jun 23;48(2):208-228. [doi: 10.1017/s0033291717001659]
46. Pedrelli P, Fedor S, Ghandeharioun A, Howe E, Ionescu DF, Bhathena D, et al. Monitoring changes in depression severity using wearable and mobile sensors. Front Psychiatry 2020 Dec 18;11:584711 [FREE Full text] [doi: 10.3389/fpsyt.2020.584711] [Medline: 33391050]
47. Guidance: Medical device stand-alone software including apps (including IVDMDs). GOV.UK. 2014 Aug 8. URL: https://www.gov.uk/government/publications/medical-devices-software-applications-apps [accessed 2022-06-15]
48. Neumann PJ, Johannesson M. From principle to public policy: using cost-effectiveness analysis. Health Aff (Millwood) 1994 Jan;13(3):206-214. [doi: 10.1377/hlthaff.13.3.206] [Medline: 7927151]
49. EUnetHTA Joint Action 2, Work Package 7, Subgroup 3, Heintz E, Gerber-Grote A, Ghabri S, Hamers FF, Rupel VP, et al. Is there a European view on health economic evaluations? Results from a synopsis of methodological guidelines used in the EUnetHTA partner countries. Pharmacoeconomics 2016 Jan 7;34(1):59-76. [doi: 10.1007/s40273-015-0328-1] [Medline: 26446858]
50. Shaw SE, Smith JA, Porter A, Rosen R, Mays N. The work of commissioning: a multisite case study of healthcare commissioning in England's NHS. BMJ Open 2013 Sep 06;3(9):e003341 [FREE Full text] [doi: 10.1136/bmjopen-2013-003341] [Medline: 24014483]
51. Evidence standards framework for digital health technologies. National Institute for Health and Care Excellence. 2019 Mar. URL: http://nice.org.uk/Media/Default/About/what-we-do/our-programmes/evidence-standards-framework/digital-evidence-standards-framework.pdf [accessed 2022-06-15]
52. Scientific advice and protocol assistance. European Medicines Agency. URL: https://www.ema.europa.eu/en/human-regulatory/research-development/scientific-advice-protocol-assistance [accessed 2022-02-28]
53. Innovative Medicines Initiative. URL: https://www.imi.europa.eu/ [accessed 2023-03-31]

Abbreviations

MS: multiple sclerosis
NHS: National Health Service
PAB: Patient Advisory Board
RADAR-CNS: Remote Assessment of Disease and Relapse–Central Nervous System
RMT: remote measurement technology

Edited by P Kubben; submitted 26.07.22; peer-reviewed by E Le Roux, M Bourgninaud; comments to author 04.01.23; revised version received 14.02.23; accepted 24.02.23; published 25.04.23.

Please cite as:
Andrews JA, Craven MP, Guo B, Weyer J, Lees S, Zormpas SI, Thorpe SE, Devonshire J, San Antonio-Arce V, Whitehouse WP, Julie J, Malins S, Hammers A, Reif A, Ruhe HG, Durbano F, Barlati S, Sen A, Frederiksen JL, Martinelli A, Callen A, Torras-Borrell J, Berrocal-Izquierdo N, Zabalza A, Morriss R, Hollis C, The RADAR-CNS Consortium
Clinical Perspectives on Using Remote Measurement Technology in Assessing Epilepsy, Multiple Sclerosis, and Depression: Delphi Study
JMIR Neurotech 2023;2:e41439
URL: https://neuro.jmir.org/2023/1/e41439
doi:10.2196/41439
PMID:

Original Paper

A Semantic Relatedness Model for the Automatic Cluster Analysis of Phonematic and Semantic Verbal Fluency Tasks Performed by People With Parkinson Disease: Prospective Multicenter Study

Tom Hähnel[1], Dr med; Tim Feige[2], Dipl (Psych); Julia Kunze[1]; Andrea Epler[1]; Anika Frank[1,2], Dr med; Jonas Bendig[1], Dr med; Nils Schnalke[1,2], Dr med; Martin Wolz[3], Dr med; Peter Themann[4], Dr med; Björn Falkenburger[1,2], Dr med

[1]Department of Neurology, University Hospital and Faculty of Medicine Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany

[2]German Center for Neurodegenerative Diseases (DZNE), Dresden, Germany

[3]Department of Neurology and Geriatrics, Elblandklinikum Meißen, Meißen, Germany

[4]Department of Neurology, Klinik am Tharandter Wald, Halsbrücke, Germany

Corresponding Author:
Tom Hähnel, Dr med
Department of Neurology
University Hospital and Faculty of Medicine Carl Gustav Carus
Technische Universität Dresden
Fetscherstraße 74
Dresden, 01307
Germany
Phone: 49 351 458 ext 11880
Fax: 49 351 458 8811880
Email: tom.haehnel@uniklinikum-dresden.de

Abstract

Background: Phonematic and semantic verbal fluency tasks (VFTs) are widely used to capture cognitive deficits in people with neurodegenerative diseases. Counting the total number of words produced within a given time frame constitutes the most commonly used analysis for VFTs. The analysis of semantic and phonematic word clusters can provide additional information about frontal and temporal cognitive functions. Traditionally, clusters in the semantic VFT are identified using fixed word lists, which need to be created manually, lack standardization, and are language specific. Furthermore, it is not possible to identify semantic clusters in the phonematic VFT using this technique.

Objective: The objective of this study was to develop a method for the automated analysis of semantically related word clusters for semantic and phonematic VFTs. Furthermore, we aimed to explore the cognitive domains captured by this analysis for people with Parkinson disease (PD).

Methods: People with PD performed tablet-based semantic (51/85, 60%) and phonematic (69/85, 81%) VFTs. For both tasks, semantic word clusters were determined using a semantic relatedness model based on a neural network trained on the Wikipedia (Wikimedia Foundation) text corpus. The cluster characteristics derived from this model were compared with those derived from traditional evaluation methods of VFTs and a set of neuropsychological parameters.

Results: For the semantic VFT, the cluster characteristics obtained through automated analyses showed good correlations with the cluster characteristics obtained through the traditional method. Cluster characteristics from automated analyses of phonematic and semantic VFTs correlated with the overall cognitive function reported by the Montreal Cognitive Assessment, executive function reported by the Frontal Assessment Battery and the Trail Making Test, and language function reported by the Boston Naming Test.

Conclusions: Our study demonstrated the feasibility of standardized automated cluster analyses of VFTs using semantic relatedness models. These models do not require manually creating and updating categorized word lists and, therefore, can be easily and objectively implemented in different languages, potentially allowing comparison of results across different languages. Furthermore, this method provides information about semantic clusters in phonematic VFTs, which cannot be obtained from traditional methods. Hence, this method could provide easily accessible digital biomarkers for executive and language functions in people with PD.

(JMIR Neurotech 2023;2:e46021) doi:10.2196/46021

KEYWORDS

cognition; executive function; language function; mild cognitive impairment; Parkinson disease; Parkinson disease dementia; semantic clusters; semantic relatedness; verbal fluency tasks

Introduction

Cognitive Deficits in People With Parkinson Disease

Parkinson disease (PD) is the fastest growing neurological disease and the second most common neurodegenerative disease [1]. Cognitive deficits are a frequent problem in people with PD. Approximately 10% to 20% of people with PD show a mild cognitive impairment [2], and approximately 46% of people with PD develop PD dementia (PDD) within 10 years after diagnosis [3]. PDD results in higher health-related costs and a reduced quality of life and, therefore, is of high importance for affected people and health care systems [2]. In addition, PDD constitutes 1 of the 4 milestones that occur, on average, 4 years prior to death and usher the terminal phase of the disease [4].

Cognitive decline in people with PD is characterized by deficits in attention, executive functions, visuospatial functions, memory, and language function [5]. Cognitive functions in people with PD are normally measured using paper-based neuropsychological tests [5,6]. These tests are time-consuming and require experienced raters.

Clusters in the Verbal Fluency Task Transcript

Verbal fluency tasks (VFTs), by contrast, require less time and can report both executive and language functions [5,7]. There are 2 types of VFTs. In the semantic VFT, participants have to produce as many words as possible from 1 specific semantic category within 1 minute. The category “animal” is used most often. In the phonematic VFT, participants have to produce as many words as possible starting with a specific letter within 1 minute. Counting the total number of words produced by the person constitutes the most common analysis for both types of VFTs.

People generally do not produce these words in an evenly spaced temporal sequence but in clusters that often share semantic or phonematic similarities [8-15]. Traditionally, these clusters have been determined manually using 2 methods described by Troyer et al [9]. Words produced in the semantic VFT tend to form clusters of semantically related words. For example, a participant could start with a cluster of pets (dog and cat) and switch to a cluster of animals from Africa (elephant, giraffe, and lion). In the traditional analysis, the identification of these clusters and switches between clusters is based on predefined lists of, for example, animals (eg, a list of African animals and a list of pets). Words produced in the phonematic VFT have traditionally been analyzed using a set of phonematic rules [9]. One of the rules, for example, is to group words starting with the same 2 letters (eg, simple, simulate, and silly).

After the identification of word clusters, characteristics such as the mean cluster size and number of switches between clusters are calculated. Several studies suggested that the size of clusters is associated with language functions [10,14,16], whereas the number of switches is more strongly associated with executive functions [10,17]. Other authors, however, obtained conflicting results [12,17,18], and some studies found the cluster size and number of clusters to be highly correlated, which means that they might not represent independent parameters at all [7,17,19].

Analysis of word clusters in the VFT has been limited by several factors. First, the lists used to analyze the semantic VFT must be created manually, which entails subjectiveness. Second, they can exclusively be used for only 1 language. Third, simple lists may not capture all the individual associations that occur during testing. Fourth, the relatedness of consecutive words can only be classified dichotomously, that is, the word either belongs to the same cluster or not. Thus, it is not possible to quantify the semantic or phonematic “distance” of consecutive words.

In recent years, several approaches have been developed to overcome these disadvantages and allow for an automated, more objective, and quantitative analysis. In general, these approaches identify semantic clusters based on the semantic relatedness of words using mathematical models trained on a large text corpus [20-31].

In some approaches, semantic relatedness is directly estimated from structured knowledge sources such as ontologies or encyclopedias. For example, databases storing hierarchical relations between words have been used to estimate semantic relatedness [20]. Thus, an ontology where *cat* and *dog* are both elements of the parent group *carnivore* leads to a higher semantic relatedness between these animals than between *cat* and *cow*. Other models estimate semantic relatedness based on the link structure between web-based encyclopedia articles [32]. These models make explicit use of knowledge created by humans, but they require complex and highly structured training sets.

A more widely used approach for estimating semantic relatedness is the *latent space analysis*, which is based on the co-occurrence of words in training texts [21-23]. Thereby, 2 words are assumed to be semantically related if they co-occur with similar words in the training texts.

Finally, recent approaches also consider the position of words in relation to each other [24-29]; *Word2Vec*, for instance, uses a sliding window to estimate semantic relatedness by analyzing surrounding words [33]. In this approach, a neural network is trained to predict a word given its surrounding words (continuous bag-of-words method) or to predict the surrounding words given a centered word (skip-gram method). On the basis of this training, semantic relatedness can be estimated from the similarity of the learned context in which these words occur.

Most previous studies analyzed VFTs performed by people with mild cognitive impairment, Alzheimer disease [20,25,28,30], or psychiatric diseases [22,23,27,29]. For people with PD, there is only very limited evidence from 1 study [17]. In this study, Farzanfar et al [17] showed that applying a semantic relatedness model to semantic VFTs performed by people with PD is feasible. Here, executive function correlated with the number of cluster switches but not with the cluster size. Whether a semantic relatedness model can also be applied to phonematic VFTs performed by people with PD has not been explored yet.

Aim of This Study

The aim of this study was to evaluate the feasibility of an automatic cluster analysis based on semantic relatedness in people with PD using voice recordings of semantic and phonematic VFTs. In addition, we aimed to validate the potential of the resulting cluster parameters as digital biomarkers for executive and language functions in people with PD. Finally, we provide our previously trained models for semantic relatedness in different languages to facilitate further research by others [34,35].

Clinical Implications

Our study provides a tool for the automatic identification of semantically related word clusters in VFTs. VFTs are a widely used assessment for capturing cognitive functions in clinical practice and research. Although counting the total number of words produced within 1 minute constitutes the commonly used analysis for VFTs, further analysis of word clusters can provide additional information about executive and language functions. However, traditional methods of cluster identification lack standardization, require considerable manual work, and are language specific, thus limiting their applicability.

We show that cluster identification is possible using a model of semantic relatedness that overcomes these limitations. We prove that this automated approach provides valid digital biomarkers for executive and language functions.

By publishing our source code together with a readily trained model, we will allow other researchers to easily use this approach in their own studies. Thus, this study will allow further trials to capture more information about executive and language impairments without requiring additional time-consuming assessments. Furthermore, this will lead to more reliable and better comparable measurements of executive and language functions as cognitive outcomes in clinical trials. This approach is not limited to people with PD and can also be applied to people with other diseases with impaired executive or language function.

Methods

Recruitment

People with PD were recruited from 3 inpatient and outpatient movement disorder clinics in east Saxony, Germany, between May 2021 and August 2022. Participants with a clinically probable diagnosis of PD according to the current clinical diagnostic criteria [36], sufficient German language skills, and a Montreal Cognitive Assessment (MoCA) score >15 were included in the study. The severity of motor symptoms was assessed using the Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS) subscale III [37]. Levodopa equivalent doses were calculated using the recommended conversion factors [38].

Ethics Approval

This study was approved by the institutional review board of Technische Universität Dresden, Germany (IRB00001473 and BO-EK-149032021). Written informed consent was obtained from all the participants before inclusion in the study.

Phonematic and Semantic VFTs

Phonematic and semantic VFTs were performed without supervision using a self-developed app on an iPad 8 (Apple Inc) running iOS version 14. The semantic VFT was added later to the app, thus leading to fewer recordings for this task. For both VFTs, words with the same word stem, word repetitions, and proper names were not allowed. Instructions for the VFTs outlining these rules were presented to the participants before the test on the tablet. The phonematic VFT was performed first, and the semantic VFT was performed second. After reading the general instructions, the participant was requested to continue to the next page. At this time, the letter "S" (for the phonematic VFT) or the category "animals" (for the semantic VFT) was shown, and the voice was recorded for 60 seconds using the tablet's internal microphone. Speech was detected and transcribed automatically using the Apple Speech Framework (Apple Inc) in iOS 14, which allows local speech processing on the device itself. The transcripts were checked manually by an investigator, and speech recognition errors were corrected. The transcripts were also checked for words that violated any of the aforementioned rules. Transcripts with >25% of violations were excluded from the analysis. In addition, recordings with no words spoken within the first 10 seconds were removed from the analysis because it was deemed unclear whether the person had understood the task.

Speech Recognition Error Rate Calculation

The error rate of the automatically transcribed VFT recordings was measured as normalized Levenshtein distance. Therefore, we counted the numbers of insertions, deletions, and substitutions of words that would be required to change the automatically transcribed word list to the correct word list. This was done using the Python package *pylev* (version 1.4) [39]. Levenshtein distance was normalized by dividing it by the number of words in the correct word list.

List-Based Clustering of the Semantic VFT

Traditional cluster analysis of the semantic VFT is based on fixed thematic lists of animals. These are based on shared features, such as geographical regions (eg, Africa), habitats (eg, water, farm, and pets), or species (eg, birds). To create these categorical lists, we translated the categories and animal lists used in the study by Troyer [11]. All animal words that were not covered by this translation were assigned to existing categories by the judgment of an investigator (TH), and additional categories were created as needed. Animal words were allowed to be part of multiple lists (eg, parrot is part of the lists "pet" and "bird"). The resulting categories and corresponding animal word lists can be found in Multimedia Appendix 1. Clusters were formed of consecutive words that occurred on at least 1 common list. The size of a cluster was calculated as the number of words within the cluster minus 1. The mean cluster size was obtained by also including clusters of single words. The number of switches was defined as the number of clusters, including clusters of single words, minus 1. To maintain consistency with the protocol of Troyer et al [9], rule violations were not excluded in the calculation of these cluster characteristics. The total word count comprised the number of words after removing rule violations.

Rule-Based Clustering of the Phonematic VFT

Traditional cluster analysis of the phonematic VFT is based on fixed phonematic rules. Usually, four rules are used to identify words belonging to a cluster: (1) words starting with the same 2 letters (eg, summer and Sunday); (2) rhyming words (eg, sand and stand); (3) words differing only in 1 vowel sound (eg, sat and seat); and (4) homonyms, if indicated by the test person (eg, some and sum) [9]. Clusters were formed of consecutive words that fulfilled at least 1 common phonematic rule. The size of a cluster was calculated as the number of words within the cluster minus 1. The mean cluster size was obtained by also including clusters of single words. The number of switches was defined as the number of clusters, including clusters of single words, minus 1. To maintain consistency with the protocol of Troyer et al [9], rule violations were not excluded in the calculation of these cluster characteristics. The total word count comprised the number of words after removing rule violations.

Semantic Relatedness Clustering

In addition to the traditional clustering methods, we implemented a semantic relatedness model based on a Word2Vec approach. In brief, this model is based on a neural network that depicts the semantic context of words in texts. To achieve this, the neural network is trained on a large text corpus in which the words surrounding each given word are analyzed. As a result, the semantic context of each word can be represented as a high-dimensional vector. The semantic relatedness (“distance”) of 2 words can be expressed as the cosine between 2 of these vectors.

The Word2Vec model was created and trained using the Python package *gensim* version 4.0.1 [40] with Python 3.9.5 [41]. We used the freely available German Wikipedia (Wikimedia Foundation) corpus for model training [42]. To obtain optimal training results, 3 hyperparameters needed to be set: the dimensionality of the semantic relatedness space, window size for the surrounding words, and training algorithm. In addition, a fixed threshold for semantic relatedness needed to be set to define the word clusters. To find the optimal hyperparameter values, we performed a grid search using the following values: (1) dimensions: 200, 500, and 1000; (2) window size: 4 and 10; and (3) algorithm: continuous bag-of-words and skip-gram. To find the best semantic relatedness threshold, this parameter was varied between 0 and 1 with a step size of 0.01. We prevented overfitting by not training the hyperparameters directly on the word sequences obtained from the participants of this study. Instead, random pairs of animals were drawn from the animal category lists described earlier. We determined the set of hyperparameters that best detected whether both animals in a given pair shared a similar category list. For the comparison with the animal category lists, a semantic relatedness threshold of 0.40 performed the best in the approach described earlier and was used for the semantic VFT. For the phonematic VFT, the semantic relatedness threshold was set to a lower value (0.30) to allow for reasonably sensitive cluster identification. The hyperparameters identified using this approach for both VFTs are summarized in Table 1. Clusters were identified as follows: the words listed by a person were analyzed as a sequence of word pairs (word 1 and word 2, word 2 and word 3, ...). A cluster was defined as a sequence of word pairs in which each sequential word pair had a semantic relatedness greater than the thresholds stated earlier. The size of a cluster was calculated as the number of words within the cluster minus 1. The mean cluster size was obtained by also including clusters of single words. The number of switches was defined as the number of clusters, including clusters of single words, minus 1. To maintain consistency with the protocol of Troyer et al [9], rule violations were not excluded in the calculation of these cluster characteristics. The mean sequential semantic relatedness was determined by calculating the mean of the semantic relatedness of the sequence of all word pairs. The exact implementation of our semantic relatedness method and both traditional methods, including formulas, hyperparameters, and source code, can be obtained from our GitHub page [35]. Furthermore, the provided source code can be easily used to train models in other languages or based on other text corpora. In addition to the German model, we provide models pretrained on the English, Spanish, and French Wikipedia corpora using the same hyperparameters as those stated earlier [34,35].

Table 1. Hyperparameters used for training the semantic relatedness model and identifying semantically related clusters.

Parameters	Semantic VFT[a]	Phonematic VFT
Dimensions of semantic relatedness space	500	500
Word2vec window size	10	10
Word2vec algorithm	Skip-gram	Skip-gram
Semantic relatedness threshold	0.40	0.30

[a] VFT: verbal fluency task.

The listed parameters are the result of hyperparameter optimization, which is described in detail in this section. Different semantic relatedness thresholds were used for the semantic and phonematic VFTs. All other hyperparameters used for model training were identical between both VFTs.

Paper-Based Neuropsychological Tests

The overall cognitive function of all the participants in this study was assessed using the MoCA [43]. In addition, the Frontal Assessment Battery (FAB) and Trail Making Test (TMT) B were used as measures of executive function [44,45]. The Boston Naming Test (BNT) [44] and Mehrfachwahl-Wortschatz-Intelligenztest B (MWT) [46] were performed to measure language function and crystallized intelligence. In the MWT, the participant has to distinguish existing words from fictive words in several word lists. The German versions of all the aforementioned tests were used.

Statistical Analyses

The correlation of the MDS-UPDRS III item "Dysarthria" with the speech recognition error rate was calculated using Spearman rank correlation. All other correlations were calculated as Pearson correlations. For the comparison of the neuropsychological test results, a Mann-Whitney *U* test was performed because the neuropsychological test results were not normally distributed. Statistical tests were performed using Python 3.10.8 [41] with the *scipy* 1.9.3 package [47]. The network graph was created using the *networkx* 2.8.8 Python package [48]. For clarity, the following data are not shown in the network graph: correlations between 2 neuropsychological test results, neuropsychological test results with no correlation *with* the clustering characteristics, and Pearson correlation coefficients between 2 cluster characteristics.

Results

Patient Characteristics and Speech Recognition

In total, 137 recordings were obtained from 85 people with PD, specifically 80 recordings (94% of participants) for the phonematic VFT and 57 (67% of participants) for the semantic VFT. Of the 137 recordings, 6 (4.4%) recordings (phonematic: n=5, 6%; semantic: n=1, 2%) were excluded because the rules of the test were violated, 1 (1%) phonematic recording was removed because not a single word was spoken within the first 10 seconds of the task, and 5 (6%) phonematic and 5 (9%) semantic recordings were excluded because the participants misunderstood the task. This resulted in 69 (out of 80, 86%) phonematic VFT and 51 (out of 80, 89%) semantic VFT transcripts, which were used for traditional and semantical relatedness analyses (Figure 1).

Clinical characteristics of the patients are listed in Table 2. The recordings were transcribed using automatic speech recognition and checked manually for errors. The total error rate, calculated as normalized Levenshtein distance for both VFTs, was 61.8%. In detail, the semantic VFT showed a somewhat lower error rate (52%) than the phonematic VFT (69%), but this difference was not statistically significant (*P*=.15; Figure S1A in Multimedia Appendix 2). Furthermore, the error rate correlated significantly with the extent of dysarthria as reported by the corresponding MDS-UPDRS III item (ρ=0.26, *P*=.005; Figure S1B in Multimedia Appendix 2).

Figure 1. Block diagram of the study design. VFT: verbal fluency task.

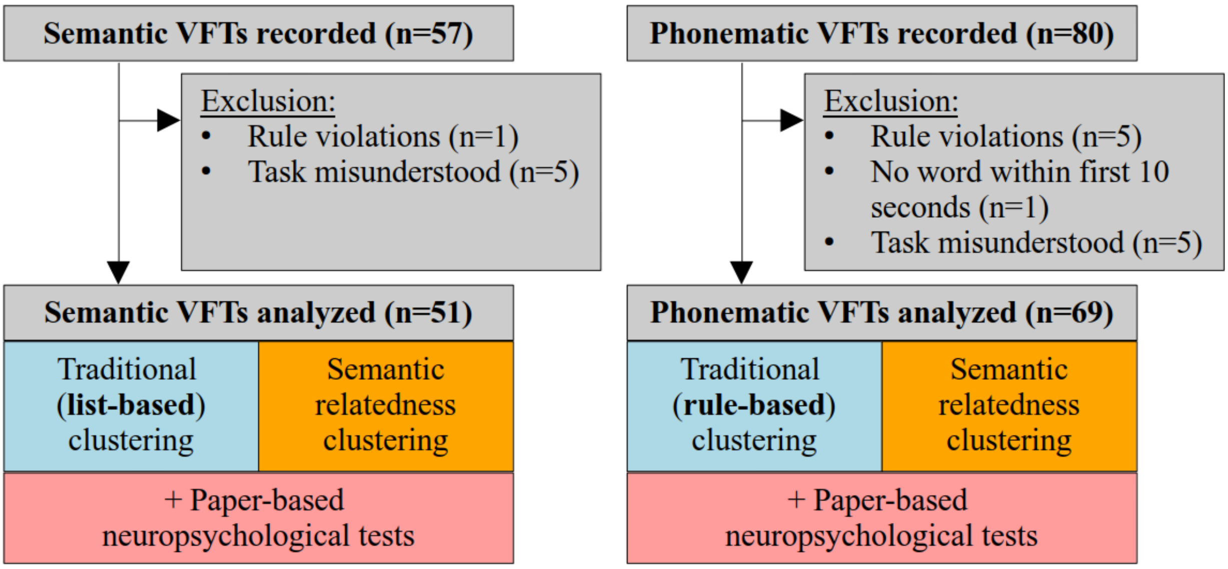

Semantic VFTs recorded (n=57)
Exclusion:
• Rule violations (n=1)
• Task misunderstood (n=5)
Semantic VFTs analyzed (n=51)
Traditional (list-based) clustering
Semantic relatedness clustering
+ Paper-based neuropsychological tests
Phonematic VFTs recorded (n=80)
Exclusion:
• Rule violations (n=5)
• No word within first 10 seconds (n=1)
• Task misunderstood (n=5)
Phonematic VFTs analyzed (n=69)
Traditional (rule-based) clustering
Semantic relatedness clustering
+ Paper-based neuropsychological tests

Table 2. Clinical characteristics of people with Parkinson disease included in the phonematic and semantic analyses.

Parameter	Phonematic VFT[a] (n=69)	Semantic VFT (n=51)
Age (years), mean (SD)	61.2 (13.1)	60.4 (12.3)
Sex, n (%)		
Female	27 (39)	16 (31)
Male	42 (61)	35 (69)
Hoehn and Yahr ON[b], n (%)		
Mild (0-2)	52 (75)	44 (86)
Moderate (2.5-3)	14 (20)	6 (12)
Severe (4-5)	3 (4)	1 (2)
Disease duration (years), mean (SD)	7.4 (5.1)	5.9 (4.0)
Subtype, n (%)		
Tremor dominant	13 (19)	11 (22)
Akinetic rigid	28 (41)	19 (37)
Mixed	28 (41)	21 (41)
LEDD[c], mean (SD)	692 (356)	670 (349)
MDS-UPDRS[d] III, mean (SD)	21 (12)	19 (10)
MoCA[e] score, mean (SD)	26.5 (2.6)	26.7 (3.0)
DBS[f], n (%)		
Yes	7 (10)	4 (8)
No	62 (90)	47 (92)

[a]VFT: verbal fluency task.

[b]People with Parkinson disease can be examined in an ON or OFF state. ON refers to the typical functional state when patients are receiving medication and have a good response.

[c]LEDD: levodopa equivalent daily dose.

[d]MDS-UPDRS: Movement Disorder Society - Sponsored Revision of the Unified Parkinson's Disease Rating Scale.

[e]MoCA: Montreal Cognitive Assessment.

[f]DBS: deep brain stimulation.

Traditional Clusters and Semantically Related Clusters

Cluster characteristics were analyzed for both phonematic and semantic VFTs using (1) traditional clustering methods and (2) the novel semantic relatedness method.

For the phonematic VFT, the traditional cluster analysis is based on phonematic rules, as described in the *Methods* section. In our data, most of the phonematic word pairs (297 word pairs) were identified as clusters because the words shared the same first 2 letters. Only a few clusters were identified by applying the remaining phonematic rules: 6 word pairs were identified as clusters because the words rhymed, 1 word pair was identified as a cluster because the words differed only in 1 vowel, and no homonyms were found. An example of rule-based phonematic clusters is shown in Figure 2A.

In contrast to these phonematic rules, the semantic relatedness method identifies clusters based on a model that can quantify the relatedness of word pairs (Figure 3). The semantic relatedness model was trained on the German Wikipedia corpus. On the basis of this large training data set, this method can identify entirely different clusters from those identified through the rule-based system, for example, the sequence salad, celery, and salami (German: *salat*, *sellerie*, and *salami*) or Zambia and Senegal (German: *Sambia* and *Senegal*), in which words did not share phonematic similarities (Figure 2C). Compared with the traditional rule-based clustering method, in the semantic relatedness method, the clusters had a smaller size, and switches between clusters occurred slightly more often (Table 3). Nonetheless, the number of switches obtained through both methods correlated strongly (r=0.77; P<.001), whereas the mean cluster size did not correlate between the clustering methods (P=.13; Figure 4), potentially because these clusters were construed differently.

For the semantic VFT, the traditional clustering method is based on lists of animals with different themes (eg, farm animals or birds). Words are recognized as a cluster if they are found on at least 1 common list. An example of such list-based clusters is shown in Figure 2B. The clusters in the semantic VFT identified through the list-based method were in general comparable with those identified through the semantic relatedness approach (Figure 2D). This is consistent with the fact that the clusters were generated in a more similar way for the semantic VFT than for the phonematic VFT. However, some additional clusters were detected through the semantic relatedness method. For instance, the cluster bunny and hedgehog may be based on a familiar German fairy tale, and the cluster fox and goose may be based on a common German nursery rhyme.

As for the phonematic VFT, switches between clusters occurred slightly more often with the semantic relatedness method than with the traditional list-based clustering method, and the clusters identified through the semantic relatedness method were slightly smaller than those identified through the traditional list-based clustering method (Table 3). The numbers of switches obtained using the 2 methods correlated significantly (*r*=0.59; *P*<.001), as did the cluster sizes (*r*=0.32; *P*=.02; Figure 4). The strength of the correlation observed in our work is comparable with that observed in a recent study that analyzed traditional and semantic clusters obtained from the semantic VFT performed by people with PD [17].

Figure 2. Phonematic and semantic clustering examples. Cluster examples for the traditional rule-based (A) and list-based (B) technique and the semantic relatedness technique (C and D). Words belonging to the same cluster are displayed in the same color (in blue or orange). For the list-based clustering (B), the common lists for clusters with >1 word is displayed next to each word pair. VFT: verbal fluency task.

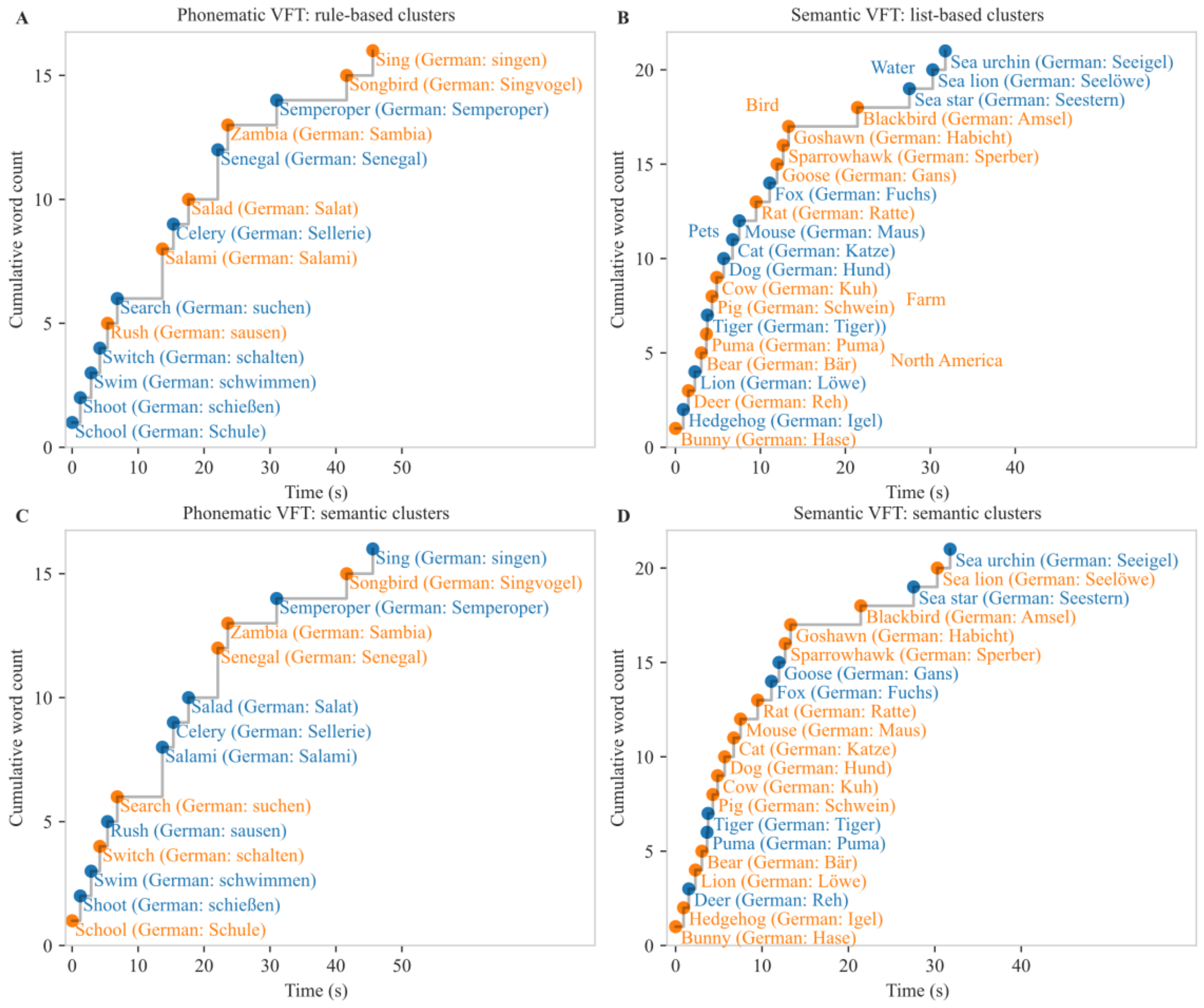


Figure 3. Identification of clusters by calculating the pairwise semantic relatedness. The figure depicts the pairwise semantic relatedness of all sequential word pairs from the phonematic verbal fluency task (VFT) shown in Figure 2C. Words with a pairwise semantic relatedness above the threshold form clusters.

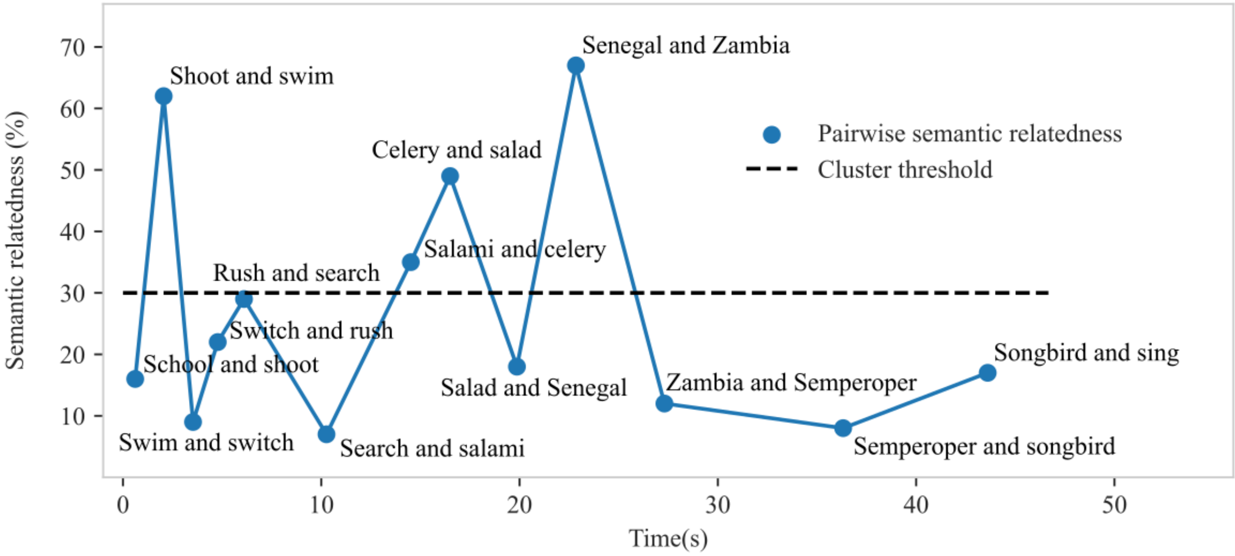


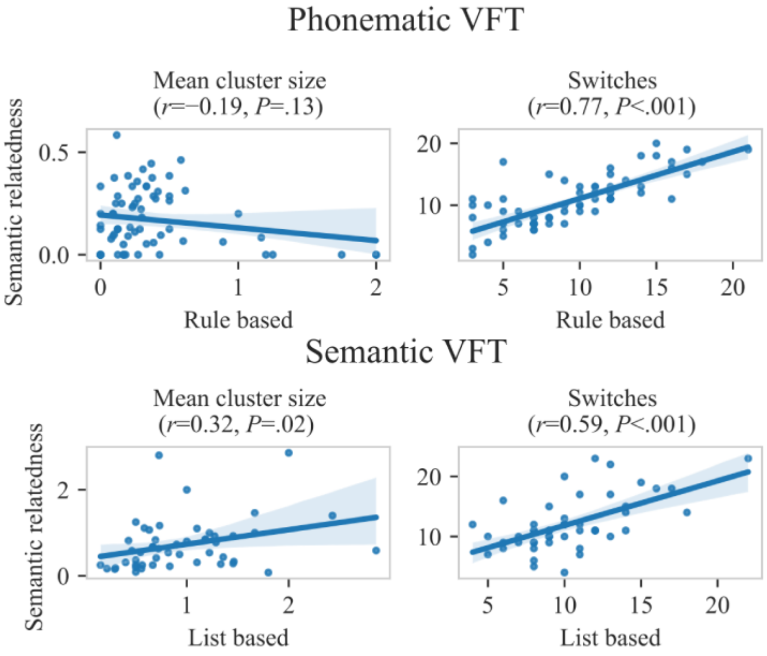
Table 3. Characteristics of the traditional (rule and list based) and semantic relatedness clusters.

	Phonematic VFT[a]		Semantic VFT	
	Rule based	Semantic relatedness	List based	Semantic relatedness
Total word count, mean (SD)	12.3 (4.6)	12.3 (4.6)	19.6 (5.7)	19.6 (5.7)
Mean cluster size, mean (SD)	0.4 (0.4)	0.2 (0.1)	0.9 (0.6)	0.7 (0.6)
Switches, mean (SD)	9.5 (4.1)	10.7 (4.0)	10.3 (3.5)	12.1 (4.4)
Mean sequential semantic relatedness (%), mean (SD)	N/A[b]	18.9 (4.4)	N/A	36.8 (5.3)

[a] VFT: verbal fluency task.

[b] N/A: not applicable.

Figure 4. Correlation of traditional rule- and list-based and semantic relatedness cluster characteristics. Correlations of the semantic-related cluster characteristics (y-axis) with the traditional rule- and list-based cluster characteristics (x-axis) for the phonematic verbal fluency task (VFT; top row) and the semantic VFT (bottom row). Pearson correlation coefficients and corresponding *P* values are shown.



Next, we investigated the effect of sex, age, and disease duration on clustering. Performing sex-specific subgroup analysis revealed no differences between male and female patients except for a slightly higher number of switches in female patients (10.9 vs 8.7; Tables S1 and S2 in Multimedia Appendix 2) in the semantic relatedness analysis of the phonematic VFT. In general, younger patients produced more total words (21.4 vs 17.5) than older patients in the semantic VFT. There were no differences in other cluster characteristics or the phonematic VFT (Tables S3 and S4 in Multimedia Appendix 2). Longer disease duration was associated with slightly more switches between semantically related clusters for the phonematic VFT (11.6 vs 9.6), whereas no difference was observed in other cluster characteristics or the phonematic VFT (Tables S5 and S6 in Multimedia Appendix 2).

Comparing the phonematic and semantic VFTs, we obtained higher total word counts (19.6 vs 12.3; *P*<.001) and larger clusters for the semantic VFT, which is consistent with previous research [49]. The semantic relatedness approach showed a much higher mean sequential relatedness between sequential words for the semantic VFT than for the phonematic VFT (36.8% vs 18.9%; *P*<.001; Table 3).

When we compared the cluster characteristics of the phonematic VFT and semantic VFT for each patient, we observed strong correlations independent of the method used for clustering (Figure S2 in Multimedia Appendix 2). Specifically, we found positive correlations between the total word count and the number of switches. A higher number of switches was associated with a lower mean cluster size, except for the semantic relatedness clusters of the phonematic VFT. Moreover, a higher mean sequential relatedness was associated with larger clusters in the semantic relatedness method, consistent with recent publications on this matter [17,19].

In summary, our semantic relatedness method produced meaningful results consistent with previous work by others. Therefore, we trained the semantic relatedness model on the English, French, and Spanish Wikipedia word corpora (see the *Methods* section) [34,35]. Examples of semantic relatedness clusters in these languages are presented in Figure S3 to S8 in Multimedia Appendix 2.

Correlations With Neuropsychological Tests of Executive and Language Functions

To investigate which cognitive domains are captured by the above-described cluster characteristics, we compared the obtained cluster characteristics with the results of paper-based cognitive tests. In general, the VFT can be seen as a measure of executive and language functions; specifically, the number of switches is considered a measure of executive function, and the cluster size is considered a measure of language function [9,10]. To assess the executive domains in more detail, we used the FAB and TMT B. A lower TMT B score indicates a better result. To assess language function in a standardized fashion, we performed the MWT and BNT. Overall cognition was measured using the MoCA. Correlations between clustering characteristics and neuropsychological tests are shown in Table 4 for the phonematic VFT and in Table 5 for the semantic VFT.

The most important readout of the VFT is the total word count. It correlated with the overall cognitive performance as measured by the MoCA for both the phonematic VFT (*r*=0.38; *P*=.002) and semantic VFT (*r*=0.45; *P*=.001). The MoCA also correlated significantly with cluster characteristics for both types of VFT obtained through the semantic relatedness method. Specifically, a higher MoCA score was associated with a higher mean sequential relatedness in the semantic VFT (*r*=0.28; *P*=.04), a higher mean cluster size (*r*=0.28; *P*=.02) in the phonematic VFT, and a higher number of switches (*r*=0.25; *P*=.04) in the phonematic VFT (Figure 5; Tables 4 and 5). Interestingly, no significant correlations with the MoCA were found for the cluster characteristics obtained through traditional clustering methods (Tables 4 and 5).

With respect to executive functions, the FAB score correlated significantly with the total word count (*r*=0.38; *P*=.005) and number of switches in the phonematic VFT obtained through the traditional rule-based clustering method (*r*=0.28; *P*=.04) and semantic relatedness method (*r*=0.28; *P*=.04). Larger clusters obtained from the semantic relatedness method were associated with a higher FAB for the phonematic VFT (*r*=0.27; *P*=.05). Regarding the semantic VFT, a higher number of switches obtained through the traditional method was associated with a higher FAB score (*r*=0.34; *P*=.04). Taken together, the FAB score thus correlated more strongly with the results of the phonematic VFT than with the results of the semantic VFT. This is consistent with previous findings by other studies [50,51]. Better TMT B results were associated with smaller clusters (*r*=0.63; *P*=.006) and a higher number of switches (*r*=−0.47; *P*=.05) for the semantic VFT as obtained through the semantic relatedness clustering method. The different correlations of FAB and TMT B demonstrate that executive function is not a homogeneous concept and support using different assessment methods. Collectively, these findings demonstrate that the semantic relatedness method can reproduce the association of VFT cluster characteristics with measures of executive function.

With respect to language function, interestingly, we observed no correlations of BNT and MWT with the clustering characteristics of the phonematic VFT. As for the semantic VFT, we found a lower number of switches to be associated with a higher MWT score (*r*=−0.54; *P*=.02) in the traditional clustering method. BNT scores correlated with the mean sequential relatedness of the semantic VFT in the semantic relatedness method (*r*=0.54, *P*=.02). These findings are consistent with the idea that clustering in VFTs is associated with language function [9,10] and demonstrate that the semantic relatedness method can reproduce associations of VFT cluster characteristics with language function.

Table 4. Correlations of the phonematic VFT[a] cluster characteristics with neuropsychological test results.

	Overall cognition	Executive function		Language function	
	MoCA[b]	FAB[c]	TMT B[d]	BNT[e]	MWT[f]
Total VFT word count					
r	*0.38* [g]	*0.38*	−0.03	0.10	0.32
P value	*.002*	*.005*	.89	.63	.12
Mean cluster size (rule based)					
r	0.16	0.05	0.00	0.25	0.12
P value	.20	.74	.99	.23	.58
Switches (rule based)					
r	0.21	*0.28*	−0.06	0.06	0.30
P value	.09	*.04*	.77	.78	.15
Mean cluster size (semantic relatedness)					
r	*0.28*	*0.27*	−0.18	0.07	−0.04
P value	*.02*	*.047*	.40	.72	.86
Switches (semantic relatedness)					
r	*0.25*	*0.28*	0.03	0.15	0.35
P value	*.04*	*.04*	.88	.47	.09
Mean sequential relatedness (semantic relatedness)					
r	0.00	0.14	−0.15	−0.08	−0.12
P value	.97	.32	.47	.71	.58

[a] VFT: verbal fluency task.

[b] MoCA: Montreal Cognitive Assessment.

[c] FAB: Frontal Assessment Battery.

[d] TMT B: Trail Making Test B.

[e] BNT: Boston Naming Test.

[f] MWT: Mehrfachwahl-Wortschatz-Intelligenztest.

[g] Significant values are in italics.

Table 5. Correlations of the semantic VFT[a] cluster characteristics with neuropsychological test results.

	Overall cognition	Executive function		Language function	
	MoCA[b]	FAB[c]	TMT B[d]	BNT[e]	MWT[f]
Total VFT word count					
r	*0.45* [g]	0.24	−0.01	0.37	0.06
P value	*.001*	.16	.96	.13	.80
Mean cluster size (list based)					
r	0.22	−0.16	0.20	0.37	0.34
P value	.12	.34	.42	.13	.17
Switches (list based)					
r	0.15	*0.34*	−0.27	−0.12	*−0.54*
P value	.30	*.04*	.28	.63	*.02*
Mean cluster size (semantic relatedness)					
r	0.19	0.10	*0.63*	0.30	0.22
P value	.18	.57	*.006*	.23	.39
Switches (semantic relatedness)					
r	0.20	0.16	*−0.47*	0.01	−0.23
P value	.16	.34	*.050*	.97	.35
Mean sequential relatedness (semantic relatedness)					
r	*0.28*	0.07	0.45	*0.54*	0.25
P value	*.045*	.69	.06	*.02*	.32

[a] VFT: verbal fluency task.

[b] MoCA: Montreal Cognitive Assessment.

[c] FAB: Frontal Assessment Battery.

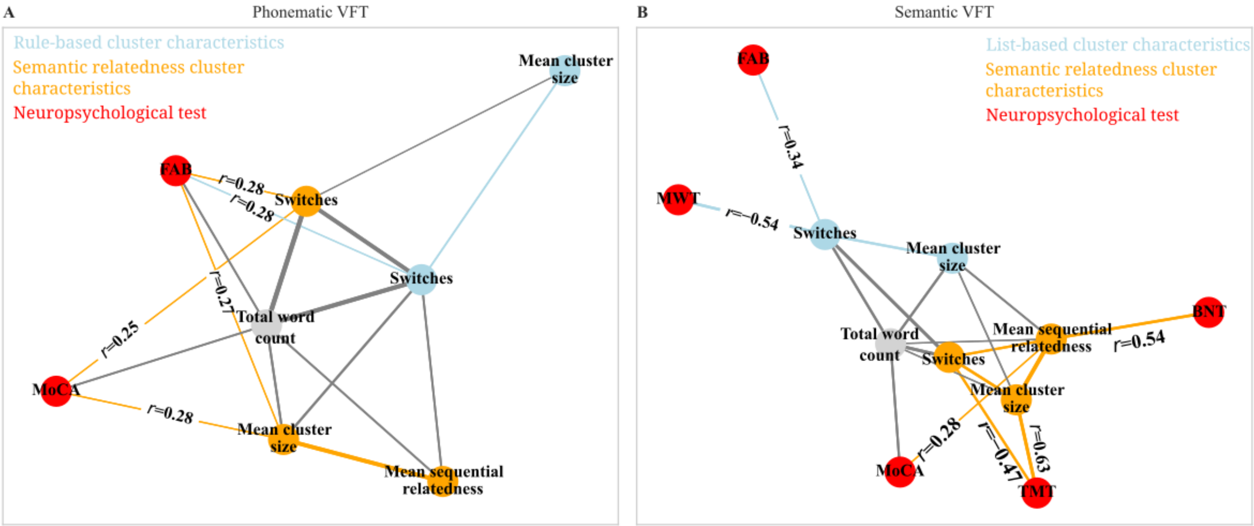
[d] TMT B: Trail Making Test B.

[e] BNT: Boston Naming Test.

[f] MWT: Mehrfachwahl-Wortschatz-Intelligenztest.

[g] Significant values are in italics.

Figure 5. Network graph of Pearson correlations between clustering characteristics and neuropsychological test results. Significant (*P*<.05) correlations for the (A) phonematic and (B) semantic verbal fluency tasks (VFTs) are shown as a network graph. The thickness of the connections and the distance between parameters indicate the magnitude of the correlation (thicker lines and shorter distances indicate stronger correlations). The Pearson correlation coefficients are shown for correlations between clustering characteristics and neuropsychological test results. BNT: Boston Naming Test; FAB: Frontal Assessment Battery; MoCA: Montreal Cognitive Assessment; MWT: Mehrfachwahl-Wortschatz-Intelligenztest; TMT: Trail Making Test.



Discussion

Principal Findings

In this study, we present an automated approach to identify semantically related clusters in VFT transcripts. Speech recordings of semantic and phonematic VFTs were generated by people with PD without supervision using a tablet computer. The obtained cluster characteristics correlated with overall cognitive, executive, and language functions. Moreover, the cluster characteristics provided additional information compared with the total word count alone.

Automatic Speech Recognition

The approach presented here allows for the automated execution and analysis of semantic and phonematic VFTs. By using a standard tablet computer and its integrated microphone, the test can be performed anywhere without the need for an experienced rater, making it a promising digital biomarker for the smartphone-based or tablet-based home monitoring of cognitive functioning.

However, the occurrence of a high percentage of speech recognition errors in automatic transcription for people with PD still limits the feasibility of completely automating this process for participants with dysarthria, consistent with previous results [52]. Advances in speech recognition technologies may help overcome this restriction in the future. The speech recognition error rate may already be lower for other languages and more advanced speech recognition algorithms [52].

Advantages of the Semantic Relatedness Method

In contrast to traditional list-based and rule-based approaches, we used a mathematical model based on the semantic relatedness of words in a large text corpus to identify clusters and calculate the semantic relatedness between words. This demonstrated that the semantic relatedness model is different from and has advantages over the traditional approaches. First, this model allows for an exact and quantitative measurement of the semantic relatedness between 2 words. This is different from traditional methods, which only allow a dichotomous distinction, that is, whether words form a cluster or not.

Second, the estimation of semantic relatedness solely relies on the presence of words in the text corpus that was used for training the model. Thus, the detected clusters do not rely on the subjective decisions of the raters who manually created the word lists. For instance, we consider the clustering of words that occur together in fairy tales or nursery rhymes appropriate. In addition, we demonstrated that the semantic relatedness model can capture more complex relationships between words that go beyond simple lists of characteristics such as geographical regions or simple phonematic rules.

Within our German cohort, we found only a minimal number of rhymes and vowel-only differences and no homonyms for the phonematic VFT. This suggests language-specific differences in rule-based clusters, which limit their usability in an international research context. This limitation does not apply to the semantic relatedness model used in this study. In our view, this method might yield results that can be easily generalized to different languages. The strong correlation of the cluster characteristics of the phonematic VFT as determined by the semantic relatedness model with MoCA and FAB scores further substantiates the validity of this approach (Figure 5; Table 4). In addition, the semantic relatedness method allows for a comparison between the cluster characteristics of the semantic VFT and the cluster characteristics of the phonematic VFT.

Advantages of the Automated Analysis

Using a semantic relatedness model as described above allows for the automation of cluster analysis in VFTs, which results in further advantages. Traditional list-based clustering requires a significant amount of manual work to create the animal lists and update them with new animals listed by the patients. If patients are to be tested again, a different category must be used for the modified test, and the 2 sets of lists might yield differing results. With the semantic relatedness approach, a modified VFT using a different category (eg, fruits instead of animals) can be analyzed using the same method without the need for extensive testing of the new word lists.

The traditional rule-based approach relies on manual work. The detection of homonyms depends on the meaning of the words, and the detection of rhymes depends on the pronunciation and not the spelling of the words. Both features would require more complex approaches, that is, databases identifying the meaning of words and algorithms incorporating the pronunciation of words. Such an automated phonetic analysis has been described, but it resulted in large differences between automated and manual cluster identification [53]. This manual work is not required when using a semantic relatedness model as described here.

As described earlier, the semantic relatedness model shows advantages when applied to different languages. Traditional list-based clustering requires the animal lists to be translated and adapted to the local and cultural circumstances. By contrast, the semantic relatedness model can be easily adapted using a freely available text corpus, such as Wikipedia in a different language. No specific adaptations or list translations need to be performed manually because all language-specific adaptations are already integrated into the text corpus used for training the model.

To further facilitate the use of the semantic relatedness method for VFT analysis, we publish with this manuscript pretrained models for the English, German, French, and Spanish languages, which are based on the corresponding language-specific Wikipedia corpora [34]. In addition, we provide a software to train the model, which will allow other researchers to apply the model to different corpora and new languages [35].

Despite these advantages, our approach also has several limitations. Most of the texts used for training the model are written text and not spoken language, some of which are written in a scientific style. Thus, a corpus of more common texts such as books or interviews may be more appropriate. Although the Wikipedia corpus is available in many languages, not all versions are as extensive as the German and English versions, which could potentially result in less accurate models. In this case, the corpus could be supplemented with books, newspaper articles, or other types of texts.

Correlations of Cluster Characteristics With Neuropsychological Parameters

We used the traditional clustering methods to identify hyperparameters for the automated semantic relatedness method that provided a good correlation with the traditional method for the semantic VFT. For the phonematic VFT, the correlation between the automated semantic relatedness method and the traditional manual method was weaker. This can be explained by the different constructs used for phonematic rules and semantic relatedness.

We observed a correlation between executive functions and cluster characteristics, specifically the number of switches in the semantic and phonematic VFTs, which is consistent with previous data [13,17] and the concept that switching in VFTs reflects executive functioning [9,10]. We were able to replicate these findings for both semantic and phonematic VFTs in the semantic relatedness clustering method. Although semantic relatedness reflects a different construct compared with traditional rule-based clustering, the number of switches between semantically related clusters in the phonematic VFT also showed significant correlations with executive function as reported by the FAB. Regarding the language function, our results do not support the idea of cluster sizes as a marker of language function [9,10]. Similarly conflicting results were also reported by other researchers, and these showed either no correlations of clustering characteristics with language function or correlations of the number of switches with language function [17,18]. The heterogeneity of these results may be caused by the subjectiveness of the animal lists required for traditional clustering and by the correlation of the mean cluster size with the number of switches, as observed in our data and described elsewhere [7,17,19]. By applying the semantic relatedness method, we were able to observe that a higher mean sequential relatedness is associated with a higher BNT score. This shows that the cluster characteristics obtained through the semantic relatedness method yield additional information about language function that cannot be inferred from the total word count or from the traditional clustering method.

Because the phonematic and semantic VFTs were conducted in the same order in all patients, we cannot rule out a negative bias toward the second task caused by fatigue. We assume that the impact of not randomizing the order of the VFTs is limited because the VFT is a very short assessment taking only 1 minute to complete.

Overall, our semantic relatedness clustering method when applied to the semantic VFT yielded results comparable with those published in a recent study [17], highlighting a robust correlation with executive function in people with PD. Our study is the first to investigate semantically related clusters for the phonematic VFT in people with PD. In this study, we showed for the first time that the semantic relatedness method can also be applied to the phonematic VFT in people with PD and that the resulting clustering characteristics are a robust marker of executive function.

Conclusions

In summary, our work demonstrates the feasibility of a standardized cluster analysis of semantic and phonematic VFT transcripts using a semantic relatedness model. This model overcomes numerous disadvantages of traditional clustering methods, allows for the automation of cluster identification, and shows strong correlations with executive functions. The presented automated approach enables a more objective identification of semantic clusters in different languages: going forward, it could help overcome the heterogeneity of previously published studies in this field. Longitudinal trials are required to determine whether cluster characteristics are associated with differences in cognitive decline or disease progression. In the future, this automated semantic relatedness method could provide easily accessible digital biomarkers for executive function in PD.

Acknowledgments

The present data analysis is part of the project ParkinsonNetzwerk Ostsachsen (PANOS). The PANOS project is funded by the German Federal Ministry of Health under the grant 2519FEP009 (Bundesministerium für Gesundheit [Federal Ministry of Health]/Bundesverwaltungsamt [Federal Office of Administration]). In addition, this project is cofinanced by tax funds based on the budget of the Saxon Ministry of Social Affairs and Social Cohesion passed by the Saxon Parliament, by the Sächsische Aufbaubank under the application number 100386587 (Sächsisches Staatsministerium für Soziales und Gesellschaftlichen Zusammenhalt [Saxon State Ministry for Social Affairs and Social Cohesion]/Sächsische Aufbaubank [Saxon Development Bank]), and by the Federal Ministry of Education and Research of Germany (Bundesministerium für Bildung und Forschung) under the application number 16DKWN1113A.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Animal lists used for traditional clustering.
[XLS File (Microsoft Excel File), 16 KB - neuro_v2i1e46021_app1.xls]

Multimedia Appendix 2
Additional figures for speech recognition error rates and correlations between cluster characteristics and subgroup analyses (age, sex, and disease duration). Furthermore, we have presented semantic relatedness cluster examples for the languages English, Spanish, and French based on the corresponding language-specific Wikipedia corpuses.
[DOC File , 2558 KB - neuro_v2i1e46021_app2.doc]

References

1. GBD 2015 Neurological Disorders Collaborator Group. Global, regional, and national burden of neurological disorders during 1990-2015: a systematic analysis for the global burden of disease study 2015. Lancet Neurol 2017 Nov;16(11):877-897 [FREE Full text] [doi: 10.1016/S1474-4422(17)30299-5] [Medline: 28931491]
2. Svenningsson P, Westman E, Ballard C, Aarsland D. Cognitive impairment in patients with Parkinson's disease: diagnosis, biomarkers, and treatment. Lancet Neurol 2012 Aug;11(8):697-707. [doi: 10.1016/S1474-4422(12)70152-7] [Medline: 22814541]
3. Williams-Gray CH, Mason SL, Evans JR, Foltynie T, Brayne C, Robbins TW, et al. The CamPaIGN study of Parkinson's disease: 10-year outlook in an incident population-based cohort. J Neurol Neurosurg Psychiatry 2013 Nov;84(11):1258-1264. [doi: 10.1136/jnnp-2013-305277] [Medline: 23781007]
4. Kempster PA, O'Sullivan SS, Holton JL, Revesz T, Lees AJ. Relationships between age and late progression of Parkinson's disease: a clinico-pathological study. Brain 2010 Jun;133(Pt 6):1755-1762. [doi: 10.1093/brain/awq059] [Medline: 20371510]
5. Emre M, Aarsland D, Brown R, Burn DJ, Duyckaerts C, Mizuno Y, et al. Clinical diagnostic criteria for dementia associated with Parkinson's disease. Mov Disord 2007 Sep 15;22(12):1689-1837. [doi: 10.1002/mds.21507] [Medline: 17542011]
6. Litvan I, Goldman JG, Tröster AI, Schmand BA, Weintraub D, Petersen RC, et al. Diagnostic criteria for mild cognitive impairment in Parkinson’s disease: movement disorder society task force guidelines. Mov Disord 2012 Mar;27(3):349-356 [FREE Full text] [doi: 10.1002/mds.24893] [Medline: 22275317]
7. Koerts J, Meijer HA, Colman KS, Tucha L, Lange KW, Tucha O. What is measured with verbal fluency tests in Parkinson's disease patients at different stages of the disease? J Neural Transm (Vienna) 2013 Mar;120(3):403-411. [doi: 10.1007/s00702-012-0885-9] [Medline: 22922998]
8. Henry JD, Crawford JR. Verbal fluency deficits in Parkinson's disease: a meta-analysis. J Int Neuropsychol Soc 2004 Jul;10(4):608-622. [doi: 10.1017/S1355617704104141] [Medline: 15327739]
9. Troyer AK, Moscovitch M, Winocur G. Clustering and switching as two components of verbal fluency: evidence from younger and older healthy adults. Neuropsychology 1997 Jan;11(1):138-146. [doi: 10.1037//0894-4105.11.1.138] [Medline: 9055277]
10. Troyer AK, Moscovitch M, Winocur G, Leach L, Freedman M. Clustering and switching on verbal fluency tests in Alzheimer's and Parkinson's disease. J Int Neuropsychol Soc 1998 Mar;4(2):137-143. [doi: 10.1017/s1355617798001374] [Medline: 9529823]
11. Troyer AK. Normative data for clustering and switching on verbal fluency tasks. J Clin Exp Neuropsychol 2000 Jun;22(3):370-378. [doi: 10.1076/1380-3395(200006)22:3;1-V;FT370] [Medline: 10855044]

12. Raoux N, Amieva H, Le Goff M, Auriacombe S, Carcaillon L, Letenneur L, et al. Clustering and switching processes in semantic verbal fluency in the course of Alzheimer's disease subjects: results from the PAQUID longitudinal study. Cortex 2008 Oct;44(9):1188-1196. [doi: 10.1016/j.cortex.2007.08.019] [Medline: 18761132]
13. Galtier I, Nieto A, Lorenzo JN, Barroso J. Mild cognitive impairment in Parkinson's disease: clustering and switching analyses in verbal fluency test. J Int Neuropsychol Soc 2017 Jul;23(6):511-520. [doi: 10.1017/S1355617717000297] [Medline: 28494819]
14. Price SE, Kinsella GJ, Ong B, Storey E, Mullaly E, Phillips M, et al. Semantic verbal fluency strategies in amnestic mild cognitive impairment. Neuropsychology 2012 Jul;26(4):490-497. [doi: 10.1037/a0028567] [Medline: 22746308]
15. Shao Z, Janse E, Visser K, Meyer AS. What do verbal fluency tasks measure? Predictors of verbal fluency performance in older adults. Front Psychol 2014 Jul 22;5:772 [FREE Full text] [doi: 10.3389/fpsyg.2014.00772] [Medline: 25101034]
16. Ober BA, Dronkers NF, Koss E, Delis DC, Friedland RP. Retrieval from semantic memory in Alzheimer-type dementia. J Clin Exp Neuropsychol 1986 Jan;8(1):75-92. [doi: 10.1080/01688638608401298] [Medline: 3944246]
17. Farzanfar D, Statucka M, Cohn M. Automated indices of clustering and switching of semantic verbal fluency in Parkinson's disease. J Int Neuropsychol Soc 2018 Nov;24(10):1047-1056. [doi: 10.1017/S1355617718000759] [Medline: 30282568]
18. Demakis GJ, Mercury MG, Sweet JJ, Rezak M, Eller T, Vergenz S. Qualitative analysis of verbal fluency before and after unilateral pallidotomy. Clin Neuropsychol 2003 Aug;17(3):322-330. [doi: 10.1076/clin.17.3.322.18081] [Medline: 14704883]
19. Tröster AI, Fields JA, Testa JA, Paul RH, Blanco CR, Hames KA, et al. Cortical and subcortical influences on clustering and switching in the performance of verbal fluency tasks. Neuropsychologia 1998 Apr;36(4):295-304. [doi: 10.1016/s0028-3932(97)00153-x] [Medline: 9665640]
20. Pakhomov SV, Hemmy LS, Lim KO. Automated semantic indices related to cognitive function and rate of cognitive decline. Neuropsychologia 2012 Jul;50(9):2165-2175 [FREE Full text] [doi: 10.1016/j.neuropsychologia.2012.05.016] [Medline: 22659109]
21. Pakhomov SV, Hemmy LS. A computational linguistic measure of clustering behavior on semantic verbal fluency task predicts risk of future dementia in the nun study. Cortex 2014 Jun;55:97-106 [FREE Full text] [doi: 10.1016/j.cortex.2013.05.009] [Medline: 23845236]
22. Nicodemus KK, Elvevåg B, Foltz PW, Rosenstein M, Diaz-Asper C, Weinberger DR. Category fluency, latent semantic analysis and schizophrenia: a candidate gene approach. Cortex 2014 Jun;55:182-191 [FREE Full text] [doi: 10.1016/j.cortex.2013.12.004] [Medline: 24447899]
23. Holshausen K, Harvey PD, Elvevåg B, Foltz PW, Bowie CR. Latent semantic variables are associated with formal thought disorder and adaptive behavior in older inpatients with schizophrenia. Cortex 2014 Jun;55:88-96 [FREE Full text] [doi: 10.1016/j.cortex.2013.02.006] [Medline: 23510635]
24. Hills TT, Todd PM, Jones MN. Foraging in semantic fields: how we search through memory. Top Cogn Sci 2015 Jul;7(3):513-534 [FREE Full text] [doi: 10.1111/tops.12151] [Medline: 26097107]
25. König A, Linz N, Tröger J, Wolters M, Alexandersson J, Robert P. Fully automatic speech-based analysis of the semantic verbal fluency task. Dement Geriatr Cogn Disord 2018;45(3-4):198-209. [doi: 10.1159/000487852] [Medline: 29886493]
26. Marggraf MP, Cohen AS, Davis BJ, DeCrescenzo P, Bair N, Minor KS. Semantic coherence in psychometric schizotypy: an investigation using latent semantic analysis. Psychiatry Res 2018 Jan;259:63-67. [doi: 10.1016/j.psychres.2017.09.078] [Medline: 29028526]
27. Holmlund TB, Cheng J, Foltz PW, Cohen AS, Elvevåg B. Updating verbal fluency analysis for the 21st century: applications for psychiatry. Psychiatry Res 2019 Mar;273:767-769 [FREE Full text] [doi: 10.1016/j.psychres.2019.02.014] [Medline: 31207864]
28. Tröger J, Linz N, König A, Robert P, Alexandersson J, Peter J, et al. Exploitation vs. exploration-computational temporal and semantic analysis explains semantic verbal fluency impairment in Alzheimer's disease. Neuropsychologia 2019 Aug;131:53-61. [doi: 10.1016/j.neuropsychologia.2019.05.007] [Medline: 31121184]
29. Lundin NB, Todd PM, Jones MN, Avery JE, O'Donnell BF, Hetrick WP. Semantic search in psychosis: modeling local exploitation and global exploration. Schizophr Bull Open 2020 Jan;1(1):sgaa011 [FREE Full text] [doi: 10.1093/schizbullopen/sgaa011] [Medline: 32803160]
30. Chen L, Asgari M, Gale R, Wild K, Dodge H, Kaye J. Improving the assessment of mild cognitive impairment in advanced age with a novel multi-feature automated speech and language analysis of verbal fluency. Front Psychol 2020 Apr 09;11:535 [FREE Full text] [doi: 10.3389/fpsyg.2020.00535] [Medline: 32328008]
31. Taler V, Johns BT, Jones MN. A large-scale semantic analysis of verbal fluency across the aging spectrum: data from the Canadian longitudinal study on aging. J Gerontol B Psychol Sci Soc Sci 2020 Oct 16;75(9):e221-e230 [FREE Full text] [doi: 10.1093/geronb/gbz003] [Medline: 30624721]
32. Kim N, Kim JH, Wolters MK, MacPherson SE, Park JC. Automatic scoring of semantic fluency. Front Psychol 2019 May 16;10:1020 [FREE Full text] [doi: 10.3389/fpsyg.2019.01020] [Medline: 31156496]
33. Bhatta J, Shrestha D, Nepal S, Pandey S, Koirala S. Efficient estimation of word representations in vector space. arXiv. Preprint posted online January 16, 2013 2020 Mar 31 [FREE Full text] [doi: 10.3126/jiee.v3i1.34327]
34. Hähnel T. Multilanguage semantic relatedness models for verbal fluency tasks. Zenodo. 2022. URL: https://zenodo.org/record/7429321 [accessed 2023-07-15]

35. Hähnel T. VFTModels: verbal fluency task analysis tool using traditional rule-based, list-based and a novel semantic relatedness method. GitHub. 2022. URL: https://github.com/t-haehnel/VFTModels [accessed 2023-07-15]
36. Postuma RB, Berg D, Stern M, Poewe W, Olanow CW, Oertel W, et al. MDS clinical diagnostic criteria for Parkinson's disease. Mov Disord 2015 Oct;30(12):1591-1601. [doi: 10.1002/mds.26424] [Medline: 26474316]
37. Goetz CG, Tilley BC, Shaftman SR, Stebbins GT, Fahn S, Martinez-Martin P, Movement Disorder Society UPDRS Revision Task Force. Movement disorder society-sponsored revision of the unified Parkinson's disease rating scale (MDS-UPDRS): scale presentation and clinimetric testing results. Mov Disord 2008 Nov 15;23(15):2129-2170. [doi: 10.1002/mds.22340] [Medline: 19025984]
38. Schade S, Mollenhauer B, Trenkwalder C. Levodopa equivalent dose conversion factors: an updated proposal including Opicapone and Safinamide. Mov Disord Clin Pract 2020 Mar 16;7(3):343-345 [FREE Full text] [doi: 10.1002/mdc3.12921] [Medline: 32258239]
39. Lindsley D. pylev 1.4.0. GitHub. 2021. URL: https://github.com/toastdriven/pylev [accessed 2022-08-20]
40. Řehůřek R, Sojka P. Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. 2010 Presented at: LREC '10; May 22, 2010; Valletta, Malta p. 46-50 URL: https://www.bibsonomy.org/bibtex/27c4ac4b3886c4d66ee336f4df6bee742/l.sz
41. Python language reference. Python Software Foundation. URL: https://www.python.org [accessed 2023-07-15]
42. Wikimedia F. Wikipedia Dump German. Wikimedia Foundation. 2021. URL: https://dumps.wikimedia.org/dewiki/20210601/dewiki-20210601-pages-articles.xml.bz2 [accessed 2021-08-01]
43. Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Collin I, et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc 2005 Apr;53(4):695-699. [doi: 10.1111/j.1532-5415.2005.53221.x] [Medline: 15817019]
44. Schmid NS, Ehrensperger MM, Berres M, Beck IR, Monsch AU. The extension of the German CERAD neuropsychological assessment battery with tests assessing subcortical, executive and frontal functions improves accuracy in dementia diagnosis. Dement Geriatr Cogn Dis Extra 2014 Aug 27;4(2):322-334 [FREE Full text] [doi: 10.1159/000357774] [Medline: 25298776]
45. Dubois B, Slachevsky A, Litvan I, Pillon B. The FAB: a frontal assessment battery at bedside. Neurology 2000 Dec 12;55(11):1621-1626. [doi: 10.1212/wnl.55.11.1621] [Medline: 11113214]
46. Lehrl S. Mehrfachwahl-Wortschatz-Intelligenztest: MWT-B. Balingen, Germany: Spitta; 2018.
47. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, SciPy 1.0 Contributors. Author correction: SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020 Mar;17(3):352 [FREE Full text] [doi: 10.1038/s41592-020-0772-5] [Medline: 32094914]
48. Hagberg A, Schult D, Swart P. Exploring network structure, dynamics, and function using NetworkX. In: Proceedings of the 7th Python in Science Conference. 2008 Presented at: SciPy '08; August 19-24, 2008; Pasadena, CA p. 11-15 URL: https://www.osti.gov/biblio/960616
49. Scholtissen B, Dijkstra J, Reithler J, Leentjens AF. Verbal fluency in Parkinson's disease: results of a 2-min fluency test. Acta Neuropsychiatr 2006 Feb;18(1):38-41. [doi: 10.1111/j.0924-2708.2006.00122.x] [Medline: 26991981]
50. Obeso I, Casabona E, Bringas ML, Alvarez L, Jahanshahi M. Semantic and phonemic verbal fluency in Parkinson's disease: influence of clinical and demographic variables. Behav Neurol 2012;25(2):111-118 [FREE Full text] [doi: 10.3233/BEN-2011-0354] [Medline: 22530265]
51. Lima CF, Meireles LP, Fonseca R, Castro SL, Garrett C. The Frontal Assessment Battery (FAB) in Parkinson's disease and correlations with formal measures of executive functioning. J Neurol 2008 Nov;255(11):1756-1761. [doi: 10.1007/s00415-008-0024-6] [Medline: 18821046]
52. Rohlfing ML, Buckley DP, Piraquive J, Stepp CE, Tracy LF. Hey Siri: how effective are common voice recognition systems at recognizing dysphonic voices? Laryngoscope 2021 Jul;131(7):1599-1607 [FREE Full text] [doi: 10.1002/lary.29082] [Medline: 32949415]
53. Ryan J, Pakhomov S, Marino S, Bernick C, Banks S. Computerized analysis of a verbal fluency test. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. 2013 Presented at: ACL '13; August 4-9, 2013; Sofia, Bulgaria p. 884-889 URL: https://aclanthology.org/P13-2153.pdf

Abbreviations

BNT: Boston Naming Test
FAB: Frontal Assessment Battery
MDS-UPDRS: Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale.
MoCA: Montreal Cognitive Assessment
MWT: Mehrfachwahl-Wortschatz-Intelligenztest
PD: Parkinson disease
PDD: Parkinson disease dementia
TMT: Trail Making Test
VFT: verbal fluency task

Edited by P Kubben; submitted 27.01.23; peer-reviewed by T Martino, E Toki; comments to author 20.04.23; revised version received 19.05.23; accepted 08.06.23; published 02.08.23.

Please cite as:
Hähnel T, Feige T, Kunze J, Epler A, Frank A, Bendig J, Schnalke N, Wolz M, Themann P, Falkenburger B
A Semantic Relatedness Model for the Automatic Cluster Analysis of Phonematic and Semantic Verbal Fluency Tasks Performed by People With Parkinson Disease: Prospective Multicenter Study
JMIR Neurotech 2023;2:e46021
URL: https://neuro.jmir.org/2023/1/e46021
doi:10.2196/46021
PMID:

Original Paper

Connect Brain, a Mobile App for Studying Depth Perception in Angiography Visualization: Gamification Study

Andrey Titov[1,2], BCompSc, MCompSc; Simon Drouin[1], BEng, MSc, PhD; Marta Kersten-Oertel[2], BA, BSc, MSc, PhD

[1]Software and Information Technology Engineering Department, École de Technologie Supérieure, Montreal, QC, Canada

[2]Gina Cody School of Computer Science and Engineering, Concordia University, Montreal, QC, Canada

Corresponding Author:
Andrey Titov, BCompSc, MCompSc
Gina Cody School of Computer Science and Engineering
Concordia University
1455 Boul. de Maisonneuve Ouest
Montreal, QC, H3G 1M8
Canada
Phone: 1 514 848 2424
Email: andrey.titov.1@ens.etsmtl.ca

Abstract

Background: One of the bottlenecks of visualization research is the lack of volunteers for studies that evaluate new methods and paradigms. The increased availability of web-based marketplaces, combined with the possibility of implementing volume rendering, a computationally expensive method, on mobile devices, has opened the door for using gamification in the context of medical image visualization studies.

Objective: We aimed to describe a gamified study that we conducted with the goal of comparing several cerebrovascular visualization techniques and to evaluate whether gamification is a valid paradigm for conducting user studies in the domain of medical imaging.

Methods: The study was implemented in the form of a mobile game, *Connect Brain*, which was developed and distributed on both Android (Google LLC) and iOS (Apple Inc) platforms. Connect Brain features 2 minigames: one asks the player to make decisions about the depth of different vessels, and the other asks the player to determine whether 2 vessels are connected.

Results: The gamification paradigm, which allowed us to collect many data samples (5267 and 1810 for the depth comparison and vessel connectivity tasks, respectively) from many participants (N=111), yielded similar results regarding the effectiveness of visualization techniques to those of smaller in-laboratory studies.

Conclusions: The results of our study suggest that the gamification paradigm not only is a viable alternative to traditional in-laboratory user studies but could also present some advantages.

(JMIR Neurotech 2023;2:e45828) doi:10.2196/45828

KEYWORDS

medical image visualization; volume visualization; depth cues; angiography; gamification; mobile games; mobile phone

Introduction

Background

In the field of medical imaging, angiography is used to visualize vascular structures inside the body. This is typically performed by injecting a contrast substance into a patient and imaging the patient via x-ray, magnetic resonance, or computed tomography [1]. For 3D x-ray, magnetic resonance, or computed tomography angiography (CTA), the result is a 3D volumetric representation of the scanned patient’s vascular anatomy. This 3D volume can be visualized using methods such as axis-aligned slicing [2], volume rendering, and surface rendering [3].

Cerebral angiography specifically depicts the blood vessels of the brain. The goal of this type of angiography is to help radiologists and surgeons understand the cerebral vasculature and detect abnormalities such as stenosis, arteriovenous malformations, and aneurisms [4]. However, visualizing angiography data such that they can be spatially well understood presents certain challenges [1,4,5]. First, the cerebral vasculature is complex, with intricate branching and many overlapping vessels, which hinders the understanding of the data in 3D [1,6]. Second, owing to variations in anatomy from patient to patient, surgeons may not always be able to rely on past experience to understand a new data set [1]. Third, depending on the environment (eg, the operating room), not all visualization methods might be suitable for rendering the data. For example, stereoscopic viewing requires specialized equipment (eg, a stereoscopic display or augmented reality glasses), which is not always available. Perspective rendering may also be inconvenient to use when displaying the data, as radiologists and surgeons may want to perform measurements on the angiographic image [4]; therefore, orthographic projection is most commonly used for 3D medical image visualization [1,4].

Motivation

To improve the depth perception and spatial understanding of vascular volumes, numerous perceptually driven vessel visualization methods have been developed [3,4,7-10]. An overview of the most related studies and their results is presented in Table 1. The studies were chosen based on whether they contained algorithms that could be implemented with direct volume rendering (DVR). In addition, we focused exclusively on static visualizations, as in some contexts (such as the rendering of virtual vessels in augmented reality during a surgical intervention), it is not possible to have dynamic transformations. Thus, to limit the number of conditions and achieve more uniformity among the conditions, we focused only on static visualizations.

In all these works, user studies for determining the effectiveness of different visualization techniques were conducted in a laboratory environment under the supervision of a researcher [12]. This type of laboratory study has a number of disadvantages: the lack of diversity between the participants (who are often young college students) [12] and a limited pool of participants or, conversely, a high monetary cost for studies that have many participants [13]. As can be seen in the table, the number of participants per study was typically between 10 and 20. To overcome these issues, alternative user study paradigms such as crowdsourcing and gamification were explored [12].

Although crowdsourcing has previously been used to evaluate medical image visualization techniques [8,9], to the best of our knowledge, gamification has not been previously used for psychophysical experiments that study the effectiveness of medical visualization techniques. In our study, we used the gamification paradigm to collect data on the effectiveness of different perceptually driven vascular volume visualization techniques. Specifically, we developed a mobile app, *Connect Brain*, with 2 different games that we distributed on the web. The app was published on Google Play (Google LLC) [14] and the App Store (Apple Inc) [15]. Using the developed game, we evaluated the possibility of using the gamification paradigm to conduct user studies on medical imaging. Specifically, the developed game had similar research questions and metrics to those in prior laboratory studies (eg, the studies by Kersten-Oertel et al [1], Ropinski et al [4], and Abhari et al [6]) that evaluated the effectiveness of diverse cerebral vessel visualization techniques. We introduced specific gamification elements, such as levels, points, and leaderboards, to engage the participants and made the games available on the App Store [15] and Google Play [14] to reach a wider participant base. This paper is based on chapter 3 of the first author's master's thesis [16].

Table 1. Related works on depth volume rendering vascular visualization techniques.

Study	Visualizations	Participants, n	Trials and sample points	Goals	Metrics
Ropinski et al [4]	Phong, stereo, chroma, pseudochroma, overlaid edges, blended edges, perspective edges; edge shading; DoF[a]; and DoF+pseudochroma	14	50 × 14 = 700	Depth comparison	Correctness, time, and user feedback
Abhari et al [6]	No cue and edge	10	60 × 10 = 600	Connectivity	Correctness, time, and expert feedback
Kersten-Oertel et al [1]	No cue, kinetic, stereo, edge, pseudochroma, and fog+combined cues (for novice experiments only)	2 studies: 13 novices and 6 experts	160 × 13 = 2080 (novice); 6 × 50 = 300 (expert)	Depth comparison	Correctness, time, and user feedback
Drouin et al [7]	Shading, pseudochroma, fog, dynamic shading, dynamic pseudochroma, and dynamic fog	20	80 × 20 = 1600	Depth comparison and targeting or reaching	Correctness, time, pointer-target distance, and user feedback
Kreiser et al [10]	Phong, chroma, pseudochroma, VSS[b] chroma, and VSS pseudochroma	19	150 × 19 = 2850	Depth comparison	Correctness and time
Titov et al [11]	Shading, pseudochroma, fog, dynamic shading, dynamic pseudochroma, and dynamic fog; all cues were visualized with a VR HMD[c]	12	80 × 12 = 960	Depth comparison and targeting or reaching	Correctness, time, pointer-target distance, head movement, and user feedback

[a] DoF: depth of field.

[b] VSS: void space surface.

[c] VR HMD: virtual reality head-mounted display.

Gamification

Gamification is similar to crowdsourcing and shares its advantages [12]. Crowdsourcing is a method of conducting user studies that distributes a given task to a larger network of participants [12]. An example of a platform for crowdsourcing is the Amazon Mechanical Turk (MTurk) [17], which has been used in studies on a variety of topics, such as the perceptual effectiveness of line drawings to depict shapes [18], natural language processing [19], and audio transcription [20]. Crowdsourcing enables a larger study population than traditional methods because the task can be distributed on the web. In addition, the participant pool becomes more diverse because the study is no longer limited to a physical environment (eg, a university laboratory). Finally, crowdsourcing is less time consuming for each individual participant and allows a lower per-participant cost [17]. This model also has some disadvantages; the main disadvantage being low data quality because researchers do not have much control over the unfolding of the experiment and because participants may be motivated only by monetary gain [12,13].

The main difference between gamification and crowdsourcing is that gamification introduces gaming elements to the study [12]. Through gamification, a study is transformed into a game that is fun to play, and the gameplay data are collected and analyzed as the results of the study. The most important advantage of gamification is that users are motivated to perform well, which consequently increases the quality of the collected data compared with crowdsourcing. Further, players are motivated to perform well not because of monetary incentives but because they enjoy playing the game [13]. As gamification scales well with a large number of participants (because players download and play the games on their own devices), these types of studies have an even lower runtime cost than crowdsourcing [13]. However, there are several disadvantages. First, not every study can be transformed into a game that is fun to play. Furthermore, developing and publishing a game requires more time and effort than creating an experimental task. Finally, for success, the researcher should develop interesting game mechanics that follow the rules of game design [13].

The goal of our work is to determine whether the gamification paradigm is a valid approach to performing user studies, specifically in the context of medical imaging.

Methods

Overview

Connect Brain was developed using the Unity engine (Unity Technologies) [21] for the Android and iOS platforms. Before starting to play the game, all players had to provide informed consent for their gameplay data to be collected anonymously and used for research purposes. They could do this by manually checking the corresponding box during the initial profile creation. In addition, an email address was provided in case players had any questions regarding the user study.

A total of 7 different visualizations were implemented in the mobile app: Blinn-Phong shading [22], edge enhancement [23], aerial perspective (also called fog) [5], chromadepth [24], pseudochromadepth [4], and chromadepth and pseudochromadepth versions of void space surfaces (VSSs) [10]. In all visualizations, the medical data set was rendered using real-time DVR. Note that all visualizations are shaded using the Blinn-Phong shading model in addition to the specified method.

Implemented Visualizations

In the following section, we describe the details of the vascular volume visualization techniques (Figure 1) that were implemented in the Connect Brain game.

Blinn-Phong shading [22] is a photorealistic illumination model that describes how a surface reflects light when illuminated by one or multiple light sources. Similar to Drouin et al [7], we used it as the baseline visualization technique. In our implementation, a single-directional light source was used whose direction was parallel to the view direction (Figure 1A). In terms of color, both the volume and the light source were white.

Edge enhancement is used to emphasize the occlusion depth cue, where a viewer determines the relative depth between different objects based on the way they overlap [23]. In vessel visualization, the contours of vessels are emphasized, typically by rendering dark lines around the edges of the vessels [25] (Figure 1E). This cue is especially helpful when the transfer function (TF) produces a translucent result. In this case, the highly contrasted black silhouettes occlude the silhouettes of the vessels that are farther away from the viewer, thus providing a better understanding of the depth ordering of vessels.

Following the work of Drouin and Collins [23], in our implementation, edge enhancement was combined with Blinn-Phong shading. To do this, the volume is rendered using Blinn-Phong shading, and each pixel that forms the silhouette is darkened based on its interpolated normal vector. Pixels with a gradient that is almost perpendicular to the viewer are considered part of the silhouette. Drouin et al [23] described the following formula for calculating the intensity of edge enhancement for a given pixel:

(1)

where α is the intensity of the edge enhancement factor, is the gradient (normal vector) of the surface, is the direction of the ray (from the volume toward the viewer), and *stepMin* and *stepMax* are user-defined parameters.

Aerial perspective (sometimes referred to as fog) is a monocular depth cue caused by the atmosphere and the way in which light scatters. Specifically, the farther the distance between an object and a viewer, the less contrast there is between the object and the background. With this technique, the vessels that are closer to the viewer appear more saturated and more contrasted, whereas farther vessels fade into the background [1,5] (Figure 1D). By comparing the saturation of 2 vessels, it is possible to deduce which one is closer and which one is farther away.

To render a data set with an aerial perspective cue, the pixels representing the color should be correctly blended with the background. Rheingans and Ebert [26] described the following formula for distance-color blending:

$C = (1 – d)\ c_o + d\ c_b$ **(2)**

where *d* is the depth of the volume at the current pixel in the range of {0,1}, c_o is the color of the object, and c_b is the color of the background. Preim et al [5] noted that the relationship between the depth of the projected vessel and saturation of the pixel does not need to be linear but can rather be exponential (by replacing *d* with an exponential function). To ensure the visualization of the entire volume (such that no vessels are blended completely into the background), Kersten et al [27] determined that the best upper bound for *d* was between 0.75 and 0.85. In our implementation, we used the original linear formulation with *d*=0.8.

Chromadepth, a technique developed by Steenblik [28], encodes depth using color. Specifically, the color of the pixels in depth follows the colors of the visible light spectrum, starting from red; progressing through orange, yellow, green, and cyan; and concluding with blue [24]. Thus, for a vascular volume, the closest vessels are red, the farthest vessels are blue, and vessels in between have a color that is linearly interpolated between these values (Figure 1B). Bailey and Clark [24] described the chromadepth TF as a 1D texture containing all colors (from red to blue), where *s* is defined as the sampling parameter. D_1 and D_2 are parameters defined by the viewer such that $D_1 \geq 0$, $D_2 \geq 1$, and $D_1 < D_2$, and for any depth *d* where $d\ ε\ \{0,1\}$, TF is defined as follows:

if $d < D_1$, then the color of the pixel is red, and if $d > D_2$, then the color of the pixel is blue; otherwise, **(3)**

Figure 2A shows the TF used in our implementation for chromadepth as well as a sample volume shaded in this manner.

Pseudochromadepth, which incorporates only 2 colors (red and blue) instead of the full color spectrum, was used by Ropinski et al [4] to deal with the large number of hues presented in a chromadepth image, which can distract the viewer from the understanding of the depth. Red and blue colors are used (Figure 1C) because of the visual phenomena of chromostereopsis [29], which is caused by the light of different colors refracting into different parts of the retina in the eye depending on the wavelength. Chromostereopsis can be used to make red objects appear closer in depth than blue objects.

When using pseudochromadepth for vasculature, the closest vessels are red; the farthest vessels are blue; and for any intermediate depth, the color of the pixel is calculated by interpolating between red and blue. Thus, using the pseudochromadepth depth cue, a depth comparison between 2 shaded objects can be simplified to a simple comparison of the hue, with warmer hues representing closer objects and colder hues representing farther objects. The pseudochromadepth cue was implemented in the same way as chromadepth, with the only difference being that the 1D rainbow-like texture was replaced by one where the color is linearly interpolated between red and blue, as shown in Figure 2B.

VSS, a technique used in vessel visualization, was developed by Kreiser et al [10] (Figures 1F and 1G). Unlike many other vessel visualization techniques that are based on shading the vessels in a certain manner, VSS concentrates on shading the area around the vessels; the background is colored to indicate the relative depth of the surrounding vessels. Therefore, to understand the relative depth of a certain vessel, one must look at the color of the background that surrounds the vessel. The motivation behind VSS is that in more traditional depth rendering methods, there is a lot of unused empty space. Therefore, instead of being limited by the area that vessels occupy on the screen, the entire screen can be used, allowing the vessel pixels to represent any other information that may be deemed necessary.

To determine the color of each pixel, a weighted average of the depths of the surrounding border pixels is calculated. To do this, a rendered image of a vessel structure in the form of a depth map on which the filled pixels (representing the volume) can be distinguished from the empty pixels (representing the background) is required. The Suzuki and Abe [30] border-following algorithm is then executed on the depth map, creating a hierarchy of the borders of the depth map. This hierarchy indicates what border pixels contribute to what part of the background. Subsequently, the interpolated depth for each background pixel is calculated using inverse distance weighting [31]:

(4)

where *Depth* is the calculated depth of the background pixel, p_i is the *i*th border pixel whose depth is used in the weighted average calculation, *N* is the total number of border pixels that affect the depth of p_b, $w\ (p_i)$, is the weight of the border pixel p_i, and $d\ (p_i)$ is the depth of the border pixel p_i.

The weight $w\ (p_i)$ of a border pixel p_i is calculated in the following manner:

(5)

where p_b is the background pixel for which the depth calculation is performed, p_i is the *i*th border pixel whose depth is used in the weighted average calculation, $m\ (p_b, p_i)$ is the magnitude of the vector between the position of the pixel p_b and p_i, and *s* is a user-defined smoothing parameter that results in closer border pixels giving exponentially more weight.

After calculating the depth of every background pixel, a TF is applied to the depths, transforming them into a color. Typically, chromadepth (Figure 1F) and pseudochromadepth (Figure 1G) are used [10]. In addition, VSS implements an approximated version of global illumination in the form of screen space directional occlusion (SSDO) [32]. SSDO darkens some regions of the generated VSS that may be occluded from the light emitted by neighboring parts of the VSS and performs an indirect light bounce. Finally, isolines are generated on the surface of the VSS in the form of black lines to improve the understanding of the generated shape by the VSS.

Owing to the hardware limitations of mobile devices, we used screen space ambient occlusion [33] instead of SSDO, which does not include indirect bounce.

Figure 1. All the implemented vessel visualization techniques: (A) shading (Blinn-Phong), (B) chromadepth, (C) pseudochromadepth, (D) aerial perspective, (E) edge enhancement, (F) void space surface (VSS) chromadepth, and (G) VSS pseudochromadepth.

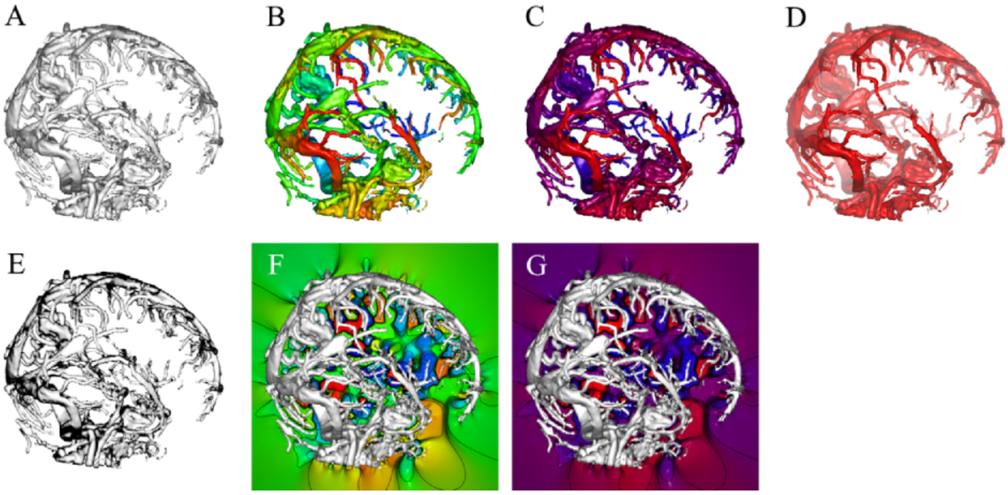
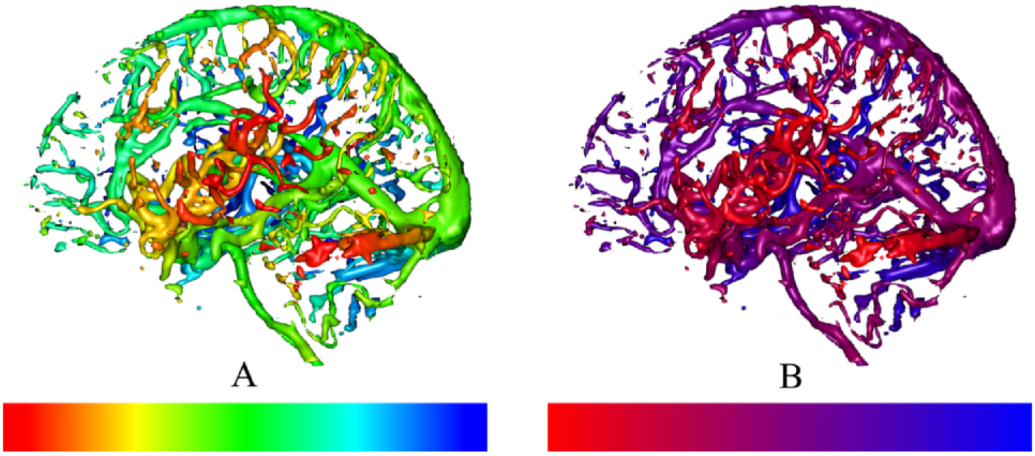


Figure 2. (A) Chromadepth and (B) pseudochromadepth with 1D transfer functions indicating near to far color mapping.



Ethics Approval

The user study was approved by the Natural Sciences and Engineering Research Council of Concordia University (certification 30016074).

DVR on the Mobile Device

To visualize the volumes, the DVR technique described by Drouin et al [23], which is based on a well-known 2-pass rendering algorithm described by Kruger et al [34], was used. This technique describes a real-time ray casting algorithm that consists of 2 rendering passes. In the first pass, the front and back faces of a colored cube representing the bounding box of the volume are rendered into 2 different textures. The red, green, and blue colors encode the start and end positions (as 3D coordinates) of the ray for each pixel. In the second pass, for each pixel, a ray is sent through the volume, and the opacity is accumulated while sampling the volume using trilinear interpolation. The ray stops, and the distance traveled by the ray is recorded into a third texture. Next, a compute shader scans the third texture to determine the smallest and highest nonzero depths of the texture such that the visible interval of the volume inside the 3D texture is known. Finally, in the second pass, the final image of the volume is rendered using the recorded pixel depths, which are adjusted using the minimum and maximum values calculated previously so that the entire range of depth values (from 0 to 1) lies within the visible part of the volume. A TF maps the adjusted depth values to the red, green, blue, and alpha colors for each pixel. This TF is encoded as a 1D texture that is passed to the shader.

As mobile device graphics processing units are typically slower than their desktop equivalents, additional optimizations were made to allow for real-time rendering. First, the ray casting algorithm was simplified so that instead of accumulating opacity at each ray step until full opacity was reached, the ray stopped immediately when the sampled value in the volume reached a given threshold, similar to the early ray termination described by Levoy [35]. Second, the ray casting algorithm was modified to reduce the frequency at which the volume was sampled. To achieve this, the 3D Chamfer distance approach described by Zuiderveld et al [36] was used. This method speeds up ray casting without compromising the quality of the rendered image by determining the distance to the closest nonzero voxel for every voxel and storing it in a 3D texture. This distance corresponds to the number of voxels that must be traversed to create a path in 3D space, assuming a 26-cell cubic neighborhood. Here, a small threshold value was defined to distinguish the *empty* voxels from the nonempty voxels. When performing ray casting, the value from the Chamfer distance 3D texture, which indicates the distance that the ray can safely travel without missing any interesting voxels, is used. Thus, the empty areas of the volume are traversed faster. It should be noted that although the algorithm does not compromise the quality of the volume, it requires more space to store the additional volume.

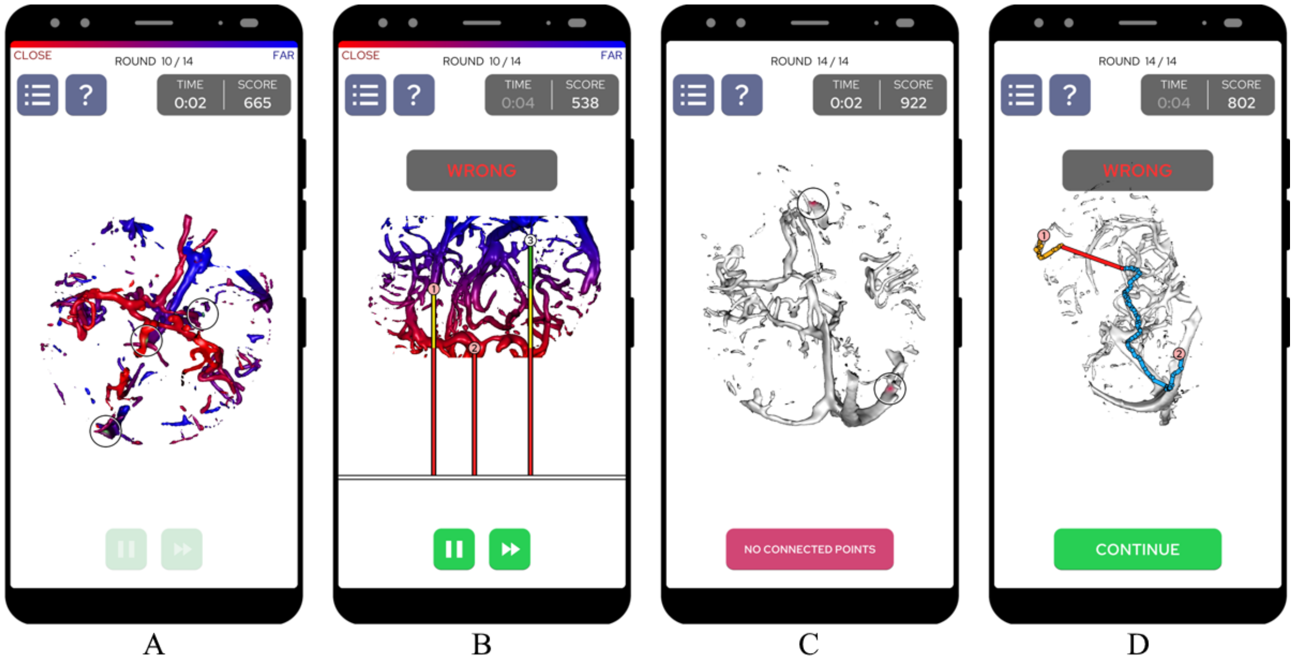
Finally, to save the battery life of the mobile device and have a smoother user interface, when the volume is not being rotated, it is rendered once to a texture and then displayed in future frames. In addition, when the volume is rotated, it is temporarily downscaled during ray casting, the smaller volume is rendered to a temporary frame buffer, and then the image obtained from this frame buffer is upscaled using linear interpolation. The intensity of the downscaling is directly proportional to the speed of the rotation of the volume, making the downsampling less perceptible to the viewer.

Using these optimizations, real-time rendering was achieved on the mobile devices tested for all cues except VSSs. Despite attempts to improve the calculation time of VSS, rendering times of only a few seconds per frame were achieved. As a result, a static version of VSS that cannot be interacted with was used in Connect Brain.

Connect Brain Gameplay

Connect Brain consists of two minigames: (1) the *Near-Far Game*, a game in which players compare the relative depth between the indicated vessels, and (2) the *Blood Circulation Game*, a game in which players must understand the connectivity between different points in the vascular volume (Figure 3, where the phone frame was adapted from Wikimedia [37]; the original uploader of the frame was MDXDave at German Wikipedia, CC BY-SA 3.0 [38]). Both minigames are split into a tutorial level that teaches the player the basics of the minigame and 11 levels that can be played in any order after the completion of the tutorial. Each level is defined by 4 parameters: the CTA data set used, threshold used for early ray termination, depth of the near and far clipping planes, and number of points selected on the volume (≥2). Each level in the game consists of 14 rounds in total, with each round showing a single visualization among those that were implemented. A legend was always present to help the players understand the color encodings for each visualization, and the player could also read the description of the visualization by pressing on a question mark icon. To avoid confusing the player and prevent biases, we decided to use the same visualization technique for 2 consecutive rounds before randomly selecting a new visualization. Videos demonstrating the gameplay of these games can be found in the multimedia appendices (see Multimedia Appendix 1 for the Near-Far Game and Multimedia Appendix 2 for the Blood Circulation Game).

Figure 3. Connect Brain screenshots: (A) gameplay of the Near-Far Game, (B) feedback for the Near-Far Game, (C) gameplay of the Blood Circulation Game, and (D) feedback for the Blood Circulation Game. Phone frame source: adapted from Wikimedia. The original uploader was MDXDave at German Wikipedia, CC BY-SA 3.0.



Near-Far Game

The *Near-Far Game* focuses on understanding the relative depth between vessels. This game is based on the experimental task described and used by Ropinski et al [4], Kersten-Oertel et al [1], and Kreiser et al [10]. The typical experimental task involves participants determining the nearest vessel between 2 selected vessels rendered using a given visualization technique. The Near-Far Game in our app uses the same principle but introduces some gameplay elements to make it more fun for players.

Players are presented with a CTA on which ≥2 points on vessels are indicated. The task of the player is to connect the points from the point closest to them to the point farthest from them using their finger. The points are indicated on the volume using a contrasting color, and to ensure that they are visible, a black and white circle is placed around them (Figure 3A). This circle also indicates the region where the player can touch the screen to select the point. To further help indicate the positions of the points, arrows appear on the screen, indicating the location of the points during the first second of each round. The selected points and view of the CTA are randomly chosen, meaning that the player cannot simply learn the correct answers. This also makes replaying a level more interesting, as the player will always have new data to view and interact with. Although random, a number of rules are applied to choose the points: (1) they are always clearly visible from the player's perspective; (2) they have a small minimum depth difference between them; and (3) there is a minimum *xy* pixel position difference between them, which is equal to the diameter of the black-and-white circle × 1.5 to avoid the overlapping of 2 indicator circles.

By connecting the points in the correct order, the player gains score points; and additional bonus points are provided for doing this quickly. The number of bonus points is calculated by applying a reciprocal function to the round time. However, if the player makes an incorrect decision, the bonus is subtracted from their current score. This gives players an incentive to complete rounds as fast as possible while simultaneously motivating them to make accurate decisions. Further, the score accumulates through the rounds and is saved on a global leaderboard where players can compare their score to others. The score of a player is only visible to other players if it is one of the top 3 scores for the current level, and this setting cannot be changed.

Some levels have rounds in which >2 points are indicated to the player. In these rounds, the player can connect any number of points at once. The goal in this case is to select all the connected points in the ascending depth order, starting from the closest point in terms of depth (similar to the work of Ritter et al [3]). However, if a point with a larger depth is selected before a point with a smaller depth, then the entire selection is considered incorrect, and the player loses the bonus time points.

If all the points are connected in the correct order, the player will receive significantly more points than if they connected each pair of points individually. Thus, selecting multiple points at once is a high-risk, high-reward strategy.

During gameplay, we enable players to rotate the volume as a last resort measure when they get stuck at a certain round. The players can rotate the volume with an offset of up to 45° from the initial position. If *x* and *y* are the rotation in degrees around the *x*- and *y*-axis from the initial position, then the rotation of the volume always follows the formula . To discourage rotation (as we wanted players to understand the data using the given visualization technique), we designed the game such that players lose score points for rotating the volume. The amount of points lost is directly proportional to the rotation of the volume in degrees. This feature was added to reduce the frustration of the player and lower the chance that they will completely abandon the game.

A preliminary in-laboratory study was conducted with 12 participants to test the gameplay aspect of Connect Brain. One of the findings of this preliminary study was that users wanted to know how they were wrong when they made an incorrect decision. Thus, a feedback feature was added; if enabled, at the end of each incorrectly completed round, the volume is rotated by 90° around the x-axis so that the points that are closer to the viewer are positioned on the bottom of this view and the points that are farther are positioned on the top. Vertical lines are then drawn like a ruler to demonstrate the relative depth between the points (Figure 3B).

Blood Circulation Game

The *Blood Circulation Game* focuses on the connectivity between different vessels in the vascular volume. This game is an adaptation of the experiment that was described by Abhari et al [6], in which participants were presented with static 2D images and asked to determine whether a path exists between 2 selected points on the visible vessel structure. We built on this experiment by adding motivating gameplay features to it.

As in the *Near-Far Game*, players are presented with ≥2 points selected on the vascular volume. However, the goal of this game is to determine which points are directly connected, in other words, whether a path exists between the 2 vessels. As each selected point on the 2D image is associated with a specific voxel in 3D, connectivity refers to the path between the 2 voxels inside the 3D volume. When the player finds 2 connected points, they link them using their finger in any order. However, if no 2 points seem to be connected with each other, the player should press the "no connected points" button that is located at the bottom of the screen (Figure 3C).

As described in the first game, the initial rotation of the volume at the beginning of each round and the selection of points are performed randomly. This means that we need to compute at runtime whether 2 voxels are connected with each other within the 3D data set. To achieve this, the A* search algorithm [39], which determines the path (if it exists) between 2 voxels inside a 3D texture, was used. A* is an informed search algorithm that considers both the distance traversed so far and an estimation (heuristic) of the remaining path, allowing it to perform very quickly and find the optimal path in case the heuristic function is admissible (never overestimates the cost to reach the goal). This algorithm requires a priority queue data structure to function, and we chose the Fibonacci heap [40] because of its efficient performance. The threshold used to define the boundaries of the vessels during path finding is the same as that used for ray casting.

The score system works in the same manner as in *Near-Far Game*, with points awarded for correct decisions about whether a path exists and for fast decision response times. The rotation of the volume also works in the same manner, resulting in a loss of points.

The Blood Circulation Game also features a feedback system; if the player decides that 2 points are connected, but in fact they are not, the feedback view shows the minimum distance that separates the 2 independent parts of the vessel structure. Conversely, if the player decides that no points are connected with each other, but some of them are, then this view demonstrates the path between the connected points (Figure 3D).

Once Connect Brain was made available on the Apple App Store and Google Play, we advertised it not only on various social media channels, such as LinkedIn [41], Twitter [42], and Facebook [43], but also through email lists to encourage users to play.

Results

Overview

At the time of our analysis, a total of 111 participants (men: n=68, 61.3%; women: n=39, 35.1%; nonbinary: n=4, 3.6%) had downloaded and played the mobile game. In addition to the 111 participants who played the game, 21 others downloaded it but did not play. Of the 111 participants, 54 (48.6%) played on Android, and the remainder (n=57, 51.4%) played on iOS. Owing to the restriction on the collection of age data on iOS apps, age was collected only from the participants who used the Android version; the age range of these participants was from 14 to 62 (mean 30, SD 11) years. Among the 111 participants, 50 (45%) had experience with medical visualization, 30 (27%) were familiar with angiography, and 36 (32.4%) had experience with vessel visualization techniques. More precisely, of the 111 participants, 26 (23.4%) had experience in all 3 previously listed domains (we refer to them as experts), and 31 (27.9%) had experience in either 1 or 2 domains (we refer to them as semiexperts). All 111 (100%) users participated in the Near-Far Game, completing, on average, 39 (SD 61) rounds, but only 44 (39. 6%) players participated in the Blood Circulation Game, completing, on average, 37 (SD 39) rounds. We hypothesize that the reason why some participants decided to quit the game too early was because they were playing the game in an environment that was not controlled, so they could stop at any moment if they were bored or did not want to continue playing. It is also possible that some players downloaded the game without knowing its purpose and were simply uninterested in playing after downloading. An ANOVA and a post hoc Tukey honest significant difference tests were used to measure and analyze correctness and response time variables. This analysis was performed using the SPSS software (version 26; IBM Corp) [44].

Similar to Kersten-Oertel et al [1] and Lawonn et al [9], for both games, in addition to correctness and response time, we examined the effect of both the distance between the indicated vessels on the screen (*xy* distance) and the distance in depth between the indicated vessels (*z* distance). Both *xy* and *z* distances were equally divided into 2 categories, *near* or *far*, measured in world coordinates. For the *xy* variable, the ranges are defined in the following manner: near (0.162-0.369) and far (0.369-0.951). For the *z* variable, the ranges are defined as follows: near (0.021-0.104) and far (0.104-0.792; note that *z* distances are distributed unequally because the close and far clipping planes in some levels greatly limit the total depth range of the volume, resulting in a larger number of entries with a small depth distance).

Owing to a lack of control over the timing and how the game was played (eg, a person might get interrupted during the game, thus increasing the decision time), we removed all extreme outliers equal to $Q_3 + 3 \times IQR$, where Q_3 represents the value at the third quartile and IQR equal to $Q_3 – Q_1$. In addition, we discarded all data completed during the tutorial levels.

Near-Far Game

A total of 5367 entries were collected for the Near-Far Game. In cases where multiple points (3 or 4) were connected simultaneously, each individual pair of connected points was treated as an individual entry.

Correctness

Correctness was represented by either 1 (correct) or 0 (incorrect) and determined based on whether the connection between points was done in the correct order. The mean correctness and SE for each visualization method are shown in Table 2. A 3-way repeated measures ANOVA was used to examine the main effects as well as the interactions of the visualization method, *xy* distance, and *z* distance, as they relate to correctness. The ANOVA showed that the visualization method had a significant effect on correctness ($F_{6,5339}$=22.404; *P*<.001). A Tukey post hoc test showed that pseudochromadepth (mean 83%, SE 1.5%), aerial perspective (mean 82%, SE 1.5%), and chromadepth (mean 81%, SD 1.5%) allowed for better depth perception than VSS chromadepth (mean 72%, SE 1.6%), VSS pseudochromadepth (mean 72%, SE 1.6%), edge enhancement (mean 66%, SE 1.6%), and shading (mean 65%, SE 1.6%). Although both VSS versions performed better than shading and edge enhancement, only the difference with shading was found to be statistically substantial according to the Tukey honestly significant difference test.

We found a significant main effect of distance on correctness ($F_{1,5339}$=24.708; *P*<.001). As expected, the near *z* distance (mean 71%, SE 0.9%) resulted in worse correctness compared with the far *z* distance (mean 77%, SE 0.8%). However, we found no main effect of the *xy* distance on correctness ($F_{1,5339}$=1.329; *P*=.25). Moreover, there was no significant 2-way interaction between *xy* distance and visualization method on correctness of depth ordering ($F_{6,5339}$=0.627; *P*=.71), between *z* distance and visualization ($F_{6,5339}$=1.836; *P*=.09), or between the *xy* and *z* distances ($F_{1,5339}$=0.619; *P*=.43). There was also no significant 3-way interaction between the variables ($F_{6,5339}$=0.595; *P*=.74).

Table 2. Mean correctness and decision time for the Near-Far Game, depending on the visualization that was used[a].

	Correctness (%), mean (SE)	Time (s), mean (SE)
Arial perspective	82 (1.5)	4.77 (0.117)
Shading	65 (1.6)	5.29 (0.120)
Chroma	81 (1.5)	5.03 (0.117)
Edges	66 (1.6)	4.98 (0.12)
Pseudochroma	83 (1.5)	4.89 (0.118)
VSS[b] chroma	72 (1.6)	5.58 (0.122)
VSS pseudochroma	72 (1.6)	5.44 (0.122)

[a]Error bars represent the SE.

[b]VSS: void space surface.

Decision Time

The decision time for levels with 2 points corresponds to the interval between the moment when the round starts, T_o, and the moment when the finger of the player reaches the second point, T_2. When >2 indicated vessels (ie, *n*) are connected in the same level, the time for connecting *n– 1* with *n* is calculated as $T_n = T_1 + T_n – T_{n–1}$. Thus, we consider the time taken to touch the first indicated vessel, which we consider the time taken by the player to make decisions about the spatial layout of the vasculature as a whole, plus the time interval to connect the 2 indicated vessels *n – 1* and *n*. The mean decision time and SE for each visualization method is shown in Table 2.

A 3-way repeated measures ANOVA was used to examine the main effects and interactions of visualization methods, *xy* distance, and *z* distance on decision time. The ANOVA showed that the visualization method had a significant effect on response time ($F_{6,5339}$=6.334; *P*<.001). A post hoc Tukey test showed that aerial perspective (mean 4.77, SD 0.117 s) and pseudochromadepth (mean 4.89, SE 0.12 s) resulted in the fastest decision times and performed better than both VSS chromadepth (mean 5.58, SE 0.12 s) and VSS pseudochromadepth (mean 5.44, SE 0.12 s). However, only aerial perspective performed better than shading (mean 5.29, SE 0.12 s), which had the third worst decision time. Chromadepth (mean 5.03, SE 0.12 s) and edge enhancement (mean 4.98, SE 0.12 s) were faster than VSS chromadepth but not VSS pseudochromadepth.

There was a significant main effect of *xy* distance ($F_{1,5339}$=12.630; *P*<.001) on decision time. Far *xy* distances (mean 5.30, SE 0.06 s) resulted in longer decision times than near *xy* distances (mean 4.98, SE 0.06 s). In addition, there was a significant main effect of *z* distance ($F_{1,5339}$=12.924; *P*<.001) on decision time. Far *z* distances (mean 4.98, SE 0.06 s) resulted in a shorter decision time than near *z* distances (mean 5.30, SE 0.07 s).

There was no significant 2-way interaction between the visualization method and the *xy* distance ($F_{6,5339}$=0.476; *P*=.83), the visualization method and the *z* distance ($F_{6,5339}$=1.190; *P*=.31), or the *xy* distance and the *z* distance ($F_{1,5339}$=0.063; *P*=.80). There was no 3-way interaction either ($F_{6,5339}$=1.455; *P*=.19).

Blood Circulation Game

The total number of entries collected for the Blood Circulation Game was 1810. A 3-way repeated measures ANOVA was used to examine the main effects as well as the interactions of visualization method, *xy*-distance, and *z*-distance, as they relate to correctness and response time for the Blood Circulation Game.

Correctness

Correctness in the Blood Circulation Game corresponds to whether the player correctly identified the indicated vessels as connected (Table 3). The ANOVA showed that there was no main effect of visualization technique ($F_{6,1782}$=1.383; *P*=.22), *xy* distance ($F_{1,1782}$=0.032; *P*=.86), or *z* distance ($F_{1,1782}$=0.004; *P*=.95) on correctness. Furthermore, there was no significant 2-way interaction between the visualization method and *xy* distance ($F_{6,1782}$=0.867; *P*=.52), between the visualization method and *z* distance ($F_{6,1782}$=1.406; *P*=.35), or between *xy* distance and *z* distance ($F_{1,1782}$=2.251; *P*=.13). No significant 3-way interaction was found either ($F_{6,1782}$=1.536; *P*=.16).

Table 3. Mean correctness and decision time for the Blood Circulation Game, depending on the visualization that was used[a].

	Correctness (%), mean (SE)	Time (s), mean (SE)
Arial perspective	80 (2.4)	3.46 (0.135)
Shading	80 (2.5)	3.18 (0.137)
Chroma	81 (2.5)	3.4 (0.138)
Edges	84 (2.4)	3.27 (0.136)
Pseudochroma	87 (2.4)	3.11 (0.133)
VSS[b] chroma	81 (2.5)	3.52 (0.138)
VSS pseudochroma	80 (2.5)	3.49 (0.141)

[a] Error bars represent the SE.

[b] VSS: void space surface.

Decision Time

The mean decision time and SE for each visualization method are shown in Table 3. ANOVA showed that there was a significant 2-way interaction between the *xy* and *z* distances on correctness ($F_{1,1782}$=4.583; *P*=.03). The combination of far *xy* and far *z* distances correspondingly resulted in a substantially longer decision time (mean 3.59, SE 0.11 s) than any other combination. There were no significant main effects of visualization method ($F_{6,1782}$=1.441; *P*=.20), *xy* distance ($F_{1,1782}$=1.550; *P*=.21), or *z* distance ($F_{1,1782}$=1.559; *P*=.21) on decision time. No significant 2-way interactions were found for the visualization technique and the *xy* distance ($F_{6,1782}$=1.409; *P*=.21) or for the visualization technique and the *z* distance ($F_{6,1782}$=1.044; *P*=.40). Finally, no 3-way interaction was found either ($F_{6,1782}$=0.708; *P*=.64).

Discussion

In general, we found that our results match those of studies that contain a larger number of participants, which suggests that the gamification paradigm is a viable alternative to conducting studies in the domain of medical imaging and, more precisely, angiography visualization.

Depth Perception and Connectivity

The analysis of the gameplay data showed that aerial perspective, chromadepth, and pseudochromadepth allow for the best relative depth perception. These techniques led to the most correct responses and the quickest times, although only aerial perspective resulted in a faster decision time than shading. For vessel connectivity, no cue performed substantially better than the others.

Similar to the study by Kersten-Oertel et al [1], we found that for depth perception, the aerial perspective and pseudochromadepth visualization techniques performed very well in terms of both correctness and decision time. However, unlike Kersten-Oertel et al [1] and Ropinski et al [4], who found pseudochromadepth to be significantly better than chromadepth, we found no difference between the cues. However, this is in line with the results reported by Kreiser et al [10], who found no difference between these 2 cues.

As for the VSS cues, we found that they performed slightly worse compared with the results obtained by Kreiser et al [10]. Although VSS chromadepth and VSS pseudochromadepth resulted in a substantially higher accuracy than shading, both performed worse than the non-VSS versions of chromadepth and pseudochromadepth. In terms of decision response time, we found a similar result to that found by Kreiser et al [10]; VSS had longer times than the directly applied visualization methods. This can be expected owing to the indirect nature of this vessel visualization technique. The correctness results may be explained by the fact that the visualized vasculature is complex, and on small devices (eg, smartphones), there is a limited amount of background, which is needed for VSS. In addition, because of the hardware limitations of mobile devices, VSS was the only cue that was not adjusted in real time when the player was rotating the volume. However, despite this constraint, VSS cues still managed to be more effective than shading, so VSS would be preferable in a context where the color of the vessels cannot be changed.

Edge enhancement was not found to be an effective cue. In terms of depth perception, it resulted in the lowest correct responses, similar to shading. In terms of decision response times, it was substantially better than only VSS chromadepth, and VSS techniques are known to require a significant amount of time to understand. In terms of vessel connectivity understanding, unlike Abhari et al [6], edge enhancement did not improve accuracy or decision time. In fact, this visualization technique had no significant impact on either correctness or response time in terms of understanding vessel connectivity. We posit that this is the case because we tended to demonstrate simpler vessel structures in the Blood Circulation Game, which was achieved by using closer clipping planes to avoid having all vessels connected with each other. The negative side effect of this was that accuracy was high across all visualizations, and decision times were generally similar. These similarities in time could be explained by the fact that players rotated the volume using their finger, but even after removing all entries where players rotated the volume, no effect was observed on the decision time.

In terms of distances between the indicated vessels, as expected, having a far *z* distance between the vessels improves relative depth perception and, surprisingly, decision time, which is different from what was observed by Kersten-Oertel et al [1]. The reason behind shorter decision times at long *z* distances could be that with shorter z distances, the players had to resort to rotating the volume with their finger to understand the depth using motion parallax. Regarding *xy* distance, although it had no effect on accuracy, it did have an effect on the decision response times, with longer *xy* distances resulting in a longer decision time. This may have been caused by the fact that for longer *xy* distances, players had to perform a longer gesture when connecting the indicated vessels. By contrast, in the Blood Circulation Game, where players had to perform a similar gesture, a long *xy* distance resulted in longer decision times only when it was combined with a long *z*-distance, which could mean that the hand gesture does not have a big impact on the decision time. Another reason for this is that players may look back and forth between indicated vessels more often in case of longer distances.

For the Blood Circulation Game, the combination of long *xy* and long *z* distances resulted in the longest decision times. This may have been because in such a combination, the vessels were the farthest apart from each other, so players had to analyze the data set more carefully to draw any conclusion about the connectivity.

Crowdsourcing and Gamification

In this paper, we describe the results of a study that compared the effectiveness of cerebral blood vessel visualization techniques, which was conducted using a mobile game, rather than in a traditional laboratory setting. Similar to previous studies, we found that aerial perspective, chromadepth, and pseudochromadepth allow for the best relative depth perception. In terms of determining the connectivity between 2 vessels, we found that the visualization method did not affect the result.

What differentiates our study from related works is the gamification paradigm that was used to conduct the study. Rather than having participants perform an experiment in a laboratory, we created a mobile game that was distributed using mobile app distribution platforms. Gamification presented multiple advantages compared with traditional in-laboratory user studies. First, it allowed us to have a high number of participants (111 at the time of analysis) with no additional per-participant cost. Second, the participants were also highly diverse, with 39 (35.1%) out of 111 participants identifying as women and 4 (3.6%) identifying as nonbinary. Third, gamification made it easier for us to recruit experts, as 16 (62%) out of 26 experts downloaded the app either from another country or another province of Canada, whereas among the semiexperts, this proportion was 18 (58%) out of 31. Finally, in cases where the study targets a broader range of participants, including nonexperts, gamification incentivizes the nonexperts to join because they might be interested in the game elements rather than the domain of the study. If we look at the average number of rounds completed by experts and semiexperts combined (mean 63, SD 107), it is approximately the same as that for nonexperts (mean 60, SD 83), which indicates that the interests of the 2 groups were approximately the same toward the game. We hypothesize that experts and semiexperts were primarily interested in continuing to play the game because of the domain of study, whereas nonexperts were interested because of the game elements, such as competing for a high score.

Limitations

Gamification also presented some important disadvantages, both during the development of the game and with data collection.

First, transforming the experiment into a game that is fun to play required more development time and additional research to create interesting game mechanics. In our case, the user study could be transformed into a game because it integrated simple visual tasks for both minigames, which were visual comparison (in the Near-Far Game) and path finding (in the Blood Circulation Game). These tasks can both be used for an experiment, but they are also common game principles. However, by themselves, these visual tasks were not interesting enough to make the game fun, so additional game elements had to be added, such as the score system or high-risk, high-reward multiple connection mechanic.

Second, implementing volume rendering such that it allows real-time rendering on mobile devices required additional optimizations of the rendering code. In addition, to ensure that the game worked on different devices and operating systems, graphics processing units, resolutions, and aspect ratios also required additional development. Even though we tested our game on a variety of Android and iOS devices, we still could not guarantee that our game worked perfectly on all hardware configurations, as we received feedback from 1 (0.9%) of the 111 participants that one of the rendering techniques crashed on their device. In addition, we did not have control over the resolution or aspect ratio of the screen, which might have had an impact on performance. However, to achieve at least some consistency, we scaled the volume such that it was proportional to the vertical resolution of the screen.

Third, the lack of a controlled environment may have impacted the collected data. As we could not observe how the game was played, we cannot be sure whether players were motivated to try to do their best. At the same time, we think that adding a competitive element to the study in the form of a leaderboard did indeed motivate most players to perform well, which should have resulted in a higher quality of the collected samples. We also had little control over the credibility of the data that users filled when creating their account and could not create a detailed pretest or posttest questionnaire, which was not possible on iOS owing to privacy concerns and in general could lead to a player abandoning the game before even starting to play.

Conclusions

Despite some of the drawbacks of gamification, using this paradigm allowed this study to collect more data samples than many similar studies [1,4,6,7,10]. Furthermore, it showed that our results were more similar to those of studies with more data samples and participants (2380 for Kersten-Oertel et al [1] and 2850 for Kreiser et al [10]) than those of studies with fewer samples (700 for Ropinski et al [4] and 600 for Abhari et al [6]). These results suggest that gamification is a viable paradigm for conducting user studies in the domain of medical imaging. Moreover, as demonstrated by our number of participants and results, if the game is fun to play and motivates the players to perform well in the study, it may lead to a higher number of participants compared with an in-laboratory user study while still maintaining a high quality of the collected data. Another advantage of web distribution-based paradigms, such as gamification, is that they make it possible to perform user studies or help with surgical education in societal situations where meeting in person is not possible [45]. Such was the case in this study, which was performed during the lockdown caused by the COVID-19 pandemic. Gamification is a promising technique for collecting large data samples; however, it is important to have fun games that users will continue to play. In the future, we could further improve the game by adding sound and music and examine whether these aspects have a positive impact on the time players spend in the game. In addition, we could pay the participants to play our game to determine how having a monetary incentive affects the behavior of the players, as they may enjoy the game more this way [46]. Regarding the study itself, in the future, illustrative techniques could be added to compare an even higher number of visualizations. Some good candidates are the hatching and distance-encoded shadows technique described by Ritter et al [3]; illustrative shadows, supporting lines, and contours technique described by Lawonn et al [9]; and anchors technique described by Lawonn et al [8]. Finally, we could compare our gamified user study to crowdsourcing, such as the EvalViz [47] wizard.

Acknowledgments

This study was supported by the Natural Sciences and Engineering Research Council of Canada (DG-N06722).

Data Availability

All the data used in the user study are available upon request from the corresponding author.

Authors' Contributions

AT contributed to conceptualization, formal analysis, investigation, methodology, validation, visualization, and the writing of the original draft. SD contributed to investigation; methodology; supervision; and the writing, review, editing of the manuscript. MK-O contributed to conceptualization; funding acquisition; investigation; methodology; supervision; visualization; and the writing, review, and editing of the manuscript.

Conflicts of Interest

AT developed the Connect Brain mobile app and its related intellectual property along with Concordia University. The app was developed for the sole purpose of drawing a scientific conclusion about depth perception in angiography visualization. It was distributed free of charge and does not have any in-app advertisements or paid content. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Gameplay video of the Near-Far Game.
[MP4 File (MP4 Video), 75413 KB - neuro_v2i1e45828_app1.mp4]

Multimedia Appendix 2
Gameplay video of the Blood Circulation Game.
[MP4 File (MP4 Video), 90143 KB - neuro_v2i1e45828_app2.mp4]

References

1. Kersten-Oertel M, Chen SJ, Collins DL. An evaluation of depth enhancing perceptual cues for vascular volume visualization in neurosurgery. IEEE Trans Vis Comput Graph 2014 Mar;20(3):391-403. [doi: 10.1109/TVCG.2013.240] [Medline: 24434220]
2. Joshi A, Scheinost D, Vives KP, Spencer DD, Staib LH, Papademetris X. Novel interaction techniques for neurosurgical planning and stereotactic navigation. IEEE Trans Vis Comput Graph 2008;14(6):1587-1594 [FREE Full text] [doi: 10.1109/TVCG.2008.150] [Medline: 18989014]
3. Ritter F, Hansen C, Dicken V, Konrad O, Preim B, Peitgen HO. Real-time illustration of vascular structures. IEEE Trans Vis Comput Graph 2006;12(5):877-884. [doi: 10.1109/TVCG.2006.172] [Medline: 17080812]
4. Ropinski T, Steinicke F, Hinrichs K. Visually supporting depth perception in angiography imaging. In: Proceedings of the 6th International Symposium, SG 2006. 2006 Presented at: 6th International Symposium, SG 2006; July 23-25, 2006; Vancouver, BC. [doi: 10.1007/11795018_9]
5. Preim B, Baer A, Cunningham D, Isenberg T, Ropinski T. A survey of perceptually motivated 3D visualization of medical image data. Comput Graph Forum 2016 Jul 04;35(3):501-525 [FREE Full text] [doi: 10.1111/cgf.12927]
6. Abhari K, Baxter JS, Eagleson R, Peters T, de Ribaupierrec S. Perceptual enhancement of arteriovenous malformation in MRI angiography displays. In: Proceedings of the Medical imaging 2012: Image perception, observer performance, and technology assessment. 2012 Presented at: Medical Imaging 2012: Image Perception, Observer Performance, and Technology Assessment; February 8-9, 2012; San Diego, CA URL: https://iacl.ece.jhu.edu/proceedings/spie2012/DATA/8318_8.PDF [doi: 10.1117/12.911687]
7. Drouin S, DiGiovanni DA, Kersten-Oertel MA, Collins L. Interaction driven enhancement of depth perception in angiographic volumes. IEEE Trans Vis Comput Graph 2018 Dec 06;26(6):2247-2257. [doi: 10.1109/TVCG.2018.2884940] [Medline: 30530366]
8. Lawonn K, Luz M, Hansen C. Improving spatial perception of vascular models using supporting anchors and illustrative visualization. Comput Graph 2017 Apr;63:37-49. [doi: 10.1016/j.cag.2017.02.002]
9. Lawonn K, Lzu M, Preim B, Hansen C. Illustrative visualization of vascular models for static 2D representations. In: Proceedings of the 18th International Conference on Medical Image Computing and Computer Assisted Intervention. 2015 Presented at: 18th International Conference on Medical Image Computing and Computer Assisted Intervention; October 5-9, 2015; Munich, Germany. [doi: 10.1007/978-3-319-24571-3_48]
10. Kreiser J, Hermosilla P, Ropinski T. Void space surfaces to convey depth in vessel visualizations. IEEE Trans Vis Comput Graph 2021 Oct 1;27(10):3913-3925. [doi: 10.1109/TVCG.2020.2993992] [Medline: 32406840]
11. Titov A, Kersten-Oertel M, Drouin S. The effect of interactive cues on the perception of angiographic volumes in virtual reality. Comput Methods Biomech Biomed Eng Imaging Vis 2021 Nov 08;10(4):357-365. [doi: 10.1080/21681163.2021.1999332]
12. Dergousoff K, Mandryk RL. Mobile gamification for crowdsourcing data collection: leveraging the freemium model. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. 2015 Presented at: CHI '15: CHI Conference on Human Factors in Computing Systems; April 18-23, 2015; Seoul, Republic of Korea. [doi: 10.1145/2702123.2702296]
13. Ahmed N, Mueller K. Gamification as a paradigm for the evaluation of visual analytics systems. In: Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization. 2014 Presented at: BELIV '14: Novel Evaluation Methods For Visualization 2014; November 10, 2014; Paris, France. [doi: 10.1145/2669557.2669574]
14. Connect brain. Google Play. URL: https://play.google.com/store/apps/details?id=ca.andreytitov.connectbrain&hl=en_CA&gl=CA [accessed 2023-01-11]
15. Connect brain. Apple Store. URL: https://apps.apple.com/ca/app/connect-brain/id1524359191 [accessed 2023-01-11]
16. Titov A. Comparing vascular visualization techniques with gamification. Concordia University. 2020. URL: https://spectrum.library.concordia.ca/987821/ [accessed 2023-08-29]
17. Mason W, Suri S. Conducting behavioral research on Amazon's Mechanical Turk. Behav Res Methods 2012 Mar;44(1):1-23. [doi: 10.3758/s13428-011-0124-6] [Medline: 21717266]
18. Cole F, Sanik K, DeCarlo D, Finkelstein A, Funkhouser T, Rusinkiewicz S, et al. How well do line drawings depict shape? In: Proceedings of the ACM SIGGRAPH 2009 papers. 2009 Presented at: SIGGRAPH09: Special Interest Group on

Computer Graphics and Interactive Techniques Conference; August 3-7, 2009; New Orleans, LA. [doi: 10.1145/1576246.1531334]

19. Snow R, O'Connor B, Jurafsky D, Ng AY. Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2008 Presented at: EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing; October 25-27, 2008; Honolulu, HI. [doi: 10.3115/1613715.1613751]
20. Marge M, Banerjee S, Rudnicky AI. Using the Amazon Mechanical Turk for transcription of spoken language. In: Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. 2010 Presented at: 2010 IEEE International Conference on Acoustics, Speech and Signal Processing; March 14-19, 2010; Dallas, TX. [doi: 10.1109/icassp.2010.5494979]
21. Unity real-time development platform. Unity Technologies. URL: https://unity.com/ [accessed 2023-01-11]
22. Blinn JF. Models of light reflection for computer synthesized pictures. In: Proceedings of the 4th annual conference on Computer graphics and interactive techniques. 1977 Presented at: SIGGRAPH '77: Computer graphics and interactive techniques; July 20-22, 1977; San Jose, CA. [doi: 10.1145/563858.563893]
23. Drouin S, Collins DL. PRISM: an open source framework for the interactive design of GPU volume rendering shaders. PLoS One 2018 Mar 13;13(3):e0193636 [FREE Full text] [doi: 10.1371/journal.pone.0193636] [Medline: 29534069]
24. Bailey M, Clark D. Using ChromaDepth to obtain inexpensive single-image stereovision for scientific visualization. J Graph Tools 1998;3(3):1-9. [doi: 10.1080/10867651.1998.10487491]
25. Lum EB, Ma KL. Hardware-accelerated parallel non-photorealistic volume rendering. In: Proceedings of the 2nd international symposium on Non-photorealistic animation and rendering. 2002 Presented at: NPAR02: Non-Photorealistic Animation and Rendering; June 3-5, 2002; Annecy, France. [doi: 10.1145/508530.508542]
26. Rheingans P, Ebert D. Volume illustration: nonphotorealistic rendering of volume models. IEEE Trans Visual Comput Graph 2001;7(3):253-264. [doi: 10.1109/2945.942693]
27. Kersten MA, Stewart AJ, Troje N, Ellis R. Enhancing depth perception in translucent volumes. IEEE Trans Vis Comput Graph 2006;12(5):1117-1123. [doi: 10.1109/TVCG.2006.139] [Medline: 17080842]
28. Steenblik RA. The chromostereoscopic process: a novel single image stereoscopic process. In: Proceedings of the True Three-dimensional Imaging Techniques and Display Technologies. 1987 Presented at: True Three-dimensional Imaging Techniques and Display Technologies; January 15-16, 1987; Los Angeles, CA. [doi: 10.1117/12.940117]
29. Thompson P, May K, Stone R. Chromostereopsis: a multicomponent depth effect? Displays 1993 Oct;14(4):227-234. [doi: 10.1016/0141-9382(93)90093-k]
30. Suzuki S, Abe K. Topological structural analysis of digitized binary images by border following. Comput Vis Graph Image Process 1985 Apr;30(1):32-46. [doi: 10.1016/0734-189x(85)90016-7]
31. Shepard D. A two-dimensional interpolation function for irregularly-spaced data. In: Proceedings of the 1968 23rd ACM national conferenc. 1968 Presented at: ACM '68: Proceedings of the 1968 23rd ACM national conference; August 27-29, 1968; New York, NY. [doi: 10.1145/800186.810616]
32. Ritsche T, Grosch T, Seidel HP. Approximating dynamic global illumination in image space. In: Proceedings of the 2009 symposium on Interactive 3D graphics and games. 2009 Presented at: I3D '09: Symposium on Interactive 3D Graphics and Games; February 27-March 1, 2009; Boston, MA. [doi: 10.1145/1507149.1507161]
33. Bavoil L, Sainz M. Screen space ambient occlusion. NVIDIA Corporation. 2008 Oct. URL: https://www.researchgate.net/publication/228576448_Screen_Space_Ambient_Occlusion [accessed 2021-10-12]
34. Kruger J, Westermann R. Acceleration techniques for GPU-based volume rendering. In: Proceedings of the IEEE Visualization. 2003 Presented at: IEEE Visualization; October 19-24, 2003; Seattle, WA. [doi: 10.1109/visual.2003.1250384]
35. Levoy M. Efficient ray tracing of volume data. ACM Trans Graph 1990 Jul;9(3):245-261. [doi: 10.1145/78964.78965]
36. Zuiderveld K, Koning, AH, Viergever M. Acceleration of ray-casting using 3-D distance transforms. In: Proceedings of the Visualization in Biomedical Computing 1992. 1992 Presented at: Proceedings Visualization in Biomedical Computing 1992; October 13-16, 1992; Chapel Hill, NC. [doi: 10.1117/12.131088]
37. HTC U12+ Mockup.png. Wikimedia. URL: https://commons.wikimedia.org/wiki/File:HTC_U12%2B_Mockup.png [accessed 2023-01-29]
38. CC BY-SA 3.0 DEED Attribution-ShareAlike 3.0 Unported. Creative Commons. URL: https://creativecommons.org/licenses/by-sa/3.0/ [accessed 2023-01-18]
39. Hart PE, Nilsson NJ, Raphael B. A formal basis for the heuristic determination of minimum cost paths. IEEE Trans Syst Sci Cyber 1968 Jul;4(2):100-107. [doi: 10.1109/tssc.1968.300136]
40. Fredman ML, Tarjan RE. Fibonacci heaps and their uses in improved network optimization algorithms. J ACM 1987 Jul 1;34(3):596-615. [doi: 10.1145/28869.28874]
41. LinkedIn: log in or sign up. LinkedIn. URL: https://www.linkedin.com/ [accessed 2023-03-23]
42. Twitter homepage. Twitter. URL: https://twitter.com/ [accessed 2023-03-23]
43. Facebook - log in or sign up. Facebook. URL: https://www.facebook.com/ [accessed 2023-03-23]
44. IBM SPSS software. IBM. URL: https://www.ibm.com/spss [accessed 2023-01-11]

45. Guérard-Poirier N, Beniey M, Meloche-Dumas L, Lebel-Guay F, Misheva B, Abbas M, et al. An educational network for surgical education supported by gamification elements: protocol for a randomized controlled trial. JMIR Res Protoc 2020 Dec 14;9(12):e21273 [FREE Full text] [doi: 10.2196/21273] [Medline: 33284780]
46. Washington P, Kalantarian H, Tariq Q, Schwartz J, Dunlap K, Chrisman B, et al. Validity of online screening for autism: crowdsourcing study comparing paid and unpaid diagnostic tasks. J Med Internet Res 2019 May 23;21(5):e13668 [FREE Full text] [doi: 10.2196/13668] [Medline: 31124463]
47. Meuschke M, Smit NN, Lichtenberg N, Preim B, Lawonn K. EvalViz – surface visualization evaluation wizard for depth and shape perception tasks. Comput Graph 2019 Aug;82:250-263. [doi: 10.1016/j.cag.2019.05.022]

Abbreviations

CTA: computed tomography angiography
DVR: direct volume rendering
MTurk: Amazon Mechanical Turk
SSDO: screen space directional occlusion
TF: transfer function
VSS: void space surface

Edited by P Kubben; submitted 19.01.23; peer-reviewed by F Blumberg, B Preim; comments to author 10.02.23; revised version received 03.04.23; accepted 20.09.23; published 20.10.23.

Please cite as:
Titov A, Drouin S, Kersten-Oertel M
Connect Brain, a Mobile App for Studying Depth Perception in Angiography Visualization: Gamification Study
JMIR Neurotech 2023;2:e45828
URL: https://neuro.jmir.org/2023/1/e45828
doi:10.2196/45828
PMID:

Original Paper

A Digital Telehealth System to Compute Myasthenia Gravis Core Examination Metrics: Exploratory Cohort Study

Marc Garbey[1,2,3,4], PhD; Guillaume Joerger[2,4], PhD; Quentin Lesport[1,3,4], MSc; Helen Girma[5], BS; Sienna McNett[5], BS; Mohammad Abu-Rub[5], BS; Henry Kaminski[5], MD

[1]Department of Surgery, School of Medicine & Health Sciences, George Washington University, Washington, DC, United States

[2]ORintelligence LLC, Houston, TX, United States

[3]Laboratoire des Sciences de l'Ingénieur pour l'Environnement (LaSIE UMR-CNRS 7356), University of La Rochelle, La Rochelle, France

[4]Care Constitution Corporation, Washington, DC, United States

[5]Department of Neurology & Rehabilitation Medicine, School of Medicine & Health Sciences, George Washington University, Washington, DC, United States

Corresponding Author:
Marc Garbey, PhD
Department of Surgery
School of Medicine & Health Sciences
George Washington University
2120 L St NW
Washington, DC, 20037
United States
Phone: 1 2815363178
Email: garbeymarc@gmail.com

Abstract

Background: Telemedicine practice for neurological diseases has grown significantly during the COVID-19 pandemic. Telemedicine offers an opportunity to assess digitalization of examinations and enhances access to modern computer vision and artificial intelligence processing to annotate and quantify examinations in a consistent and reproducible manner. The Myasthenia Gravis Core Examination (MG-CE) has been recommended for the telemedicine evaluation of patients with myasthenia gravis.

Objective: We aimed to assess the ability to take accurate and robust measurements during the examination, which would allow improvement in workflow efficiency by making the data acquisition and analytics fully automatic and thereby limit the potential for observation bias.

Methods: We used Zoom (Zoom Video Communications) videos of patients with myasthenia gravis undergoing the MG-CE. The core examination tests required 2 broad categories of processing. First, computer vision algorithms were used to analyze videos with a focus on eye or body motions. Second, for the assessment of examinations involving vocalization, a different category of signal processing methods was required. In this way, we provide an algorithm toolbox to assist clinicians with the MG-CE. We used a data set of 6 patients recorded during 2 sessions.

Results: Digitalization and control of quality of the core examination are advantageous and let the medical examiner concentrate on the patient instead of managing the logistics of the test. This approach showed the possibility of standardized data acquisition during telehealth sessions and provided real-time feedback on the quality of the metrics the medical doctor is assessing. Overall, our new telehealth platform showed submillimeter accuracy for ptosis and eye motion. In addition, the method showed good results in monitoring muscle weakness, demonstrating that continuous analysis is likely superior to pre-exercise and postexercise subjective assessment.

Conclusions: We demonstrated the ability to objectively quantitate the MG-CE. Our results indicate that the MG-CE should be revisited to consider some of the new metrics that our algorithm identified. We provide a proof of concept involving the MG-CE, but the method and tools developed can be applied to many neurological disorders and have great potential to improve clinical care.

(JMIR Neurotech 2023;2:e43387) doi:10.2196/43387

KEYWORDS

telehealth; telemedicine; myasthenia gravis; ptosis; diplopia; deep learning; computer vision; eyes tracking; neurological disease

Introduction

With the COVID-19 pandemic, there was a rapid increase in the use of telemedicine in routine patient care [1] and in clinical trials that moved to video evaluations to maintain subject follow-up [2]. Telemedicine was already commonly used for acute stroke care and was in development for Parkinson disease, but the vast majority of neurologists were not using such approaches and were suddenly thrust into unfamiliar territory [3-5]. Diagnosis and monitoring of neuromuscular disorders, in particular, rely on a nuanced physical examination, and specialists would be particularly reticent to use telemedicine. However, telemedicine has great potential to provide improved assessment of aspects of neurological examinations, and facilitate patient monitoring and their education [6], while reducing patient burden in attending in-person clinic visits and potentially increasing access. Further, there is great potential for rigorous video assessment to enhance clinical trial performance, which could reduce the burden on study participants and thereby enhance recruitment and retention.

The Myasthenia Gravis Core Examination (MG-CE) [7] was recommended for telemedicine evaluation of patients with myasthenia gravis (MG), and it involves specific aspects of neurological examinations critical to the comprehensive assessment of patients with MG. The National Institutes of Health Rare Disease Clinical Research Network dedicated to MG, MGNet, initiated an evaluation to assess the feasibility and validity of MG-CE for use in future clinical trials. These assessments were video recorded using the software Zoom (Zoom Video Communications), and we used the evaluations performed at George Washington University with the following 2 objectives: (1) assess workflow efficiency by making the data acquisition and analytics fully automatic and (2) evaluate the potential to quantitate the evaluations.

Methods

MG-CE and Automatic Data Acquisition

The study used recorded telemedicine evaluations of individuals with a clinical- and laboratory-confirmed diagnosis of MG. The patients were provided instructions regarding their position in relation to the cameras and level of illumination, and were told to follow the examiner's instructions. We used videos of 6 subjects recorded twice within 7 days to develop our algorithms. One normal control subject was used to evaluate the methodology prior to evaluating MG subject videos.

The MG-CE is summarized in Table 1, and a full description has been provided previously [7]. The examination required 2 broad categories of processing: (1) the computer vision algorithm to analyze video focusing on eye or body motions and (2) the analysis of the voice signal, which requires a completely different category of signal processing methods. We describe successively each of the techniques used in these categories and summarize the digitalization process in Table 2.

Table 1. Myasthenia Gravis Core Examination exercises and evaluation metrics [7].

Variable	Normal (0)	Mild (1)	Moderate (2)	Severe (3)
Eyelid droop (ptosis)	No ptosis	Eyelid above the pupil	Eyelid at the pupil	Eyelid below the pupil
Double vision (right/left)	No diplopia with a gaze of 61 seconds	Diplopia with a gaze of 11-60 seconds	Diplopia with a gaze of 1-10 seconds	Immediate diplopia
Cheek puff	Normal "seal"	Transverse pucker	Opposes lips but air escapes	Cannot perform the exercise
Tongue to cheek	Normal: full convex deformity in the cheek	Partial convex deformity in the cheek	Able to move the tongue to the cheek, but no deformity	Cannot perform the exercise
Counting to 50	No dysarthria at 50	Dysarthria at 30-49	Dysarthria at 10-29	Dysarthria at 1-9
Arm strength	No drift for >120 seconds	Drift at 90-119 seconds	Drift at 10-89 seconds	Drift at 0-9 seconds
Single-breath count	Count of ≥30	Count of 25-29	Count of 20-24	Count of <20
Sit-to-stand maneuver	No difficulty	Slow with effort but no hands	Need to use hands	Unable to stand unassisted

Table 2. Summary of our algorithm tool box to assist the clinician with the Myasthenia Gravis Core Examination.

Exercise	Description	Observation	Metric	Digital tool
Ptosis	Patients hold their gaze up for 60 seconds.	Weakness of the upper eyelid and eyelid going above the pupil.	Distance between the eyelid and the pupil, and distance between the upper and lower eyelids.	High-definition camera and eye segmentation.
Double vision	Patients hold their gaze right and then left for 60 seconds.	Misalignment of the eyes and moment of double vision.	Track the distance between anatomic landmarks such as the upper/lower lid, and pupil and iris left and right boundaries.	High-definition camera and eye segmentation.
Cheek puff	Patients puff their cheeks and hold it.	• Assess muscle strength and fatiguability. • Extent of puffiness at baseline and versus external pressure placed on the cheeks. • Symmetry of cheek puff (left vs right).	Track face feature variation, mouth curvature, and dimension in particular.	• Depth camera or Lidar. • High-definition camera with face landmark monitoring. • Track change of illumination in the region of interest.
Tongue pushing	Patients use their tongue to push the cheek.	Tongue muscle strength and symmetry.	Track face feature variation, mouth curvature, dimension, and orientation in particular.	• Depth camera or Lidar. • High-definition camera with face landmark monitoring. • Track change of illumination in the region of interest.
Counting to 50	Patients count out loud from 1 to 50.	Assess for respiratory muscle fatigue and shortness of breath.	• Loudness of the voice. • Various types of spectral analysis of the voice and mouth motion. • Energy metric of the voice.	Lip tracking and sound analysis of the exercise clip.
Arm strength	Patients hold their arms straight.	Assess for muscle fatigue via sustained abduction of the arm.	• Track body pose and different angles. • Length of time the patient can hold the arm in the pose. • Trajectory of the arm over time.	Pose detection on high-definition images.
Single-breath test	Patients count with only 1 breath.	Assess for respiratory muscle fatigue.	Length of the breath.	Lip tracking and sound analysis of the exercise clip.
Sit-to-stand maneuver	Patients stand up with and without crossing their arms.	• Assess for muscle fatigue. • Ability of the patient to stand without using the arms for assistance.	• Body pose tracking. • Compare standing up speed between clips.	Pose detection on high-definition images.

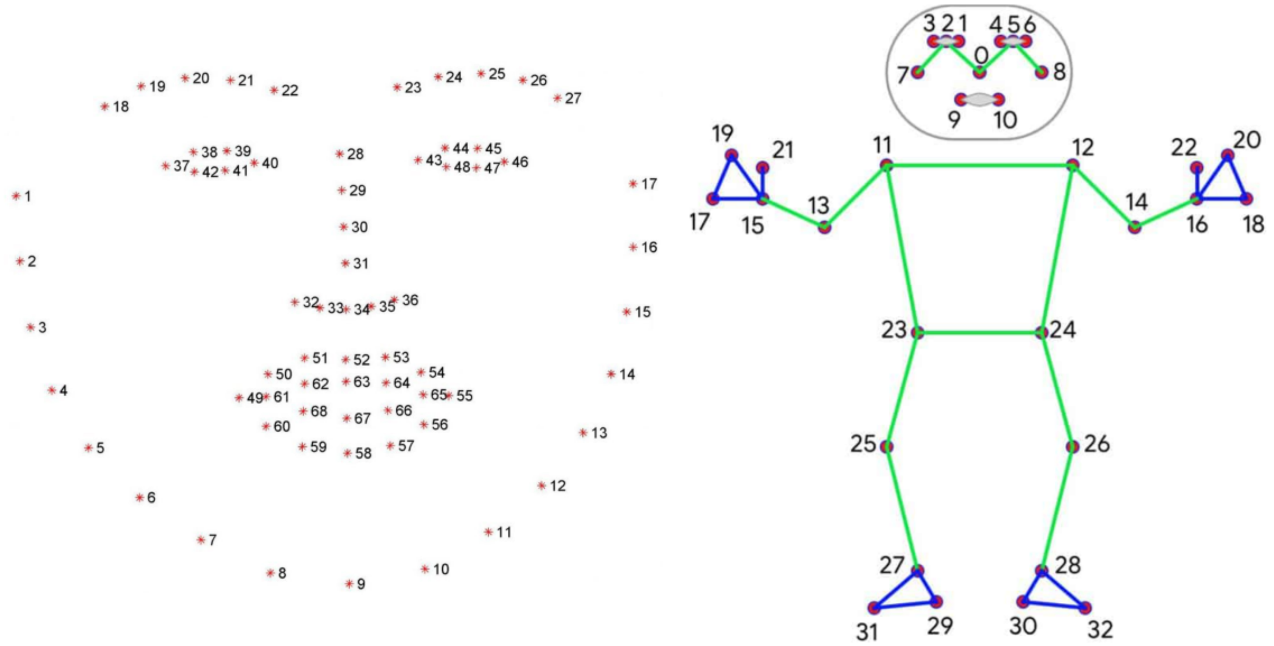
Deep Learning and Computer Vision Analysis

Machine Learning to Track Body Landmarks and Face Landmarks

Tracking faces or all body motions has become a standard tool [8] thanks to publicly available deep learning libraries with a standard model (Figure 1). To track body positions during the test of arm position fatigue and the sit-to-stand maneuver (Figure 1), we used a deep learning model that is publicly available (the pretrained machine learning model BlazePose GHUM 3D from MediaPipe) (Figure 1) [9]. For eye detection, we first needed to localize the face in the video frame.

Among the most commonly used algorithms [10,11], we chose OpenCV's implementation of the Haar Cascade algorithm [12], based on the detector from Lienhart et al [13]. Our criteria to select the method were speed and reliability for real-time detection.

Figure 1. Pretrained machine learning models used with characteristic points.



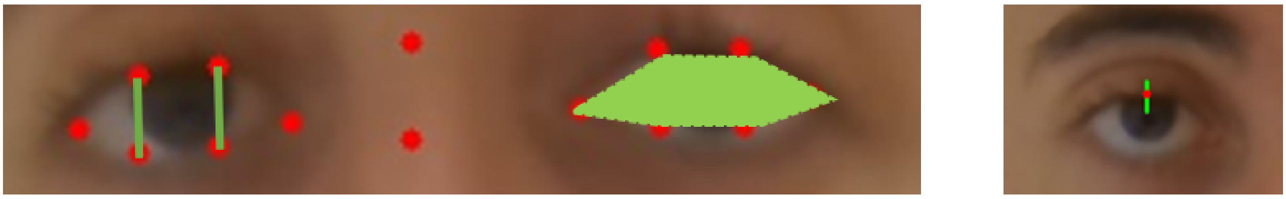
To focus on the regions of interest (ROIs) of the eyes and lids, we used the pretrained DLib 68 points facial landmark detector that is based on the shape regression approach [14,15]. It is a machine learning algorithm that places 68 characteristic points on a detected face. The model is pretrained on the I-BUG 300-W data set, which is comprised of 300 face pictures (Figure 1) [16]. This software was used for the assessment of ptosis and eye position, as well as for the test of counting to 50 and the single-breath test in order to document lip reading and tracking of jaw motion (Table 1).

Overall, both libraries provided robust results and could be used to annotate the video in real time for the ROIs. However, we found that the accuracy of the landmark points in the model of Figure 1 obtained by this library was not adequate to provide metrics that could be used in eye motion assessment in the context of a standard telehealth session. Therefore, we developed a hybrid method that began from deep learning to identify the ROIs and refine the search for the pupil, eyelid, and iris as described next.

Eye and Lid Image Segmentation

Assessment of ptosis and ocular motility requires precise tracking of the eyelid, pupil, and iris. Precise metrics of these measures have been developed [17-20]. Established techniques to detect the iris location [21] are the circular Hough transform [22] and the Daughman algorithm method [23]. However, we found that these methods lack robustness due to their insensitivity to the low resolution of the ROIs of the eyes, poor control for illumination of the subject, and specific eye geometry consequent to ptosis. The eye image in a standard Zoom meeting may not be bigger than about 40 pixels wide and 20 pixels high. Liu et al [24] assessed eye movements for a computer-aided examination, but with highly controlled data and a highly controlled environment. We did not have optimum control of telehealth footage with patients at home, and the eye region has only one-tenth, at best, of the image frame dimension. Therefore, we took a more versatile approach that began with the ROI given by the previous deep learning library that we had used and then concentrated on a local search of the iris boundary, pupil center, and upper/lower eyelid (Figure 2). Since we started from a good estimate of the ROI for the eye, we used a combination of a local gradient method and clustering technique to compute the spatial coordinate and distance between landmarks of interest, and we have described this in the Results section. There are 2 classes of assessment depending on whether we compute the geometric dimension on an individual image or the dynamic of eye motion on video clips. We retrieved, for example, the relaxation time of the eyelid versus equilibrium, with some of the patients performing both eye exercises (Figure 2). However, there is no mention of such a metric in the core examination [11]. The incorporation of this new information in the standard core remains to be determined.

Figure 2. Approximations on ptosis to assess the field of view: distance between the upper and lower eye lids (left), eye area opening (center), and distance from the upper lid to the pupil (right).



Body Image Segmentation

To have reproducible results with the entire view of the body during the examination, we tested our telehealth platform Inteleclinic on 1 patient and several control subjects. The pretrained machine learning model BlazePose GHUM 3D from MediaPipe [9] has been evaluated extensively, so we only provide some examples of the results obtained with the MG-CE. The arms of the patient are extended for 2 minutes during the exercise, and we used the segments joining the landmark point (12) to (14) to track the right arm position and the landmark point (11) to (13) to track the left arm position (Figure 1B). We computed the angle formed by the arm's segment as described above and the horizontal line going through the landmark points (11) and (12) of the upper torso in the model (Figure 1B). If the arms stay horizontal, the 2 angles we track for the model (Figure 1B) should be approximately zero. As the arm strength of the patient may fatigue during the exercise, the arms fall from the horizontal position, and the angle would decrease and become negative. A similar approach was used for the sit-to-stand exercise by tracking the hip landmarks (23) and (24) of the body motion model (Figure 1B).

Cheek Deformation

The ROI for cheek deformation was the polygon delimited by points (3), (15), (13), and (5) of model 1 (Figure 1A) for the cheek puff exercise. We could restrict this ROI to one half of the polygon for the tongue-to-cheek push exercise that is only performed on one side. As we aimed to reconstruct the local curvature of the cheek during the test involving (3) and (4) that can lead to cheek deformation, we used a depth camera and computed the depth map to assess the contour of the deformation in the ROI. When it came to the depth map, our first approach was to use a depth camera that could directly reconstruct the local curvature of the surface seen. The depth camera Intel Realsense D435 (Intel) has, according to the vendor, a relative accuracy below 2% for a distance less than 2 meters. This technology uses infrared and stereo cameras to analyze the deformation of a projected pattern of a scene and reconstruct from this information the depth, but requires camera calibration [25-28]. All tests were performed in realistic conditions for telehealth, that is, the distance of the face from the camera was 1 meter at minimum and the patient was directly facing the camera.

The second approach we used was to assess a pure computer vision technique that works on a standard video. Our objective was to define basic information regarding when the cheek deformation starts, when it ends, and how it may become weaker during the examination period. In practice, this is what the medical doctor may grade during a telehealth consultation.

The first solution exploits the local skin appearance alterations as the cheek becomes dilated [29]. We could then compute the ROI "centered" on the cheek area where we expect the deformation to be most significant and the average pixel value of the blue dimension of the RGB code. To track the ROI, we used the mouth location and external boundary of the cheek that can be recovered from the model (Figure 1A). We could then track the average value over time during the exercise, that is, before the push to its end. We show in the Results section the limitation of this method that is a priori not robust with respect to light conditions and may depend on skin color.

The second solution is based on the observation that cheek deformation impacts the mouth geometry. For example, in the cheek puff exercise, the mouth is closed and invariably the lip shape features change from those in the rest position. In the one-side tongue-to-cheek push, the upper lip is deformed. All these changes can be monitored in time easily by tracking the relative position of the points in the facial model that mark the mouth (Figure 1A).

We describe our computer vision methods based on an analysis performed with 3 different formats of videos. The first was acquired with our new telehealth platform using a high-definition camera with a patient who has a normal cheek puff response. The second was acquired on a control subject with a cell phone camera (Apple 13 system, Apple Inc), and the third was extracted from the MGNet data set. We tested the impact of diversity with White subjects, subjects with dark sun tan, and subjects who were African American. We demonstrate in the Results section which metrics appeared to provide the best assessment.

Voice Analysis

Our goal was to assess breathing and change in speech in patients with MG from analyzing counting to 50 and single-breath count. Dysarthria is not a simple concept and is classified in several ways [30]. Shortness of breath was easier to define but could be compromised by multiple factors. Shortness of breath and pulmonary function can be assessed from speech as appreciated by others [29,31]. Previous studies have used machine learning and artificial intelligence (AI) techniques that require large training sets, and they are not specific to any neurological disorder or specific to a voice acquisition protocol.

A good example of dysarthria detection has been published previously [32]. The rate of success of a neural network is modest, that is, about 70% when competing with standard diagnostic performance. An alternative solution is to use a fractal feature as reported previously [33]. This methodology seems to reach a greater accuracy of about 90% and does not require a training set.

Lip and jaw movements are related to dysarthria [34]. We are not aware of any systematic study that combines automatic lip motion tracking and speech digital analysis to assess breathing and dysarthria in patients with MG. We assessed more than half a dozen algorithms producing various sound metrics to check for the potential best voice analysis candidate to assess MG patients. As the analysis of the pitch of voice did not show any outliers in the data set and the energy metric analysis was impacted by the environment and control of the exercise, we restricted the description to the most promising algorithm. To compute voice features, we used the following steps. We separated the interval of time when the subject spoke from when the subject was silent. We used the MATLAB function "detectSpeech" [35] on the original signal. The function "detectSpeech" provides the start and end times of each so called "speech segment." The frequency of signal acquisition was about 1000 Hz. For comparison, we used our own custom-made algorithm to extract speech segments using sampling of size 60 of the voice signal. The signal now had an equivalent frequency of acquisition of about 17 Hz. We then used averaging on each sample of the original signal to dampen noise. The signal was then smoother, and we could use a threshold to filter out noise without building up a large number of small gaps corresponding to "no sound." We looked in the sound track of "counting to 50" exercises for the largest 50 time intervals of sound above noise level. All voice features presented below were computed on the sound track that contained speech only.

We present below the list of voice features we computed systematically for each of the sound tracks for both voice exercises. All these individual metrics or combinations of metrics were candidates to grade the severity of symptoms. The Results section reports which metric worked the best. The features are as follows:

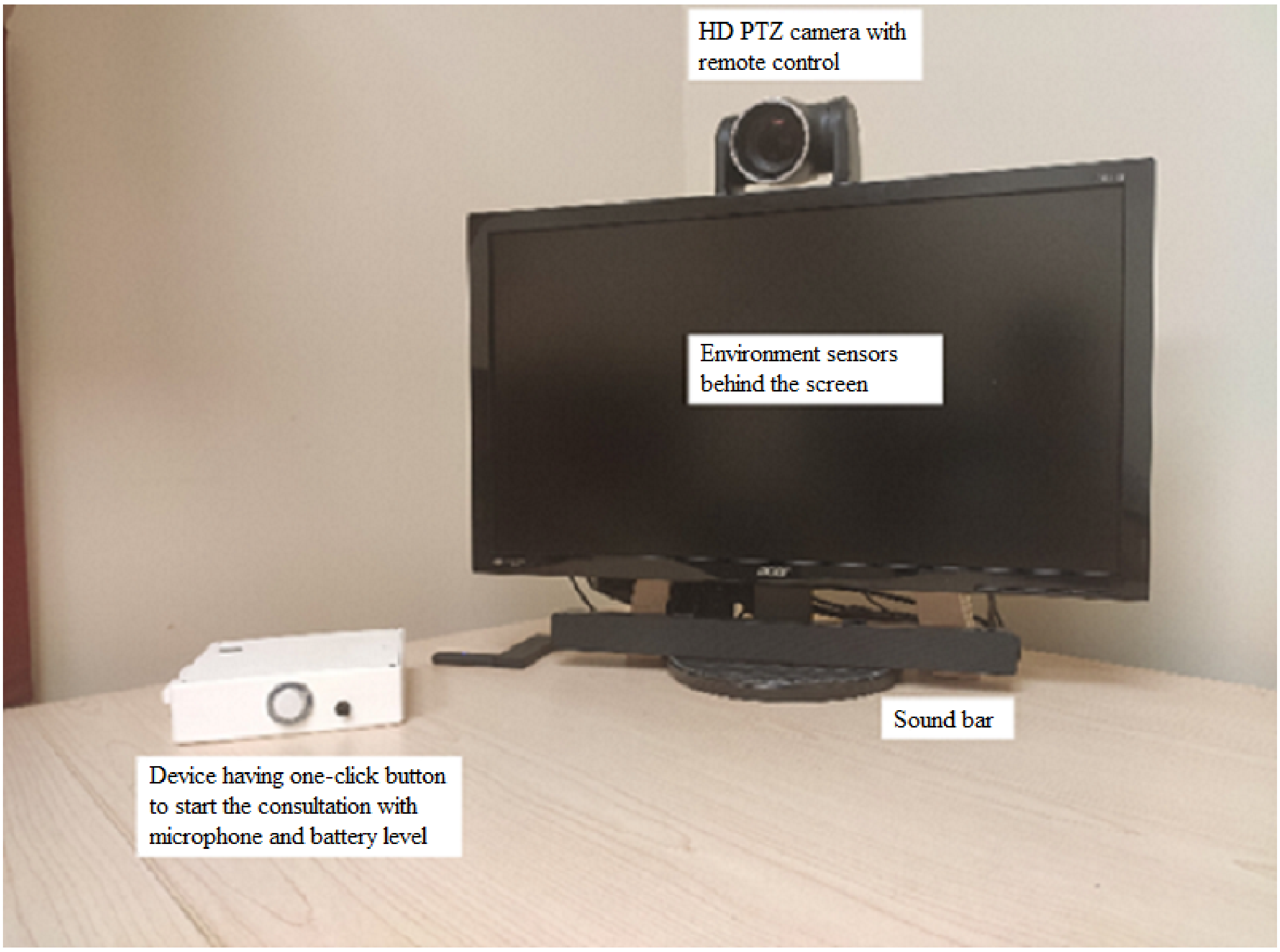
- Loudness of voice: Loudness was computed based on the algorithms defined in the ITU-R BS.1770-4 and EBU R 128 standards. The loudness of voice was integrated over all speech segments.
- Pitch or fundamental frequency of voice: The pitch was computed for each speech segment. The speech of a typical adult man will have a fundamental frequency from 85 Hz to 155 Hz and that of a typical adult woman will have a fundamental frequency from 165 Hz to 255 Hz.
- Spectral energy on a frequency interval: Both voice exercises were considered as breathing exercises, so we computed the L2 norm spectral energy of the voice signal over all speech segments in a frequency window that focused on the breathing rate (5 Hz to 25 Hz).
- Teager-Kaiser energy: It was used in tone detection [36].
- Spectral entropy of the voice signal: Spectral entropy is a measure of spectral power distribution. Spectral entropy's concept is based on Shannon entropy or information entropy. Spectral entropy treats the signal's normalized power distribution in the frequency domain as a probability distribution and calculates the Shannon entropy of it. The Shannon entropy has been used for feature extraction in fault detection and diagnosis [37,38]. Spectral entropy has also been widely used as a feature in speech recognition [39] and biomedical signal processing [40].
- Special feature of the single-breath count: The airflow volume expansion during speech is in first approximation related to the square of the amplitude of the sound wave [41]. We computed the integral of the square of the amplitude of the sound wave during the time window of the patient's speech. Since there is no calibration of the microphone, the metric might be biased. There was considerable variability of diction during this exercise. Some subjects counted more slowly, while others appeared anxious and pronounced words quickly. We computed as an additional feature the percentage of time with vocal sound versus total time.

For the voice analysis test in particular and for tests in general, there was significant variability in the parameters of data acquisition under clinical conditions, such as sound level. Providing guidance in real time to the patient will be essential to improve the ability to quantitate the telehealth examination.

The Need for a Novel Telehealth Platform to Support the Protocol and Improvement of Data Acquisition

Reproducibility requires that the various examinations are run in similar conditions. While we have mainly evaluated our algorithms on an existing data set of standard Zoom video evaluations with 6 patients, we next describe our new hardware and software solution named "Inteleclinic" (Figure 3) designed to improve data acquisition.

Figure 3. View on the patient side of our cyber-physical system named “Inteleclinic” used to uniformize the sessions and improve the quality of the metrics. HD: high-definition.


HD PTZ camera with remote control
Environment sensors behind the screen
Sound bar
Device having one-click button to start the consultation with microphone and battery level

Controlling the Setting and Hardware

To avoid changes in the quality of the recording (frames and audio), it is important that the hardware used is identical for all sessions and is calibrated. In an attempt to improve the quality of data acquisition with the future development of clinical studies, we built a new telehealth system with a high-definition camera and microphone that can be controlled remotely by the examiner. The recording is now performed on the patient side to obtain the raw footage of the video and audio in order to optimize maximum resolution and avoid any issues with network connection quality during the examination. Different interfaces and tools are added compared to Zoom’s control system to assist the patient and the doctor to focus on the consultation and not the technology. We demonstrate in the Results section the benefits of Inteleclinic compared to a standard Zoom video call of a patient at home using various technologies.

Controlling Time

Some of the tests, such as lid and eye positions as well as arm position, require precise timing from start to end. To avoid any manual entry errors, our digital heath system automatically computes start and end times. The single-breath test as practiced now has no control on loudness and timing. Breath capacity is the product of the air outflow flux and the duration of the exercise. To be more precise, breath capacity is the time integration of the time-dependent output flow on the time interval of the exercise. We found that the maximum number counted is weakly correlated with the duration of the counting exercise. The maximum number reached is dependent on diction and may not be used as a valuable metric. Outflow during speech is proportional in the first approximation to the square of the energy source of sound [41]. Depending on how loud the count is, one may expect different airflow output values for the patient. We tested on our platform a visual aid on the telehealth display to guide patient counting with a consistent rhythm of about 1 number counted per second in both the count to 50 and single-breath counting exercises.

Controlling the Framing of the ROI and the Distance From the ROI to the Camera

The digitalization of the tests involving ptosis, diplopia, cheek puff, tongue to cheek, arm strength, and sit-to-stand movement depends heavily on vision through the telehealth system. While we used a standard Zoom video that was preregistered in this study, it is straightforward with our system to provide guidance on the quality of video acquisition to make sure that distance between landmarks of interest use approximately the same number of pixels in order to provide quality and consistency for the results. In practice, we can provide a mark on the display of the patient with Inteleclinic to make sure the individual is properly centered and distanced from the camera.

Controlling Sounds

Every telehealth session may have different loudness of the sound track depending on the microphone setting and how loud or how soft the patient is speaking. Loudness is computed based on the algorithms defined in the ITU-R BS.1770-4 and EBU R 128 standards. If the microphone is calibrated with a benchmark sound, the loudness of the sound track can be computed continuously and guidance can be provided to the patient on how loud or soft they should try to keep their voice during the exercises.

Ethics Approval

All participants provided written consent for inclusion in the study. The study that provided the data has been approved by the George Washington University Institutional Review Board (IRB# NCR224008).

Results

Standardization of the Data Acquisition

We evaluated our methods and identified large variability in the data acquisition and mode of operation for the assessments performed by the single examiner. The purpose of the primary study from which we obtained the videos was to assess test/retest variability and interexaminer variation in performance of telemedicine evaluations (manuscript in preparation). Our goal was to assess accurate and robust measurements from the MG-CE in order to remove human bias. The videos from the clinical study are quickly showing some limitations as the hardware used to film was not identical and the recording was performed on the doctor side, linking the quality of the frames to the quality of the network on both sides. We will next present our methodology and the platform. One of the purposes of our project was to standardize the data acquisition during the telehealth session and provide real-time feedback of the quality of the assessments for the examiner.

Eyelid Position and Eye Movements

We have used a data set of images of 6 patients and 3 healthy subjects with broad diversity in skin color, eye color, ocular anatomy, and image frame resolution to test the accuracy of our approach. We identified 72 ROIs of patients' eye movements and then annotated them with ground true measures obtained by manually zooming in the computer images. On average, we found the pupil location within 3 pixels and the distance from the pupil to the upper eyelid within 2 pixels independent of the image frame resolution. The relative accuracy in pixels was independent of the camera used. The standard Zoom video has 450×800 pixels per frame, a smartphone has 720×1280 pixels, and the Lumens B30U PTZ camera (Lumens Digital Optics Inc) has a resolution of 1080×1920. Overall, Inteleclinic doubles the resolution of a standard Zoom call and provides a submillimeter accuracy of lid position and eye motion. Figure 4 provides an example of the localization of the upper lid, lower lid, and iris lower boundary detected automatically with our hybrid method using digital zooming of the face of the patient with both ROIs. We underline that the patient is sitting about 1 meter from the camera during the telehealth eye exercises, and one can see the patient's face and shoulders. No particular effort was made to focus on the eyes of the patient in the video. The 6 red circles in each eye correspond to the markers of the ROI obtained with the deep learning library of model 1 (Figure 1A). The bottom markers are slightly off, and our local computer vision technique provides the ability to correct the position of the lower lid.

In Figure 5, we show tracking of the distance between the lower boundary of the iris and the upper lid with a black curve, and the distance between the bottom of the iris and the lower lid with a red curve. One can check that the patient performs the exercise properly and can measure a 15% decay of the ptosis distance during the 1-minute exercise. As shown in the green least square fit with the green line, this decay is both linear and statistically significant. This decay is in fact difficult to notice during a medical examination without our method.

In Figure 6, we report on the second exercise that tests diplopia. The red circle locations of the deep learning model of Figure 1A in the ROIs are accurate. We tracked the vertical border of the iris and computed the barycentric coordinate of the most inner points of the boundaries to compute any eventual misalignment of both eyes. The patient did not report double vision, and the quasisteady variation of the barycentric coordinates, as reported in Figure 7, confirmed this.

However, the positions of the eyes of patients might be so extreme that some of the pupils might be partially obstructed during the exercise, which limits the value of the conclusion. In addition, we clearly observed ptosis during the exercise as the vertical dimension of the eye opening reached about half of what it was during the ptosis exercise.

Figure 4. Image during the ptosis exercise. Digital zoom on the view of the patient obtained with the Inteleclinic system showing anatomic markers obtained by computer vision in green, starting from the landmarks of the regions of interest obtained by deep learning.

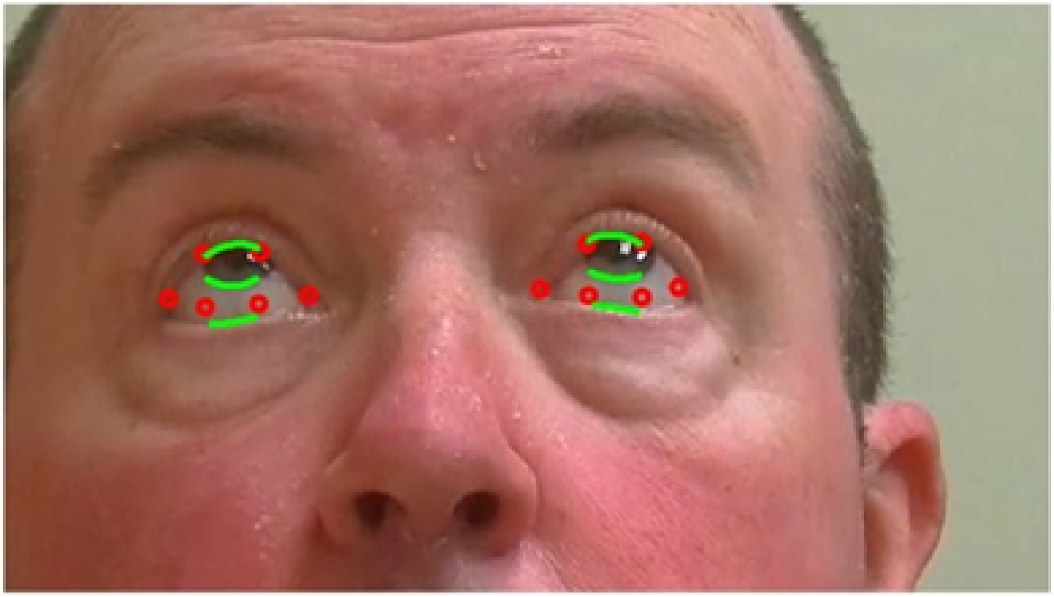


Figure 5. Graphic representation of the distance between anatomic landmarks to asses ptosis dynamically during the first eye exercise.

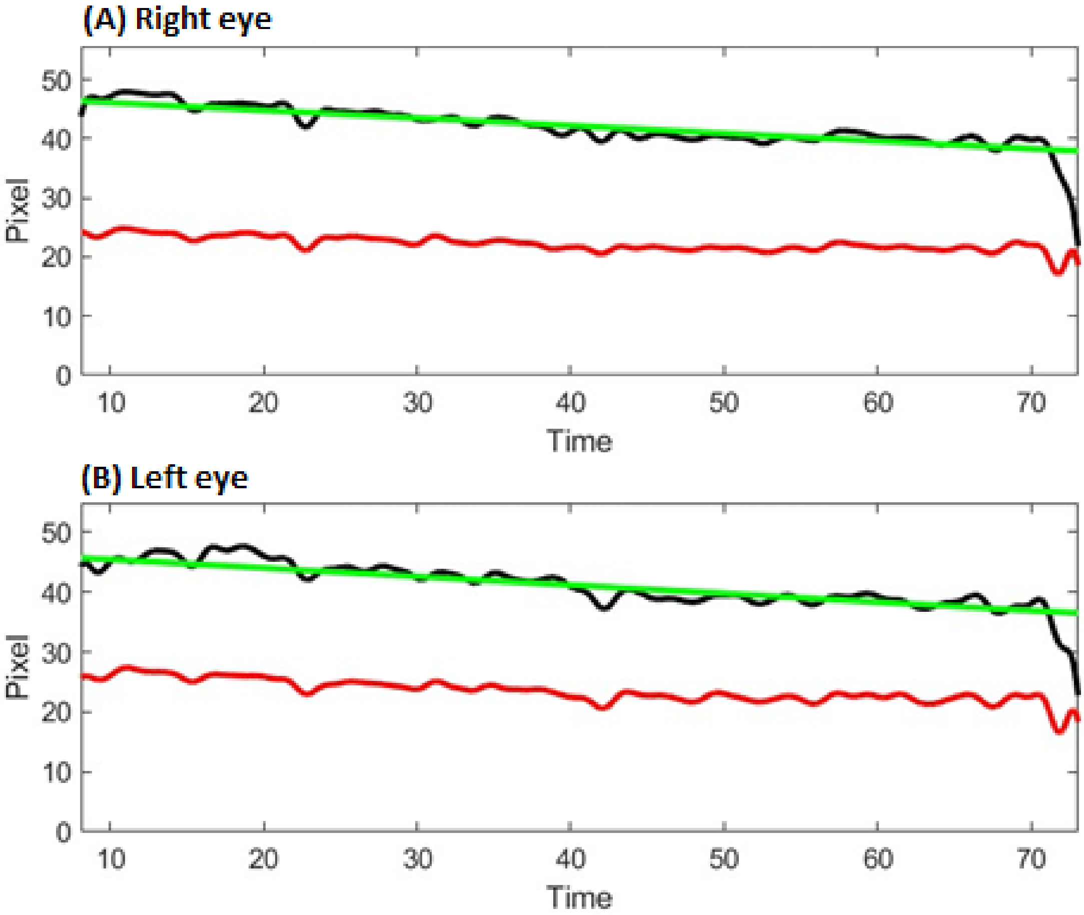


Figure 6. Image during the diplopia exercise. Digital zoom on the view of the patient obtained with the Inteleclinic system showing anatomic markers obtained by computer vision in green, starting from the landmarks of the regions of interest obtained by deep learning.

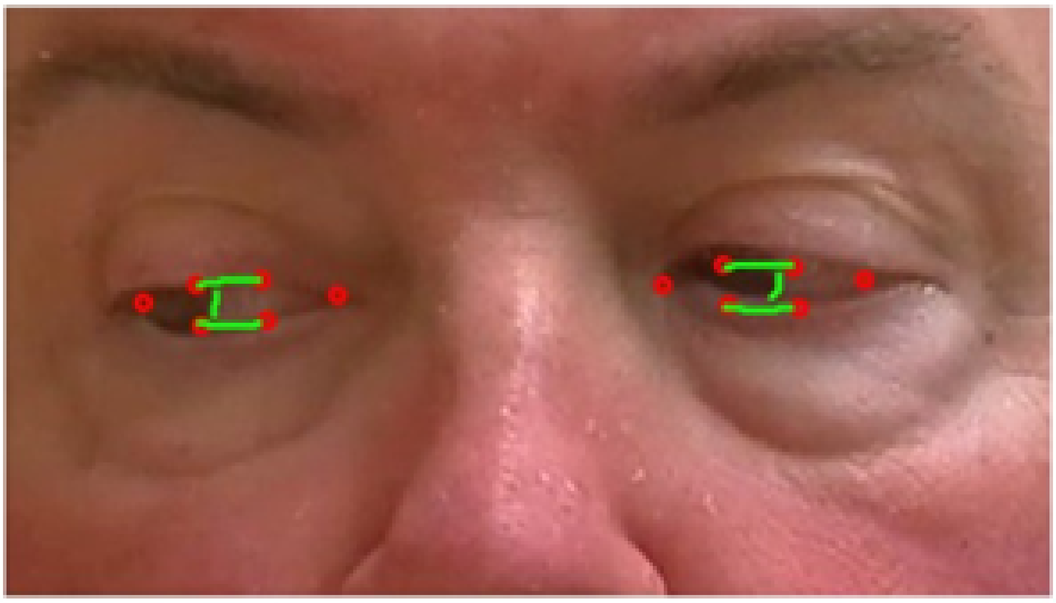
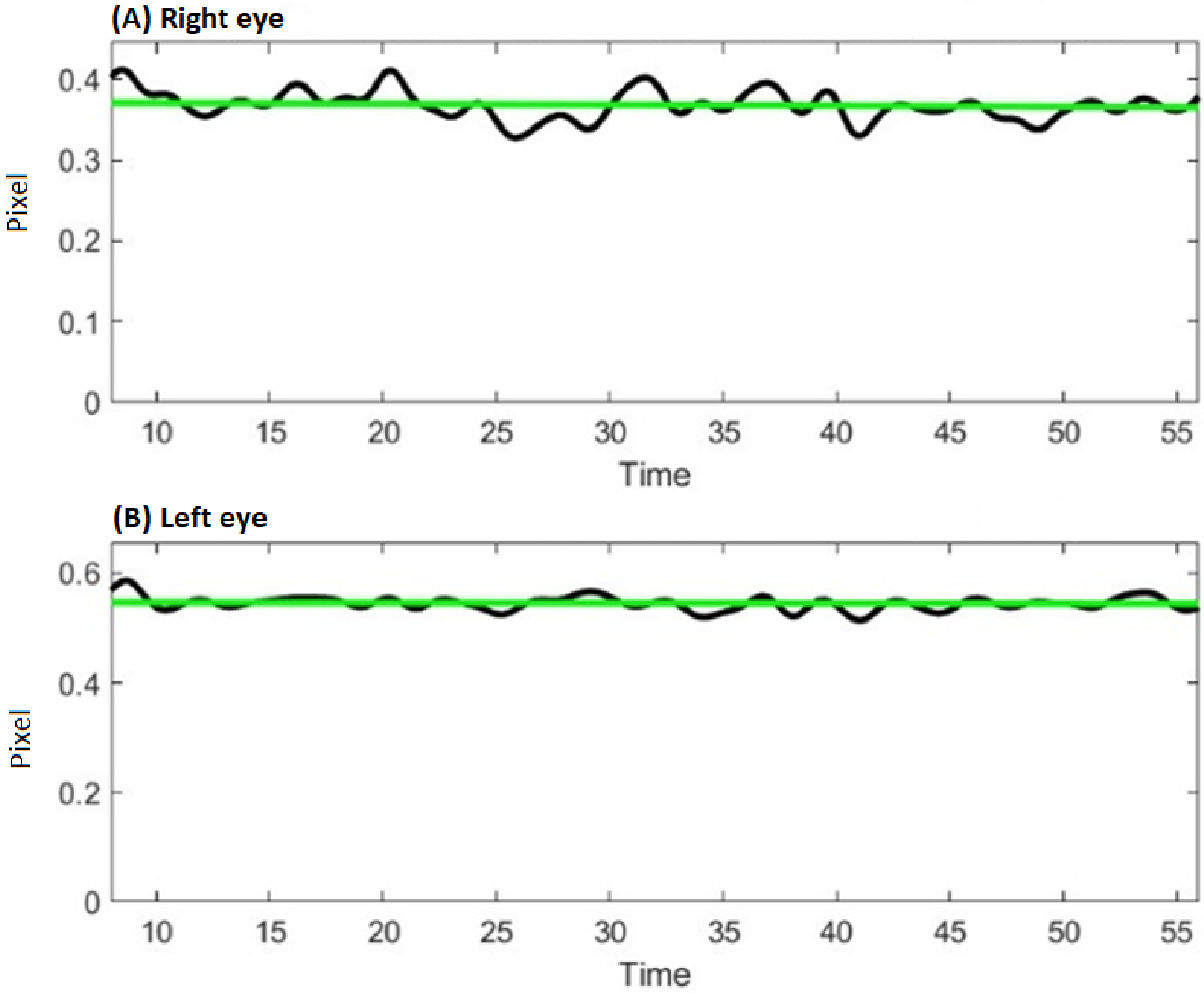


Figure 7. Graphic representation of the bariatric coordinate of the anatomic landmarks used to assess eye alignment dynamically during the third eye exercise.



Cheek Puff and Tongue to Cheek

We used a low-cost depth camera from Intel to reconstruct the local curvature of the cheek in laboratory conditions with a healthy subject who produced a large deformation, which was at the noise level of the signal. This evaluation would have failed for any patient who has difficulty to push the tongue into the cheek. Better depth accuracy could be obtained by sensors that use time-of-flight technology [42].

The variety of videos demonstrates the limits and potential of our approach. In one video of the cheek puff exercise, the patient was told to blow his cheeks for about 2 seconds. The video was cut after 15 seconds because the patient was asked to test the stiffness of the skin with his fingers. The placement of the fingers on the cheek completely confused the AI tracking algorithm. The change in the mean value of the third component (blue) of the RGB classification inside the ROI on both sides of the cheek of the patient is not reliable unless the left cheek or right cheek ROI has good illumination. The detection is usually far less reliable on one of these ROIs because it is difficult to achieve good illumination on both sides of the face of the patient.

Tracking mouth deformation during the exercise was a superior approach. First, we easily detected if and when the patient had the ability or did not have the ability to keep the mouth closed. Second, we tested several features, such as the distance between the corners where the upper and lower lip meet, that is, the segment delimited by the points (49) and (55) in the model (Figure 1A), the deformation of the mouth in the vertical direction, and the mean curvature of the upper lip and lower lip. Figure 8 shows the feature that measures the normalized distance between the upper lip and the bottom of the nose during the cheek puff exercise using a standard Zoom video with 450×800 pixels per frame in an ADAPT (Adapting Disease Specific Outcome Measures Pilot Trial) patient. We obtained a curve that was close to the step function in this ADAPT patient, which accurately detected when the deformation of the cheek started and ended, and indicated how strong the deformation was. Not all features work all the time for all patients. As expected, variability in the anatomy of patients causes differences in which features work the best. Form our experience, we found that the combination of several features helps identify the extent of cheek puff during the exercise.

We obtained very similar results for the tongue-to-cheek push exercise. In Figure 9, an ADAPT patient pushes the left cheek with the tongue and then pushes the right cheek at 5.6 seconds.

Figure 8. Normalized distance of the upper lip to the lower part of the nose during the cheek puff exercise.

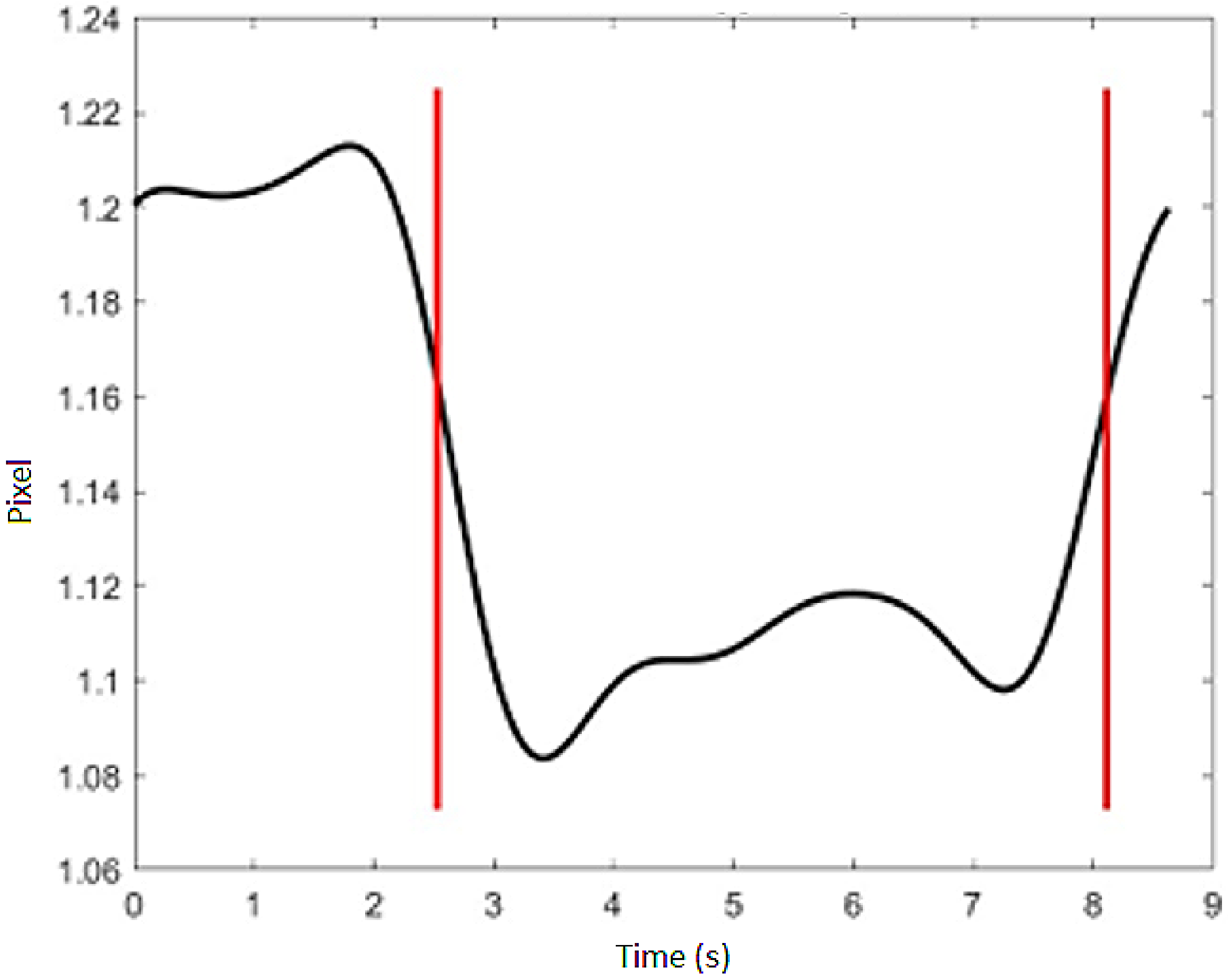
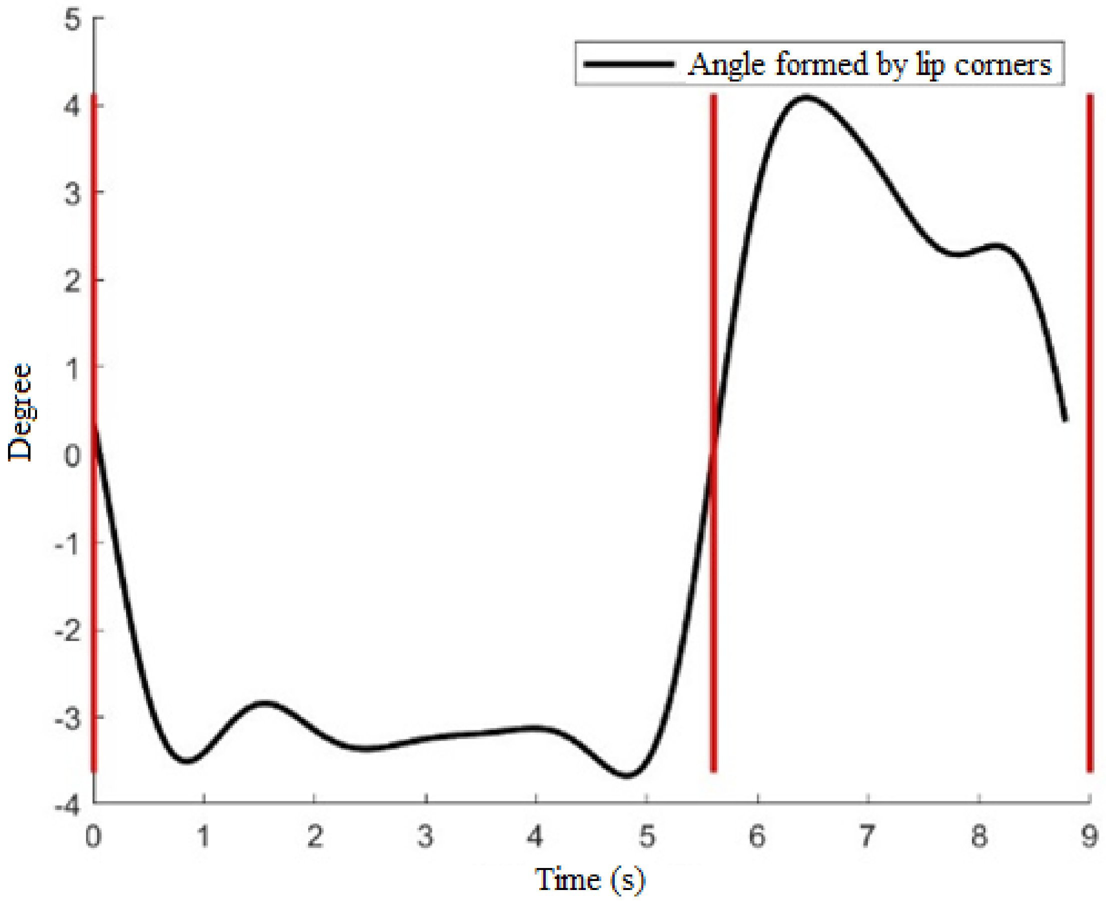


Figure 9. Exercise involving the tongue pushing the left cheek and then the right cheek with an ADAPT (Adapting Disease Specific Outcome Measures Pilot Trial) patient and a standard Zoom video. Tracking modifications of the lip shape orientation during the exercise. The red vertical bar in the middle corresponds to the patient switching from pushing the left cheek to pushing the right cheek.



The geometric feature we used was the angle formed by the mouth and the horizontal axis. The exercise breaks the symmetry of the face, and this feature is particularly adapted to capture the one-side deformation of the cheek. The illumination figure shows only marginal change for the second part of the exercise and is therefore not very robust. One may expect however that better control of the light during the telehealth session will resolve this issue.

These techniques will not work for a subject with a moustache or beard. The shape of the face of patients with a high BMI may also impact the quality of the results. More work needs to be done on the digitalization of this specific test. As mentioned before, the depth camera would need to be highly accurate in order for the signal to be above the noise level, which is not the case with entry-level and low-cost systems.

Arm Position and Sit-to-Stand Movement

Most videos of the MGNet data set offered only partial views of the body during these exercises and showed great variability. The model (Figure 1) failed under such conditions.

Figure 10 shows a representative example of the arm angle decay due to weakening during the 120-second assessment of one of the ADAPT patients. The measurement exhibited some minor noise. We used a high-order filtering method [43] to provide a meaningful graphic to limit the noise of the method and maintain the trend that could be used for the physical examination assessment. The decay of both arms was linear and significant. It was however difficult for the medical examiner to quantify the slope or even notice it.

Figure 11 shows an example of the vertical elevation examination with respect to time for both hips, involving tracking the elevation of landmark points (23) and (24) (Figure 1B) as a function of time. From this measurement, we could not only compute acceleration and speed as indicators of muscle function but also assess the stability of the motion by measuring lateral motion in the x-coordinate.

One of the benefits of having the whole body tracked during these MG-CE evaluations is the ability to access additional information, such as the ability of the patient to stay stationary and keep their balance. While all measures are in pixels in the video itself, we recovered a good approximation of the physical dimension using the known dimension of the seat.

Figure 10. Patient performing one of the exercises of the protocol with movement of the arms. Tracking the angle of right and left arm lowering during the exercise.

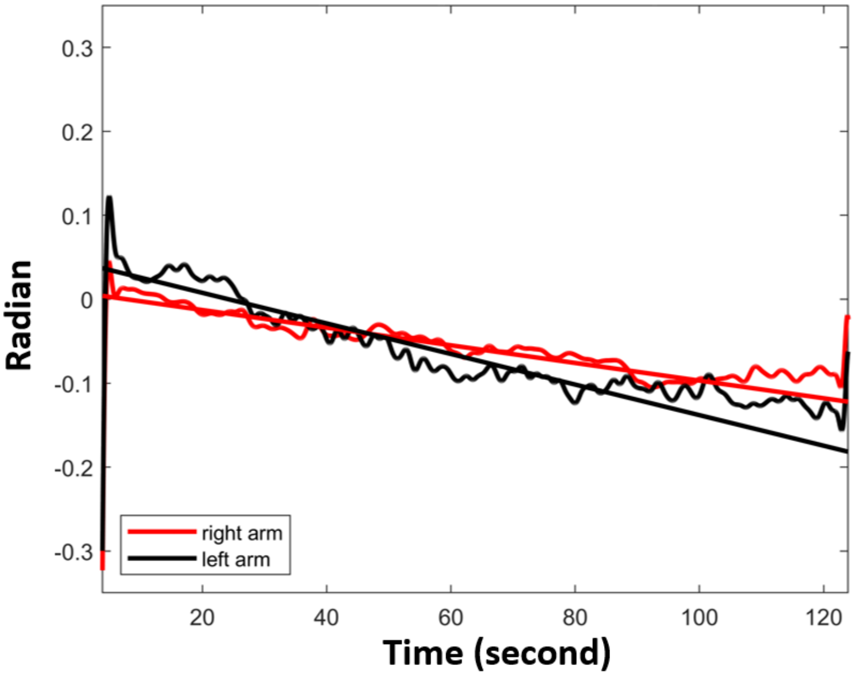
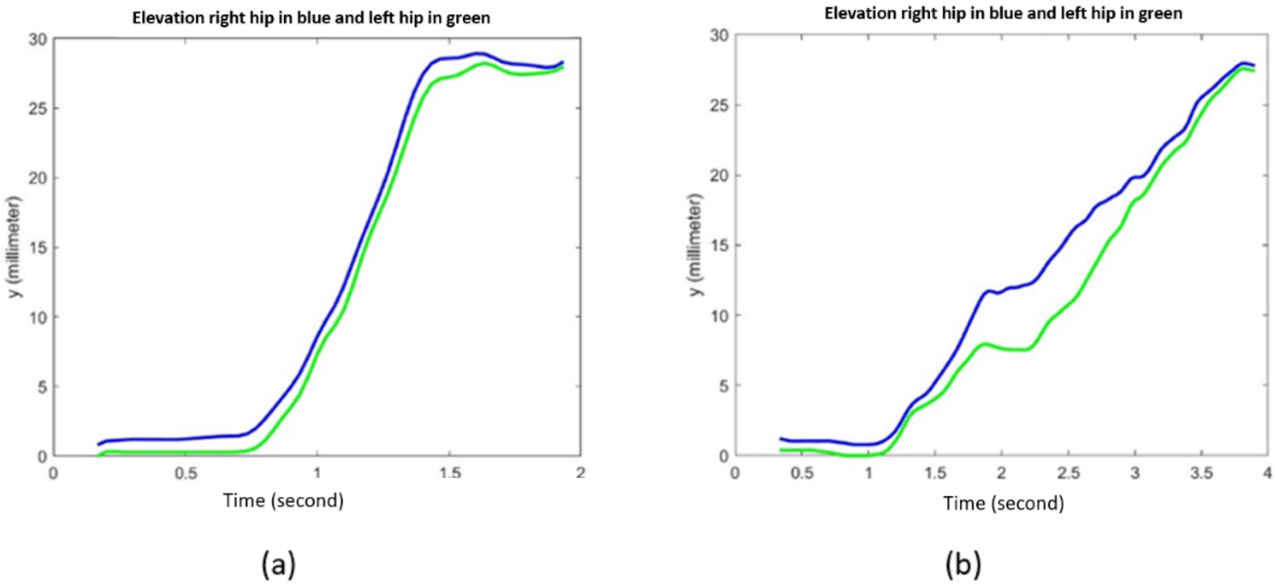


Figure 11. Elevation of both hips during the exercises: (A) normal stand up; (B) weak stand up. The right hip is indicated in blue, and the left hip is indicated in green.



Counting Exercise

This evaluation is used to assess breathing and speech quality. We used a data set of 6 patients involving 2 sessions and 9 additional healthy subject voice exercises. Audio files were cut to start and end approximately within 1 to 3 seconds of initiation and the end of counting. Based on the evaluation of patients performed by the physician according to the protocol [7], there were rarely differences between the first visit and the second controlled visit for ptosis and diplopia grading. We found however that most of the metrics described above had some variability from one visit to another, and might be considered as more sensitive metrics than the current physician examination. We will report here on our main findings with these metrics.

Instead of using the maximum number reached during the single-breath counting exercise, we used the duration of the exercise itself to grade the exercise. We found that the maximum number counted was weakly correlated with the duration of the single-breath counting exercise. The maximum number reached was indeed dependent on the speed of diction that varies greatly from one patient to another. To be more precise, one may expect that the airflow output value for the patient depends on the loudness of the voice and the pitch of the voice. In fact, we found that loudness and pitch computed with our algorithm varied dramatically from one patient to another. There had been no calibration of the microphone at the patient’s home, so we used the duration of the single-breath counting exercise as an indicator of MG severity. We suspected that a lower duration of the single-breath counting exercise is associated with more severe shortness of breath symptoms. We formulated the hypothesis that breathing difficulty might be detected by analyzing the signal in a range of frequencies concentrating on the typical breathing rate window. We used a fast Fourier transform to obtain the spectrum of the voice signal during the complete duration of the counting to 50 exercise and computed the energy of the signal restricted in the low-frequency window (5 Hz to 25 Hz). We found a weak correlation between the energy and MG severity estimated as described above (Figure 12). There were 2 outliers corresponding to 1 of the 6 patients who had severe symptoms according to the examiner annotation (Figure 12). The voice of the patient was so weak in the acquisition that the breathing signal information might have been at the noise level of the method.

We did not proceed with the identification of dysarthria per say, but looked for a relationship involving one of the generic metrics that could be computed such as spectral entropy or Teager-Kaiser energy. An example of the mean entropy of the voice signal (Figure 13) shows that this criterion is promising and may separate patients from healthy controls. Counting the number of singular picks in entropy during the examination provides better separation between patients and healthy subjects. The argument would be that an MG patient has a more monotonic voice than a healthy subject. More evaluations will be needed to confirm if entropy is a good metric. In contrast, Teager-Kaiser energy did not clearly separate MG patients.

Figure 12. Weak correlation between the duration of counting and the energy of the signal in the low-frequency bandwidth corresponding to the breathing range. Three patient sessions with voice loudness below the threshold were not counted in the fitting. The 2 outliers are from 1 patient who had a very weak voice acquisition.

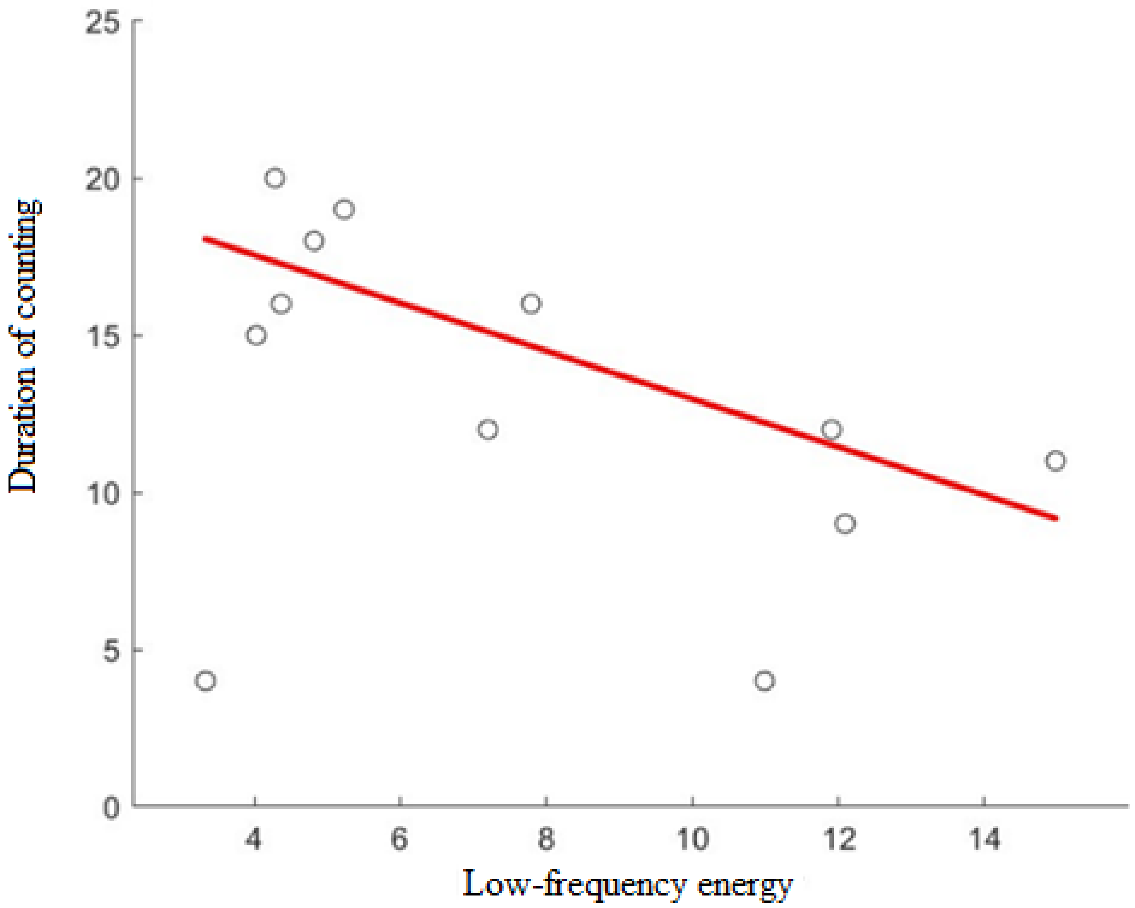
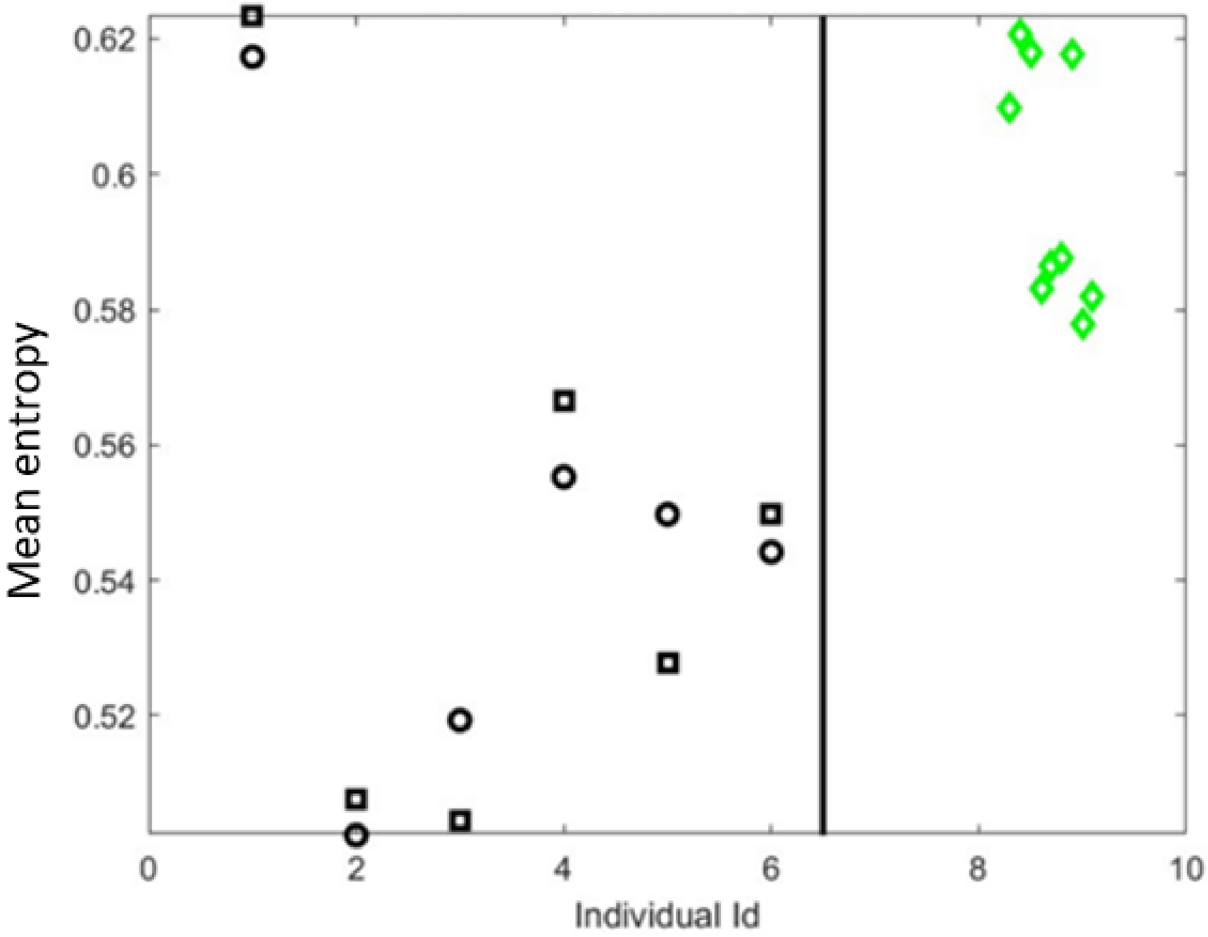


Figure 13. Mean Entropy of the voice signal during Exercise 7.



Discussion

Principal Findings

We have systematically built a series of algorithms that can automatically compute the metrics of the MG-CE, which is a standardized telemedicine physical examination for patients with MG. This effort was motivated by the increasing use of telemedicine and the appreciation of inherent limitations of presently used clinical outcome measures [44,45]. For the MG-CE, the examiner ranked the subjective observation of each examination item into categories, but this separation among classes was performed a priori and was not the result of data mining in a large population. In that context, the threshold numbers used to separate metrics, such as duration, are likely to be artificial. The data collection for these tests during a teleconsultation is tedious, repetitive, and demanding for the physician. We demonstrated a methodology, which can accelerate the data collection and provide the rational for a posteriori classification of MG severity based on a large population of patients.

Other ranking of the test might be intuitive, for example, how to define or compare difficulty in standing up. It may involve tedious motion due to muscle weakness, arthritis, or obesity. Currently, the duration of cheek deformation is not counted, but our methodology may eventually provide a precise measurement. Based on the new data set that our method provides, one should investigate further if the MG-CE classification, as well as all other categorical measures in MG, should be revisited to consider the new metrics that our algorithm can provide. In particular, the dynamic component of muscle weakness, which is a hallmark of MG and an important factor in quality of life, is not captured well by existing clinical outcome measures and not at all in routine clinical practice [46].

Our study exposed limitations in aspects of neuromuscular examination. The ability to deform the cheek does not say much about the ability to hold pressure and for how long. The cheek deformation exercise did prove to be the most difficult for achieving proper digitalization. The scoring of this exercise in the original medical protocol appeared particularly limited. We have refined mouth deformation monitoring under laboratory conditions with our Inteleclinic system to better apply computer learning techniques. Moreover, the counting exercises can be used to assess respiration function, but the number achieved does not fully equate to the severity of respiratory insufficiency.

Another challenge for our evaluation is that patients can compensate for some level of weakness and reduce the apparent severity assessed by the examination. For example, the ability to precisely compute the trajectory of the patient’s hip movement during the sit-to-stand exercise may identify if there is compensation by one leg supporting the movement more than the other. This situation could be particularly difficult for a human examiner to identify. Overall, our algorithms should give unbiased results and remove any potential subjectivity from the medical examination.

The accuracy of every computer algorithm must be constantly interrogated. Every metric should, in principle, come with an error estimate, which is not frequently the case in the current solutions, including that of the human examiner. One key component to ensure such quality of results is to control the condition of the acquisition of video and sound during the telemedicine session. With voice analysis, we would need to ensure proper calibration of the microphone at the patient's home, as well as check during the sound registration that the patient speaks with a loudness within acceptable bounds. The later can be done automatically in order to provide guidance during the examination. Similarly, the AI and computer vision aspects of the data acquisition require the patient's distance from the camera and the light condition to always be consistent with the exercise requested. This is technically feasible because the telehealth system can compute in real time the dimension in pixels of any ROI and the quality of segmentation in order to correct any obvious mistake in the data acquisition. For example, the AI model of Figure 1B that tracks the sit-to-stand exercise fails if the patient's head leaves the video frame. This kind of problem can be immediately reported to the examiner during the test. Another example is that the single-breath counting test may poorly define the initial state, speed, and loudness of speech, as counting greatly varies between patients and has an impact on breath performance evaluation.

Conclusion

Systematic digitalization and control of quality of the MG-CE are advantageous and would allow trained medical assistants to perform standardized examinations, allowing the physician to concentrate on patient questions and education instead of managing the logistics of the test. We also assessed our hardware-software "Inteleclinic" solution for telehealth consultation, which appears to be able to enhance data quality (described in a provisional patent; number 63305420; Garbey M and Joerger G, 2020). Our methods and technology would be particularly applicable to clinical trials, which are limited in requiring a large number of examiners who all perform assessments in slightly different manners. A trial could substitute present operations with a central telemedicine facility. We envision that our telehealth approach can be applied to other neuromuscular diseases beyond MG and will provide objective, reproducible, and quantitative health care assessments that go beyond the present capabilities.

Acknowledgments

The work to gather clinical data was supported in part by the MGNet, a member of the Rare Disease Clinical Research Network Consortium (RDCRN) (NIH U54 NS115054). Funding support for the Data Management Coordinating Center is provided by the National Center for Advancing Translational Sciences (NCATS) and the National Institute of Neurological Disorders and Stroke (NINDS).

Conflicts of Interest

None declared.

References

1. Feldman J, Szerencsy A, Mann D, Austrian J, Kothari U, Heo H, et al. Giving Your Electronic Health Record a Checkup After COVID-19: A Practical Framework for Reviewing Clinical Decision Support in Light of the Telemedicine Expansion. JMIR Med Inform 2021 Jan 27;9(1):e21712 [FREE Full text] [doi: 10.2196/21712] [Medline: 33400683]
2. Faget KY. The role of telehealth in decentralized clinical trials. Journal of Health Care Compliance 2021:1-5 [FREE Full text]
3. Giannotta M, Petrelli C, Pini A. Telemedicine applied to neuromuscular disorders: focus on the COVID-19 pandemic era. Acta Myol 2022 Mar;41(1):30-36 [FREE Full text] [doi: 10.36185/2532-1900-066] [Medline: 35465343]
4. Spina E, Trojsi F, Tozza S, Iovino A, Iodice R, Passaniti C, Digital Technologies, WebSocial Media Study Group of the Italian Society of Neurology (SIN). How to manage with telemedicine people with neuromuscular diseases? Neurol Sci 2021 Sep 25;42(9):3553-3559 [FREE Full text] [doi: 10.1007/s10072-021-05396-8] [Medline: 34173087]
5. Gmunder KN, Ruiz JW, Franceschi D, Suarez MM. Factors to effective telemedicine visits during the COVID-19 pandemic: Cohort study. JMIR Med Inform 2021 Aug 27;9(8):e27977 [FREE Full text] [doi: 10.2196/27977] [Medline: 34254936]
6. Alhajri N, Simsekler MCE, Alfalasi B, Alhashmi M, AlGhatrif M, Balalaa N, et al. Physicians' attitudes toward telemedicine consultations during the COVID-19 pandemic: Cross-sectional study. JMIR Med Inform 2021 Jun 01;9(6):e29251 [FREE Full text] [doi: 10.2196/29251] [Medline: 34001497]
7. Guidon AC, Muppidi S, Nowak RJ, Guptill JT, Hehir MK, Ruzhansky K, et al. Telemedicine visits in myasthenia gravis: Expert guidance and the Myasthenia Gravis Core Exam (MG-CE). Muscle Nerve 2021 Sep 07;64(3):270-276 [FREE Full text] [doi: 10.1002/mus.27260] [Medline: 33959997]
8. Clark RA, Mentiplay BF, Hough E, Pua YH. Three-dimensional cameras and skeleton pose tracking for physical function assessment: A review of uses, validity, current developments and Kinect alternatives. Gait Posture 2019 Feb;68:193-200. [doi: 10.1016/j.gaitpost.2018.11.029] [Medline: 30500731]

9. Bazarevsky V, Grishchenko I, Raveendran K, Zhu T, Zhang F, Grundmann M. BlazePose: On-device Real-time Body Pose tracking. arXiv. URL: https://arxiv.org/abs/2006.10204 [accessed 2023-03-16]
10. Jain V, Learned-Miller E. FDDB: A Benchmark for Face Detection in Unconstrained Settings. University of Massachusetts. URL: http://vis-www.cs.umass.edu/fddb/fddb.pdf [accessed 2023-03-16]
11. Kabakus AT. An experimental performance comparison of widely used face detection tools. ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal 2019;8(3):5-12 [FREE Full text] [doi: 10.14201/ADCAIJ201983512]
12. opencv. GitHub. URL: https://github.com/opencv/opencv/blob/master/data/haarcascades/haarcascade_eye.xml [accessed 2023-03-16]
13. Lienhart R, Maydt J. An extended set of Haar-like features for rapid object detection. In: Proceedings of the International Conference on Image Processing. 2002 Presented at: International Conference on Image Processing; September 22-25, 2002; Rochester, NY, USA. [doi: 10.1109/icip.2002.1038171]
14. Cao X, Wei Y, Wen F, Sun J. Face alignment by explicit shape regression. Int J Comput Vis 2013 Dec 13;107(2):177-190. [doi: 10.1007/s11263-013-0667-3]
15. Kazemi V, Sullivan J. One millisecond face alignment with an ensemble of regression trees. 2014 Presented at: IEEE Conference on Computer Vision and Pattern Recognition; June 23-28, 2014; Columbus, OH, USA. [doi: 10.1109/cvpr.2014.241]
16. Sagonas C, Antonakos E, Tzimiropoulos G, Zafeiriou S, Pantic M. 300 Faces In-The-Wild Challenge: database and results. Image and Vision Computing 2016 Mar;47:3-18. [doi: 10.1016/j.imavis.2016.01.002]
17. Putterman A. Margin reflex distance (MRD) 1, 2, and 3. Ophthalmic Plast Reconstr Surg 2012;28(4):308-311. [doi: 10.1097/IOP.0b013e3182523b7f] [Medline: 22785597]
18. Callahan MA. Surgically mismanaged ptosis associated with double elevator palsy. Arch Ophthalmol 1981 Jan 01;99(1):108-112. [doi: 10.1001/archopht.1981.03930010110014] [Medline: 7458735]
19. Struck MC, Larson JC. Surgery for supranuclear monocular elevation deficiency. Strabismus 2015 Dec 15;23(4):176-181. [doi: 10.3109/09273972.2015.1099710] [Medline: 26669423]
20. Yurdakul NS, Ugurlu S, Maden A. Surgical treatment in patients with double elevator palsy. Eur J Ophthalmol 2009 Jan 24;19(5):697-701. [doi: 10.1177/112067210901900502] [Medline: 19787584]
21. Singh HJ. Human eye tracking and related issues: A review. International Journal of Scientific and Research Publications 2012:1-9 [FREE Full text]
22. Kunka B, Kostek B. Non-intrusive infrared-free eye tracking method. 2009 Presented at: Signal Processing Algorithms, Architectures, Arrangements, and Applications SPA 2009; September 24-26, 2009; Poznan, Poland.
23. Toennies K, Behrens F, Aurnhammer M. Feasibility of Hough-transform-based iris localisation for real-time-application. 2002 Presented at: International Conference on Pattern Recognition; August 11-15, 2002; Quebec City, QC, Canada p. 1053-1056. [doi: 10.1109/icpr.2002.1048486]
24. Liu G, Wei Y, Xie Y, Li J, Qiao L, Yang J. A computer-aided system for ocular myasthenia gravis diagnosis. Tsinghua Sci Technol 2021 Oct;26(5):749-758. [doi: 10.26599/TST.2021.9010025]
25. Scharstein D, Szeliski R. High-accuracy stereo depth maps using structured light. 2003 Presented at: IEEE Computer Society Conference on Computer Vision and Pattern Recognition; June 18-20, 2003; Madison, WI, USA. [doi: 10.1109/cvpr.2003.1211354]
26. Trucco E, Verri A. Introductory Techniques for 3-D Computer Vision. Englewood Cliffs, NJ, USA: Prentice Hall; 1998.
27. Geng J. Structured-light 3D surface imaging: a tutorial. Adv Opt Photon 2011 Mar 31;3(2):128. [doi: 10.1364/aop.3.000128]
28. Siena FL, Byrom B, Watts P, Breedon P. Utilising the Intel RealSense Camera for measuring health outcomes in clinical research. J Med Syst 2018 Feb 05;42(3):53 [FREE Full text] [doi: 10.1007/s10916-018-0905-x] [Medline: 29404692]
29. Hassan S, Aronovich D, Kotzen K, Mohan D, Tal-Singer R, Simonelli P. Detecting shortness of breath remotely and accurately using smartphones and vocal biomarkers. European Respiratory Journal 2020;56:4715. [doi: 10.1183/13993003.congress-2020.4715]
30. Enderby P. Disorders of communication: dysarthria. Handb Clin Neurol 2013;110:273-281. [doi: 10.1016/B978-0-444-52901-5.00022-8] [Medline: 23312647]
31. Alam MZ, Simonetti A, Brillantino R, Tayler N, Grainge C, Siribaddana P, et al. Predicting pulmonary function from the analysis of voice: A machine learning approach. Front Digit Health 2022 Feb 8;4:750226 [FREE Full text] [doi: 10.3389/fdgth.2022.750226] [Medline: 35211691]
32. Ijitona TB, Soraghan JJ, Lowit A, Di-Caterina G, Yue H. Automatic detection of speech disorder in dysarthria using extended speech feature extraction and neural networks classification. 2017 Presented at: IET 3rd International Conference on Intelligent Signal Processing (ISP 2017); December 04-05, 2017; London p. 1-6. [doi: 10.1049/cp.2017.0360]
33. Spangler T, Vinodchandran NV, Samal A, Green JR. Fractal features for automatic detection of dysarthria. 2017 Presented at: IEEE EMBS International Conference on Biomedical & Health Informatics (BHI); February 16-19, 2017; Orlando, FL, USA p. 437-440. [doi: 10.1109/bhi.2017.7897299]

34. Mefferd AS, Lai A, Bagnato F. A first investigation of tongue, lip, and jaw movements in persons with dysarthria due to multiple sclerosis. Mult Scler Relat Disord 2019 Jan;27:188-194 [FREE Full text] [doi: 10.1016/j.msard.2018.10.116] [Medline: 30399501]
35. Warden P. Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition. arXiv. 2018. URL: https://arxiv.org/abs/1804.03209 [accessed 2023-03-26]
36. Boudraa A, Salzenstein F. Teager–Kaiser energy methods for signal and image analysis: A review. Digital Signal Processing 2018 Jul;78:338-375. [doi: 10.1016/j.dsp.2018.03.010]
37. Pan YN, Chen J, Li XL. Spectral entropy: A complementary index for rolling element bearing performance degradation assessment. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science 2008 Dec 11;223(5):1223-1231. [doi: 10.1243/09544062JMES1224]
38. Sharma V, Parey A. A review of gear fault diagnosis using various condition indicators. Procedia Engineering 2016;144:253-263. [doi: 10.1016/j.proeng.2016.05.131]
39. Shen JL, Hung JW, Lee LS. Robust entropy-based endpoint detection for speech recognition in noisy environments. In: Proceedings of the 5th International Conference on Spoken Language Processing. 1998 Presented at: 5th International Conference on Spoken Language Processing; November 30-December 4, 1998; Sydney, Australia. [doi: 10.21437/ICSLP.1998-527]
40. Vakkuri A, Yli-Hankala A, Talja P, Mustola S, Tolvanen-Laakso H, Sampson T, et al. Time-frequency balanced spectral entropy as a measure of anesthetic drug effect in central nervous system during sevoflurane, propofol, and thiopental anesthesia. Acta Anaesthesiol Scand 2004 Feb;48(2):145-153. [doi: 10.1111/j.0001-5172.2004.00323.x] [Medline: 14995935]
41. Tong JY, Sataloff RT. Respiratory function and voice: The role for airflow measures. J Voice 2022 Jul;36(4):542-553. [doi: 10.1016/j.jvoice.2020.07.019] [Medline: 32981809]
42. Horaud R, Hansard M, Evangelidis G, Ménier C. An overview of depth cameras and range scanners based on time-of-flight technologies. Machine Vision and Applications 2016 Jun 16;27(7):1005-1020 [FREE Full text] [doi: 10.1007/s00138-016-0784-4]
43. Garbey M, Sun N, Merla A, Pavlidis I. Contact-free measurement of cardiac pulse based on the analysis of thermal imagery. IEEE Trans Biomed Eng 2007 Aug;54(8):1418-1426. [doi: 10.1109/TBME.2007.891930] [Medline: 17694862]
44. McPherson T, Aban I, Duda PW, Farzaneh-Far R, Wolfe GI, Kaminski HJ, of the MGTX Study Group. Correlation of Quantitative Myasthenia Gravis and Myasthenia Gravis Activities of Daily Living scales in the MGTX study. Muscle Nerve 2020 Aug 04;62(2):261-266 [FREE Full text] [doi: 10.1002/mus.26910] [Medline: 32369631]
45. Cleanthous S, Mork A, Regnault A, Cano S, Kaminski HJ, Morel T. Development of the Myasthenia Gravis (MG) Symptoms PRO: a case study of a patient-centred outcome measure in rare disease. Orphanet J Rare Dis 2021 Oct 30;16(1):457 [FREE Full text] [doi: 10.1186/s13023-021-02064-0] [Medline: 34717694]
46. Benatar M, Cutter G, Kaminski HJ. The best and worst of times in therapy development for myasthenia gravis. Muscle Nerve 2023 Jan 12;67(1):12-16. [doi: 10.1002/mus.27742] [Medline: 36321730]

Abbreviations

ADAPT: Adapting Disease Specific Outcome Measures Pilot Trial
AI: artificial intelligence
MG: myasthenia gravis
MG-CE: Myasthenia Gravis Core Examination
ROI: region of interest

Edited by P Kubben; submitted 12.10.22; peer-reviewed by V Ochs, M Banf, PP Zhao; comments to author 30.12.22; revised version received 27.02.23; accepted 02.03.23; published 19.04.23.

Please cite as:
Garbey M, Joerger G, Lesport Q, Girma H, McNett S, Abu-Rub M, Kaminski H
A Digital Telehealth System to Compute Myasthenia Gravis Core Examination Metrics: Exploratory Cohort Study
JMIR Neurotech 2023;2:e43387
URL: https://neuro.jmir.org/2023/1/e43387
doi:10.2196/43387
PMID:37435094

Original Paper

A Novel System to Monitor Tic Attacks for Tourette Syndrome Using Machine Learning and Wearable Technology: Preliminary Survey Study and Proposal for a New Sensing Device

Agni Rajinikanth[1*]; Davis Kevin Clark[2*]; Marianna Evangelia Kapsetaki[3], MD, MSc, PhD

[1]Rutgers Preparatory School, Somerset, NJ, United States

[2]St George's School, Vancouver, BC, Canada

[3]Faculty of Life Sciences, Division of Biosciences, University College London, London, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Marianna Evangelia Kapsetaki, MD, MSc, PhD
Faculty of Life Sciences
Division of Biosciences
University College London
Gower Street
London, WC1E 6BT
United Kingdom
Phone: 44 07392970494
Email: marianna.kapsetaki.15@ucl.ac.uk

Abstract

Background: Tourette syndrome is a neurological disorder that is characterized by repeated unintentional physical movement and vocal sounds, better known as tics. Cases of mild Tourette can have tics numerous times throughout the day, while severe cases may have tics every 5 to 10 seconds. At certain times, typically during high levels of stress, tics become chained in an incessant, continuous fashion—this is known as a tic attack. Tic attacks incapacitate the patient, rendering it difficult for them to move, perform daily actions, and even communicate with others. Caretakers—usually guardians, family members, or nurses—can help reduce the time tic attacks last with their presence and by providing emotional support to the patient.

Objective: We describe TSBand, a wearable wristband that uses machine learning algorithms and a variety of sensors to monitor for tic attacks and notify caretakers when an attack occurs.

Methods: We conducted a research survey with 70 Tourette patients to determine the usability and functionality of TSBand; internal review board approval was not required.

Results: This study has resulted in a smart wristband prototype that costs US $62.74; it uses movement, heart rate, sweat, and body temperature to detect tic attacks using a hybrid local outlier factoring and regression algorithm. An audio tic attack detection mechanism is also included, using recurrent neural networks, and a manually activated backup button and backup audio mechanism are fitted to alert caretakers on the personalized companion app.

Conclusions: TSBand enables the caretaker to provide support faster and prevent excessive self-harm or injury during the attack. It is an affordable and effective solution, solving a problem that many Tourette patients, often children, face. This study has not had the opportunity to test TSBand with any Tourette patients, and we aim to perform rigorous testing and analysis after grant funding is secured.

(JMIR Neurotech 2023;2:e43351) doi:10.2196/43351

KEYWORDS

Tourette syndrome; neurological diseases; tic attacks; wearable technology; movement disorders; tremor monitoring; biosensing technology; automatic tic detection

Introduction

Background

An estimated 350,000 to 450,000 children and adults in the United States alone have Tourette syndrome (TS), a neurological disorder that causes involuntary movements or vocal sounds known as tics, and about 1 million children and adults in the United States have other persistent tic disorders [1]. Tics are usually most prevalent in adolescents, and the severity of the disorder tends to decrease with age, making children with TS a community that needs medical and technological innovation. Research has shown that around 1% to 3% of children in mainstream schools are affected by TS [2]. Specific tics and their severity vary from person to person (eg, hand gestures vs whole-arm movements), but nearly all tics can be categorized as uncontrollable movements or audible noises. Specific tics can both develop and disappear at random. People who have TS often also experience symptoms of obsessive-compulsive disorder and attention deficit hyperactivity disorder in addition to tics [3].

Under certain situations of high stress or anxiety, a tic attack may occur, an event characterized by nonstop tics of higher severity that often incapacitate the physical motion and verbal communication of the patient. The frequency of tic attacks can range from occasional (less than once a month) to daily; attacks can vary dramatically from person to person, and they can last from a few minutes to several hours [4]. There is no cure for tics or tic attacks, but the severity of tic attacks can potentially be reduced by medications such as neuroleptics or fluphenazine [5]. Providing medicine and emotional support is typically the job of caretakers, such as parents, guardians, and nurses, but constant surveillance is not always possible.

Product/Need

Few products currently exist for patients with TS. There are some products designed to mitigate tics through slight electrical pulses, but these do not help solve the problem of detection and alerting someone for help. TSBand is a wearable wrist device that aims to automatically detect when a tic attack is occurring and then alert a nearby caretaker with a mobile notification. The device serves as a bridge between caretakers and patients, allowing those in need to receive help without physically needing to call for it. This can potentially decrease the span of a tic attack from hours to just a few minutes with help from the caretaker, limiting the time of struggle and the possibility of self-harm. This is helpful for adults but is especially helpful for children, who are more inexperienced with handling attacks, as caretakers can be notified instantly via TSBand and help the child by quickly practicing behavioral techniques to calm the child down, reducing the time of the attack. Additionally, studies have shown that medical technologies integrating mobile phones and wearable sensors need improvement and further scrutiny [6]. This paper proposes the development of TSBand to detect tic attacks and notify caretakers.

Methods

Sensors Used for Detection

To facilitate the detection mechanism of the device, common patterns that people with TS usually exhibit must be analyzed. TSBand is equipped with a triaxial gyroscope and accelerometer, a pulse sensor, and a body temperature and humidity sensor. When tic attacks occur, patients usually exhibit uncontrollable and repetitive upper limb movement. The gyroscope and accelerometer help in providing data to perform the necessary computations to determine tic attacks from arm movement. The accelerometer is used to gather data on the speed and acceleration of the arm during tic attacks, while the gyroscope is used to provide information on roll, pitch, and yaw to account for the variation in angle and thus calculate the angular speed and acceleration. Furthermore, a microphone is used to assist with the detection and account for vocal tics as well (described in the Audio Analysis section).

In one previously published paper, researchers described a detailed clinical study showing that high blood pressure, stress, and an increased heart rate during tic attacks are all signs and symptoms commonly reported by patients with TS [7]. By using these vital signs in combination with movement, the tic detection algorithm can be enhanced for a higher degree of accuracy. To measure these factors and symptoms, TSBand includes a pulse sensor that actively measures the heart rate of the user. As temperature and sweat can increase because of both movement and stress, the body temperature and humidity sensor monitors for increasing fluctuations in these 2 vital signs. Combined, these sensors are used to detect changes in the vital signs of patients with TS and serve as another method to monitor for tic attacks.

Machine Learning

Because tics and tic attacks vary from person to person, the detection model cannot be generalized, rather, it needs to be custom-tailored to each individual. To monitor for and detect tic attacks, TSBand requires a 2-day calibration period in which the user’s movement patterns are stored to determine the tic patterns of the individual patient. The wristband will need to be worn both during the day and at night for calibration, as severe tics persist throughout the night and during sleep in some people. The device calibration is used primarily to detect the regular movement patterns of the user to determine a common baseline and allow the wristband to adapt to the unique baseline of the patient. There does not need to be any tic attack during the calibration period for the algorithm to work as intended; if there are any tic attacks, the algorithm will likely classify them as outliers based on the other data collected during this period. Using the local outlier factor (LOF) algorithm, TSBand separates tic attacks from normal, everyday movements. The variables that the LOF model analyzes are speed and acceleration; these were chosen because traditional models such as random forest and regression are less accurate at detecting tic attacks with these parameters. LOF models use a vast amount of data to form clusters and, when optimized, can be used to find outliers in the data set.

LOF uses unsupervised machine learning techniques that use the density distribution and standard deviation of data points to detect outliers in the data. Considering reachability distance and the density of the data values, TSBand uses a fine-tuned algorithm to limit the nearest neighbor's parameter to ensure that clusters will not be formed near a tic attack, helping reduce the number of false positives:

By obtaining data for 2 days to train the model, TSBand can uniquely tailor the detection system to each individual user for ease of customization and personalization. The 2-day calibration period also helps limit both the number of false positives that occur and the number of clusters around tic attacks, thus increasing the overall accuracy of the model. The data during the 2-day calibration are the only data used to determine outliers; all later data are not recorded or stored for detecting tic attacks.

In addition to the outlier algorithm, a linear regression model—used to find trends over data points—was implemented as a measure to enhance the detection process. The regression model measures data from the last hour to search for dramatic increases and changes in vital signs parsed through a certain threshold value. The threshold is passed when the rate of change at the data point is greater than or equal to 4/3 or less than or equal to –4/3. The ±4/3 threshold value was chosen because vital signs such as heart rate are usually stable, and when variations occur, the changes are dramatically visible, allowing ±4/3 to be a sensible rate of change to use. However, as this threshold range was chosen through a reasonable estimate, it is subject to change after refinement of the algorithms with patient testing:

$$y=b_0+b_1x_1$$

$$\text{Threshold Range: } -43 \leq y' \leq 43$$

The algorithm uses this principle to determine whether vital signs such as stress, sweat, body temperature, and heart rate are increasing or are predicted to increase for the wristband bearer. This mechanism can be used to find general trends in vital signs that are clear indicators of tic attacks. The combined algorithms output a percentage that conveys the likeliness of a tic attack occurring. Although variations in heart rate, temperature, and humidity might signify an attack, these parameters are not always correlated with a tic attack, while movements are. To place greater emphasis on movement, 70% of the final prediction relies on this parameter, while the other values equally share the remaining 30%. These ratios are subject to change after testing among real patients with TS is completed and data are aggregated:

$$P_{Total}=0.7P_{Movement}+0.1P_{HeartRate}+0.1P_{Temperature}+0.1P_{Sweat/Humidity}$$

Once the probability of the total attack (P_{Total}) is calculated, any result over 80% will register an alert and notify the caretaker. When the threshold is passed for the linear regression models, the attack is considered "true" (ie, P_{Vital}=1), and the model will stop analyzing new data until the patient can confirm the attack is over on the mobile app or there has not been a tic attack registered (ie, P_{Total}>80%) for an hour. When either of these conditions is met, the model will then assess the data from the last hour and continue monitoring new vital sign data from thenceforth. The value of 80% for P_{Total} is a default value and can be changed manually by the caretaker and the patient to suit their specific needs through our companion mobile app. This feature was added in consideration of people who may not experience large fluctuations in health vital signs during tic attacks; however, the value should remain above 70% so as to not rely only on movement.

Band Design and Schematics

The wristband has a sleek design to enhance the experience of the user. The band is powered using an ESP32 Wi-Fi chip (Espressif Systems), which is used to send data to a live-updated database hosted on Firebase (Google Inc). To program the ESP32, the Arduino C++ language was used. All the sensor data collected from the user are securely stored in Google's real-time Firebase server, where the algorithms, hosted on Google Cloud, execute decisions and determine whether a tic attack is occurring. The wristband may also potentially be able to predict tic attacks before they occur. Although movement cannot be used as the sole factor for prediction because stress is often a precursor to tic attacks, increased heart rate and body temperature can potentially be used as a predictive mechanism. When a tic attack is detected, an alert is immediately sent to the caretaker's phone through the companion mobile app. A buzzer is built into the band that sounds to act as an audio cue to notify the user that an alert has been sent.

A backup method exists in case a tic attack occurs but the algorithm does not detect the attack; this allows the user to manually send an alert to the caretaker. A button on the device can be pressed and held for 5 seconds to send a request to the caretaker for help. A potential issue with the button is that a user could develop a motor tic and press the button without the intention to call a caretaker for help. Therefore, the button must be held for 5 seconds to give the user time to reconsider the action and let go of the button if necessary. Additionally, a cancellation feature was implemented in case a faulty alert was sent for any reason. If the user notices the buzzer going off, they can press the button 3 times (or more) in the span of 1 second to cancel the request to the caretaker. Figures 1 and 2 show a computer-aided design model of the prototype, and Figure 3 displays the prototype being worn. Because of budget constraints, we were not able to scale down the size of our central processing unit, which is why wires can be seen coming out of the device in Figure 3.

Figure 1. Computer-aided design model of the TSBand prototype (bottom view).

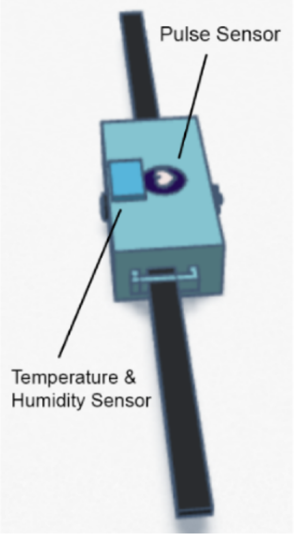


Figure 2. Computer-aided design model of the TSBand prototype (bottom view inside the compartment).

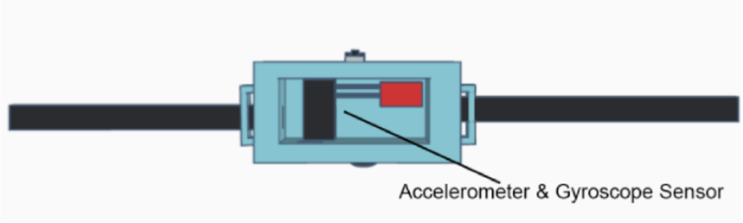
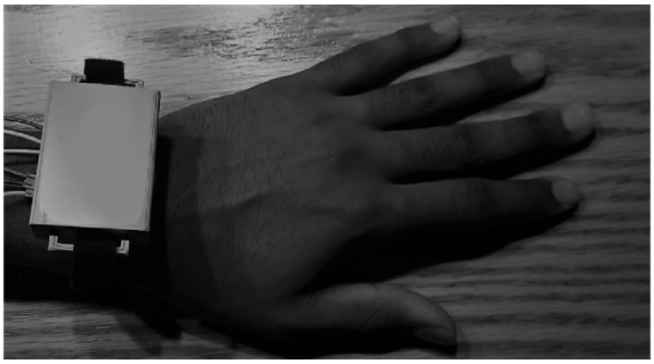


Figure 3. TSBand prototype being worn on the wrist.



Audio Analysis

Tics do not only occur in the form of motor movements—they can also appear as verbal noises. Vocal tics are common and words or specific phrases are typically blurted out at random moments. During tic attacks, patients with TS usually repeat those specific phrases or other sounds over a very short period. On the device, audio files obtained through a microphone are passed through a high-pass filter to limit background noise. By converting speech to text and analyzing patterns in speech over a certain time interval, the wristband is able to detect and process vocal tics. The model used to train the speech recognition software is a recurrent neural network model that uses a long short-term memory algorithm. The data were compared with the Google Speech data set and had an accuracy rate of 92%. To limit the amount of error in this trial, the algorithm was designed to listen for and catch the frequency of a specific word repeated over a period of 10 seconds. If the repetition occurs for a duration that is longer than average for the user, a notification is sent to the caretaker. This audio detection mechanism is independent of the physical detection mechanism; if one detects an attack while the other does not, an alert will still be sent.

Furthermore, a backup vocal alert system is also included in the band as a supplement to the audio analysis. Much like smart home devices such as Amazon Alexa or Google Home Mini, a prerecorded audio cue phrase can be set to instantly send an alert notification to the caretaker. For example, if the cue phrase is “notify TSBand,” a notification is sent to the caretaker alerting them of an ongoing attack any time the wristband detects this specific phrase. The phrase only needs to be said once, which is why it is important to pick a specific, uncommon phrase. This is intended for an absolute emergency backup case, such as when the automatic detection and repeated phrase detection fail and the backup button is not pressable due to an excessive number of tics that limit movement. There is potential that the user develops a regular verbal tic for this exact phrase, causing the alert to be triggered when there is no tic attack. Thus, the audio phrase that triggers the alert can be changed and prerecorded again through the mobile app to preserve the use of this feature. Again, the buzzer will sound to notify the user that an alert has been sent, and the cancellation feature is still valid in this scenario if it is necessary to cancel the request. The flowchart in Figure 4 summarizes the process. Figure 5 shows the vocalization diagram of the phrase “notify TSBand” in a signal diagram. Figure 6 shows the vocalization in a spectrum diagram. Figures 7 and 8 show the audio vocalization under a spectrogram and a mel spectrogram, respectively.

Figure 4. Flowchart diagram and schematics of the TSBand in operation. LOF: local outlier factor.

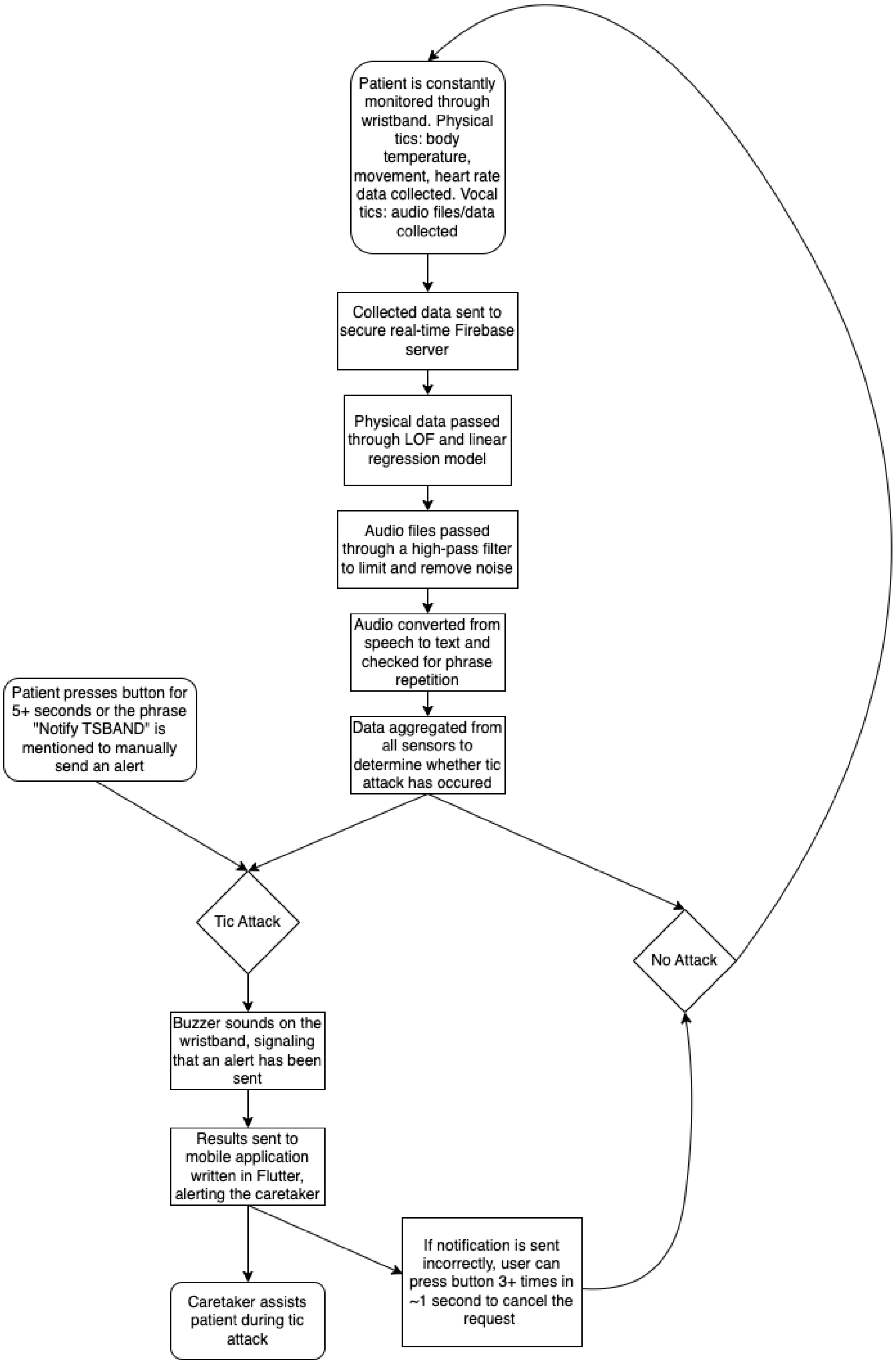

Patient is constantly monitored through wristband. Physical tics: body temperature, movement, heart rate data collected. Vocal tics: audio files/data collected
Collected data sent to secure real-time Firebase server
Physical data passed through LOF and linear regression model
Audio files passed through a high-pass filter to limit and remove noise
Audio converted from speech to text and checked for phrase repetition
Data aggregated from all sensors to determine whether tic attack has occured
Patient presses button for 5+ seconds or the phrase "Notify TSBAND" is mentioned to manually send an alert
Tic Attack
No Attack
Buzzer sounds on the wristband, signaling that an alert has been sent
Results sent to mobile application written in Flutter, alerting the caretaker
Caretaker assists patient during tic attack
If notification is sent incorrectly, user can press button 3+ times in ~1 second to cancel the request

Figure 5. “Notify TSBand” vocalization signal diagram.

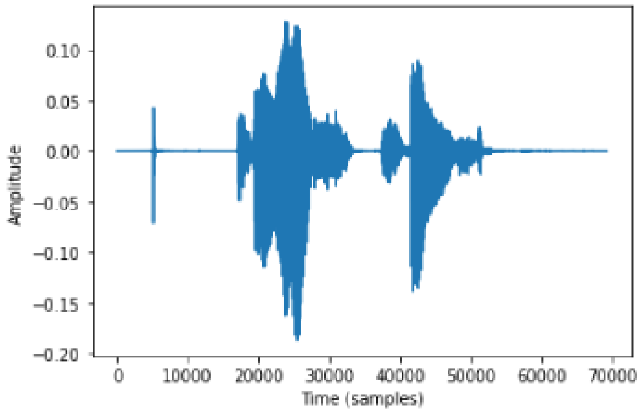


Figure 6. TSBand spectrum diagram.

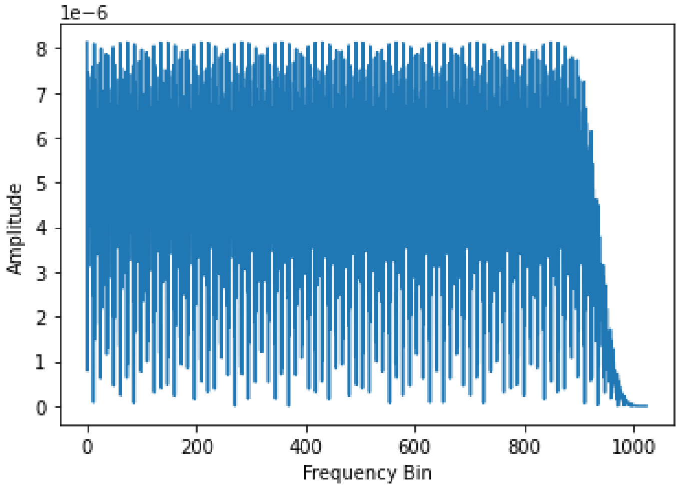


Figure 7. TSBand spectrogram diagram.

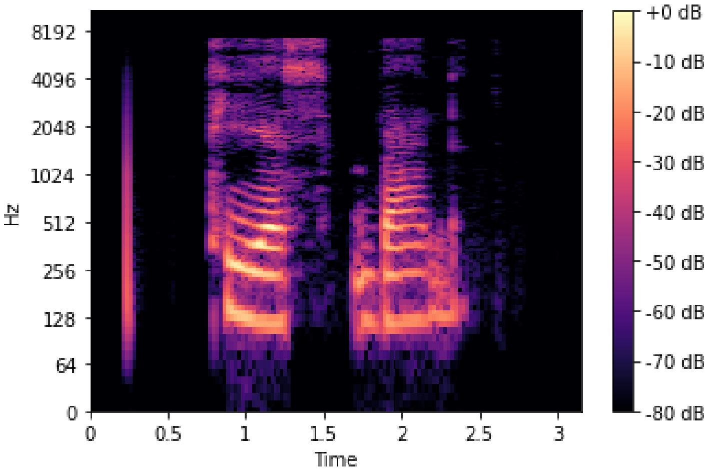
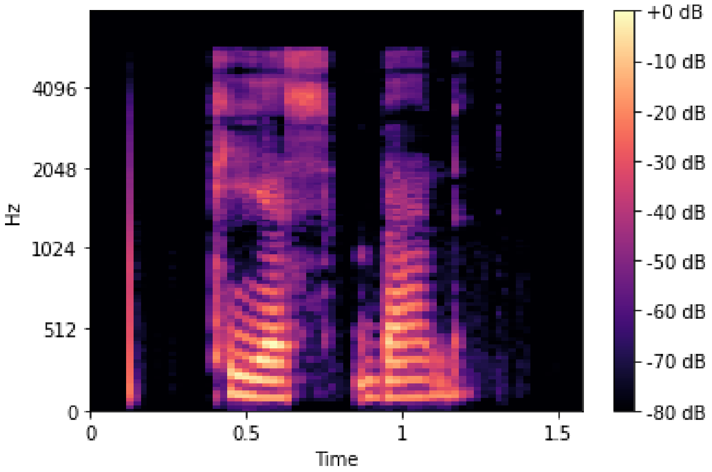


Figure 8. TSBand mel spectrogram diagram.

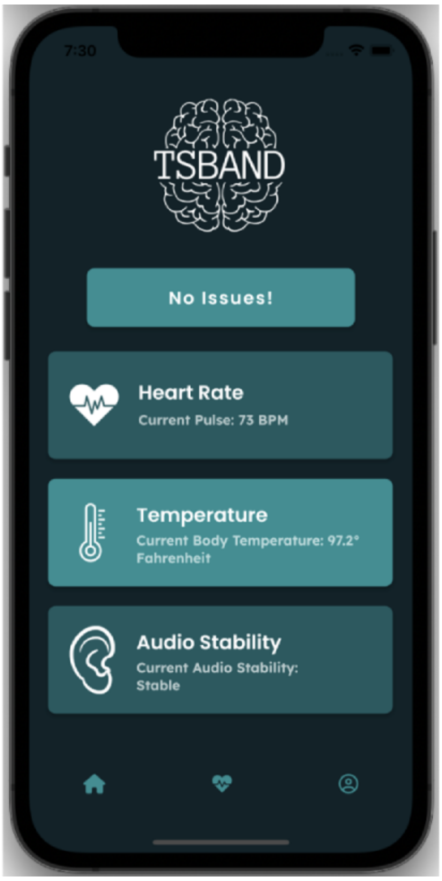


Mobile App

TSBand has a companion mobile app to send notifications to caretakers. After downloading the app, a caretaker or a group of caretakers can pair with the wristband—this can be done as each band has a unique hash code associated with it that can be used to accurately pair up with the app. The app displays certain health vitals, namely heart rate and body temperature, and an audio monitor that shows the frequency of words and reports if there is anything abnormal. The value is shown as either “stable” or “unstable” depending on the count of the repetition of words (it is determined the same way as described in the Audio Analysis section). Heart rate and body temperature are read directly from Firebase and are displayed live on the app.

As shown in Figure 9, typical activity will result in the bar at the top showing the status “no issues.” If a tic attack is detected by the algorithm, a push notification (pop-up notification) will appear on the phone to alert the caretaker. Upon opening the app, as shown in Figure 10, the “no issues” bar at the bottom will have changed to the “possible attack” status. The app will not send any new push notifications on the phone while the status is “possible attack.” If the situation has been handled, the “possible attack” button can be pressed on the app to revert the status back to the normal “no issues” value, and at that point, the app will once again send push notifications in case another tic attack happens. Pressing this button will lead to the restart of the regression models.

Figure 9. Mobile app when tic attack is not occurring.



7:30

TSBAND

No Issues!

Heart Rate

Current Pulse: 73 BPM

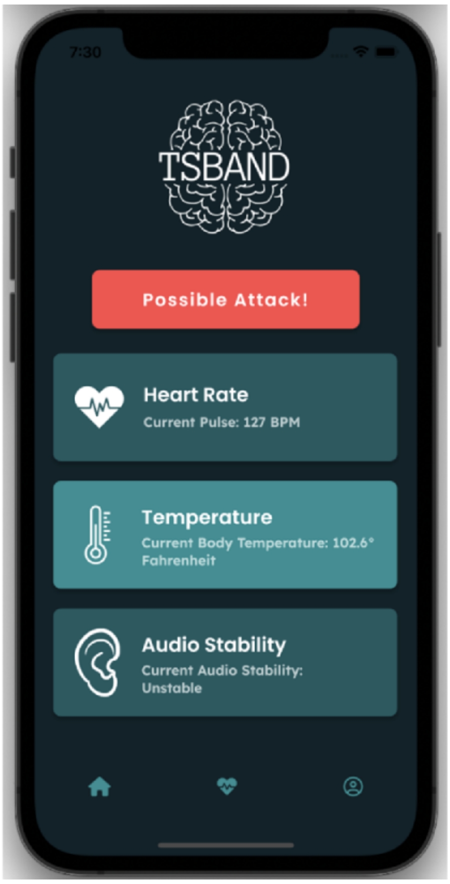
Temperature

Current Body Temperature: 97.2° Fahrenheit

Audio Stability

Current Audio Stability: Stable

Figure 10. Mobile app when tic attack is occurring.



7:30

TSBAND

Possible Attack!

Heart Rate

Current Pulse: 127 BPM

Temperature

Current Body Temperature: 102.6° Fahrenheit

Audio Stability

Current Audio Stability: Unstable

Additional Procedure

We have not yet had the opportunity to evaluate or test our device with physical participants or patients due to funding and time constraints. This paper focuses on introducing the technology behind the device and analyzes its potential introduction into the market. We have performed said analysis through a preliminary research survey to obtain perspective on the potential of the device. This survey—mainly consisting of questions inquiring about the personal experiences of tic attacks among patients with TS and asking for feedback and new features for the wristband—was filled out by 70 people with TS and is detailed in the sections below.

Ethical Considerations

The survey did not contain any questions that would induce physiological stress or anxiety and collected only nonsensitive data; thus, we did not need ethics approval. A link to the quantitative questions and results is available in the Data Availability section.

Results

Finances

For the economic and financial analysis, we determined the current cost of our prototype in US dollars (Table 1), but since parts and materials are subject to change, we did not predict a potential final price. In the future, since the current prototype is bulky, we plan to further size it down. We hypothesize that the overall prototype cost will be reduced with mass manufacturing, but it is difficult to estimate the price of a smaller central processing unit and other final parts, as there are many variables involved with finding new components.

With a final cost of $62.74, TSBand will be suitable for potential users. Furthermore, on the mobile app side of this service, it is estimated that server upkeep and fees for application programming interfaces used for the software behind the device will cost an additional US $5.00 per user per year. If the wristband is mass manufactured with a custom-printed circuit board, both the cost and size of TSBand can be reduced.

Table 1. Prototype cost structure.

Components	Cost (US $)
ESP32 feather	9.95
Pulse sensor	10.99
Humidity and temperature sensor (DHT11)	5.00
Gyroscope and accelerometer sensor (GY-521)	6.00
Audio microphone (AOM-5035L-HD3-R; PUI Audio)	4.45
Buzzer (WT-1614T; Soberton Inc)	1.25
Battery source (130 maH)	19.99
Velcro	0.11
3D printing	5.00
Total cost	62.74

Research Survey

The questionnaire was an online survey consisting of 11 questions, both qualitative and quantitative, directly addressing the TS community. The survey was a completely blinded review survey performed online on social media platforms that posed both quantitative and open-ended questions. To ensure anonymity and to abide by Health Insurance Portability and Accountability Act and the General Data Protection Regulation, data on the participants were not tracked so that responses would be more transparent. The collected data are held securely.

Survey Findings

Notably, 64 of 70 (91%) respondents said that they struggle to communicate with caretakers when having a tic attack, and 67 of 70 (96%) respondents believed that stress and anxiety are contributing causes of tic attacks. In qualitative responses, one common comment was that an alert could help bring forth a trusted caretaker who could help calm the patient down and ensure they do not cause self-harm. Some mentioned that the material of the wristband is also extremely important due to skin irritation that some face. In addition, the structural integrity of the band must be strong to ensure longevity and so the band does not break when exposed to movement during an attack. We found that 30 of 70 (43%) respondents believed that having a person near them would help them cope with the attack, while 13 of 70 (19%) respondents said it would only help if the figure was trusted. Overall, 53 of 70 (76%) respondents stated that such a device would be useful for them. Those who stated that the device would not be beneficial to them were adults who did not have caretakers or did not use medication, as their tic attacks were of shorter duration.

Discussion

Limitations

Certain limitations still exist with wristband detection. For example, it is difficult to differentiate exercise from a tic attack, as both include excessive movement of the body. If the machine-learning model is not trained with exercise data, it may consider simple tasks like running to be a tic attack, but if the model is trained with exercises like running, it may potentially consider a tic attack to be another aspect of a daily routine and not trigger an alert. A model trained without exercise would force the user to take off the wristband during physical activity and leave them vulnerable to a sudden tic attack with no automatic detection. Ideally, someone else would supervise them in this scenario, but adding measures such as the ability to turn off the automatic detection algorithm of the band while still keeping backup elements active would be optimal for users when exercising. Alternatively, if the model is trained with exercise, there is potential that with vast data on previous patterns of the patient and an exercise mode built into the wristband, the algorithm can stay powered on during activities like running and assess differences between tic attacks and exercise. Creating an exercise mode would need much experimentation and previous data on the user's personal tic attacks to accurately differentiate between physical activity and tic attacks. We also recognize that the survey of 70 respondents does not represent the entire population of people with TS.

Comparisons to Similar Wristbands

Similar studies have been conducted for tic movements and other tremors, such as those caused by Parkinson disease (Table 2).

Table 2. Comparison with similar wearable technologies.

Paper	Project purpose	Measure of movement	Measure of vital signs	Recognition method	Audio analysis
This paper	Monitor for and detect tic attacks	Accelerometer, gyroscope	Heart rate, body temperature, sweat	Local outlier factor, regression, threshold	Long short-term memory
Bernabei et al [8]	Detect tic movements	Accelerometer	None	Time variant threshold	None
Fraiwan et al [9]	Monitor Parkinson tremors	Mobile phone accelerometer	None	Deep-learning neural networks	None
Kim et al [10]	Differentiate upper limb tremors from regular movement	Accelerometer, gyroscope	None	Statistical pattern recognition	None
Cole et al [11]	Detect tremor and dyskinesia in Parkinson patients	Accelerometer	Surface electromyography	Support vector machine and Markov models	None

The most similar wristband to TSBand is the band developed by Bernabei et al [8], which uses a triaxial accelerometer to determine tic movements. It corroborates video analysis with the wristband data to detect tic activities, but it cannot detect tic attacks. Furthermore, it does not use any other factors, such as angular acceleration, vital signs fluctuations, or audio analysis to determine tic movements, and is not tailored to each individual that uses it. Kim et al [10] applied an inertial measurement unit on a worn device to differentiate typical daily movement from tremors in the upper limbs. Artificial tremors were generated, and the data were passed through statistical pattern recognition to determine if a reading was a tremor or standard activity. The studies conducted by Fraiwan et al [9] and Cole et al [11] both monitored for tremors in Parkinson patients, with the former using the accelerometer in a smartphone and the latter using a separate accelerometer sensor. Both used neural networks for the detection of tremors, but Cole et al also used surface electromyography for a higher rate of accuracy. The machine-learning techniques used in both studies were used only for tremor detection accuracy and did not incorporate customizability for individual patients, as all Parkinson tremors are quite similar. The machine learning models used in TSBand, however, are intended to adapt to each person's unique tics and tic attacks for better detection purposes. This wristband serves as a complete solution that could immediately be used by those in need, with backup and emergency situations accounted for in its design. Because of the personalization TSBand provides, it could also potentially be used for similar conditions with similar symptoms, including myoclonus, tremors, chorea, athetosis, dystonia, akathisia movements, paroxysmal dyskinesias, and ballistic movements.

Future Work

Initially, TSBand was designed to be a thin bracelet strapped around the wrist, but because the sensors took up a large amount of space, a storage compartment was added to hold them. However, this storage compartment now gives the potential to add a light emitting diode screen, among other options, on top of the compartment, which could be used to enhance the user experience. Based on survey responses, certain features could be added to the band and mobile app to improve user benefit. One common suggestion was a GPS system to determine the location of the attack to better notify the caretakers, as well as a screen on the wristband itself that shows how far away the caretaker is to give an estimate of how long it would take for them to arrive. Another feature was for the app to not only notify caretakers, but also inform teachers or coworkers that a tic attack is occurring. Several individuals who were students mentioned that they were embarrassed to raise their hands to tell teachers that they needed to leave the classroom during a tic attack, so this mechanism would effectively avoid said embarrassment by providing a tacit signal to the teacher that the user needs some time and space away from class to calm down.

We would like to test the current accuracy of the band among patients with TS to determine specific values for the regression model, see which areas need to be improved on, and understand how new features can be incorporated (eg, the exercise mode). A common request from the survey was a way to also send information on the severity of the attack so as not to worry the caretaker if the attack is only mild. Adding this would be extremely helpful but would require a great deal of testing and experimentation to measure severity. We are also looking to enhance our algorithm to potentially predict tic attacks before they occur based on previous behavior patterns and also scale down the size and cost of TSBand through a custom-printed circuit board to fit better on the wrist. In terms of sensors, upgrading the band to include more accurate and precise sensors, such as galvanic skin response and blood pressure sensors, would allow it to detect tic attacks more effectively and accurately. We filed a provisional patent with the United States Patent and Trademark Office on October 8, 2022, and we hope to finalize these upgrades and finish conducting this research by September 2023.

Conclusion

In this work, machine learning and audio analysis are used to detect tic attacks. LOF is used to individualize the detection process, with a regression model used to ensure greater accuracy. To reiterate, we were not able to test these machine learning components with actual patients due to funding and other concerns; we aim to do this in the future. Audio analysis is used to check for repeated phrases over a time interval to detect vocal tics. In addition to these features, we also have an emergency audio backup method along with our standard backup button. A recurrent neural network was trained on a Google Speech data set to obtain a 92% speech accuracy rate. Within the next year, we hope to add new features, scale down the size of the device, add new sensors, and test the wristband among patients with TS. We conducted a survey involving 70 participants with TS to gather data on commonalities in tic attacks, the efficacy of having a caretaker nearby, and qualitative feedback to help determine limitations and solicit suggestions on the development of the wristband. The benefit of a platform that connects patients at risk with caretakers is not limited to those with TS. The elderly or people at risk of seizures can benefit from having an alert automatically sent to a caretaker when they are not able to physically move or request help themselves. Any person that requires an easy method of requesting assistance can use this wristband and receive help faster. This wristband has vast potential and can not only be applied to TS but used in many different fields of health care and for patient treatment.

Acknowledgments

We would like to thank the individuals who completed our survey.

Data Availability

The data sets generated during and/or analyzed during this study are available in the Open Science Framework repository [12].

Conflicts of Interest

DKC, AR, and MEK were involved in the production and development of software and hardware for the device and filed a provisional patent with the United States Patent and Trademark Office on October 8, 2022.

References

1. Tinker SC, Bitsko RH, Danielson ML, Newsome K, Kaminski JW. Estimating the number of people with Tourette syndrome and persistent tic disorder in the United States. Psychiatry Res 2022 Aug;314:114684. [doi: 10.1016/j.psychres.2022.114684] [Medline: 35724469]
2. Robertson MM. Diagnosing Tourette syndrome: is it a common disorder? J Psychosom Res 2003 Jul;55(1):3-6. [doi: 10.1016/s0022-3999(02)00580-9] [Medline: 12842225]
3. Chowdhury U, Heyman I. Tourette's syndrome in children. BMJ 2004 Dec 11;329(7479):1356-1357 [FREE Full text] [doi: 10.1136/bmj.329.7479.1356] [Medline: 15591541]
4. Robinson S, Hedderly T. Novel psychological formulation and treatment of "Tic Attacks" in Tourette syndrome. Front Pediatr 2016;4:46 [FREE Full text] [doi: 10.3389/fped.2016.00046] [Medline: 27242975]
5. Eddy CM, Rickards HE, Cavanna AE. Treatment strategies for tics in Tourette syndrome. Ther Adv Neurol Disord 2011 Jan;4(1):25-45 [FREE Full text] [doi: 10.1177/1756285610390261] [Medline: 21339906]
6. Seppälä J, De Vita I, Jämsä T, Miettunen J, Isohanni M, Rubinstein K, M-RESIST Group, et al. Mobile phone and wearable sensor-based mHealth approaches for psychiatric disorders and symptoms: systematic review. JMIR Ment Health 2019 Feb 20;6(2):e9819 [FREE Full text] [doi: 10.2196/mental.9819] [Medline: 30785404]
7. Shapiro A. Gilles de la Tourette Syndrome. 2nd edition. New York, NY: Raven Press; 1987.
8. Bernabei M, Andreoni G, Mendez Garcia MO, Piccini L, Aletti F, Sassi M, et al. Automatic detection of tic activity in the Tourette Syndrome. Annu Int Conf IEEE Eng Med Biol Soc 2010;2010:422-425. [doi: 10.1109/IEMBS.2010.5627374] [Medline: 21096762]
9. Fraiwan L, Khnouf R, Mashagbeh AR. Parkinson's disease hand tremor detection system for mobile application. J Med Eng Technol 2016;40(3):127-134. [doi: 10.3109/03091902.2016.1148792] [Medline: 26977823]
10. Kim J, Parnell C, Wichmann T, DeWeerth SP. Longitudinal wearable tremor measurement system with activity recognition algorithms for upper limb tremor. Annu Int Conf IEEE Eng Med Biol Soc 2016 Aug;2016:6166-6169. [doi: 10.1109/EMBC.2016.7592136] [Medline: 28269660]

11. Cole BT, Roy SH, De Luca CJ, Nawab SH. Dynamical learning and tracking of tremor and dyskinesia from wearable sensors. IEEE Trans Neural Syst Rehabil Eng 2014 Sep;22(5):982-991. [doi: 10.1109/TNSRE.2014.2310904] [Medline: 24760943]
12. Tourette's. OSF. URL: https://osf.io/c3qp6/?view_only=2e3a16e630ee42cf8e1b63daae5102f0 [accessed 2023-04-10]

Abbreviations

GDPR: General Data Protection Regulation
HIPAA: Health Insurance Portability and Accountability Act
LOF: local outlier factor
TS: Tourette syndrome

Edited by P Kubben; submitted 09.10.22; peer-reviewed by K Khan, X Cheng; comments to author 21.12.22; revised version received 11.03.23; accepted 16.03.23; published 25.04.23.

Please cite as:
Rajinikanth A, Clark DK, Kapsetaki ME
A Novel System to Monitor Tic Attacks for Tourette Syndrome Using Machine Learning and Wearable Technology: Preliminary Survey Study and Proposal for a New Sensing Device
JMIR Neurotech 2023;2:e43351
URL: https://neuro.jmir.org/2023/1/e43351
doi:10.2196/43351
PMID:

Original Paper

Application of a Low-Cost mHealth Solution for the Remote Monitoring of Patients With Epilepsy: Algorithm Development and Validation

Natarajan Sriraam[1], PhD; S Raghu[1,2], PhD; Erik D Gommer[3], PhD; Danny M W Hilkman[3], MD; Yasin Temel[2], MD; Shyam Vasudeva Rao[2], PhD; Alangar Satyaranjandas Hegde[4], MD; Pieter L Kubben[2], MD

[1]Center for Medical Electronics and Computing, Ramaiah Institute of Technology, Bengaluru, India

[2]Department of Neurosurgery, Maastricht University, Maastricht, Netherlands

[3]Department of Clinical Neurophysiology, Maastricht University Medical Centre, Maastricht, Netherlands

[4]Institute of Neuroscience, Ramaiah Medical College and Hospitals, Bengaluru, India

Corresponding Author:
Natarajan Sriraam, PhD
Center for Medical Electronics and Computing
Ramaiah Institute of Technology
MSRIT Post, M S Ramaiah Nagar
Bengaluru, 560054
India
Phone: 91 9632294999
Email: sriraam@msrit.edu

Abstract

Background: Implementing automated seizure detection in long-term electroencephalography (EEG) analysis enables the remote monitoring of patients with epilepsy, thereby improving their quality of life.

Objective: The objective of this study was to explore an mHealth (mobile health) solution by investigating the feasibility of smartphones for processing large EEG recordings for the remote monitoring of patients with epilepsy.

Methods: We developed a mobile app to automatically analyze and classify epileptic seizures using EEG. We used the cross-database model developed in our previous study, incorporating successive decomposition index and matrix determinant as features, adaptive median feature baseline correction for overcoming interdatabase feature variation, and postprocessing-based support vector machine for classification using 5 different EEG databases. The Sezect (Seizure Detect) Android app was built using the Chaquopy software development kit, which uses the Python language in Android Studio. Various durations of EEG signals were tested on different smartphones to check the feasibility of the Sezect app.

Results: We observed a sensitivity of 93.5%, a specificity of 97.5%, and a false detection rate of 1.5 per hour for EEG recordings using the Sezect app. The various mobile phones did not differ substantially in processing time, which indicates a range of phone models can be used for implementation. The computational time required to process real-time EEG data via smartphones and the classification results suggests that our mHealth app could be a valuable asset for monitoring patients with epilepsy.

Conclusions: Smartphones have multipurpose use in health care, offering tools that can improve the quality of patients' lives.

(JMIR Neurotech 2023;2:e50660) doi:10.2196/50660

KEYWORDS

Android; epileptic seizures; mobile health; mHealth; mobile phone–based epilepsy monitoring; support vector machine; seizure; epileptic; epilepsy; monitoring; smartphone; smartphones; mobile phone; neurology; neuroscience; electroencephalography; EEG; brain; classification; detect; detection; neurological; electroencephalogram; diagnose; diagnosis; diagnostic; imaging

Introduction

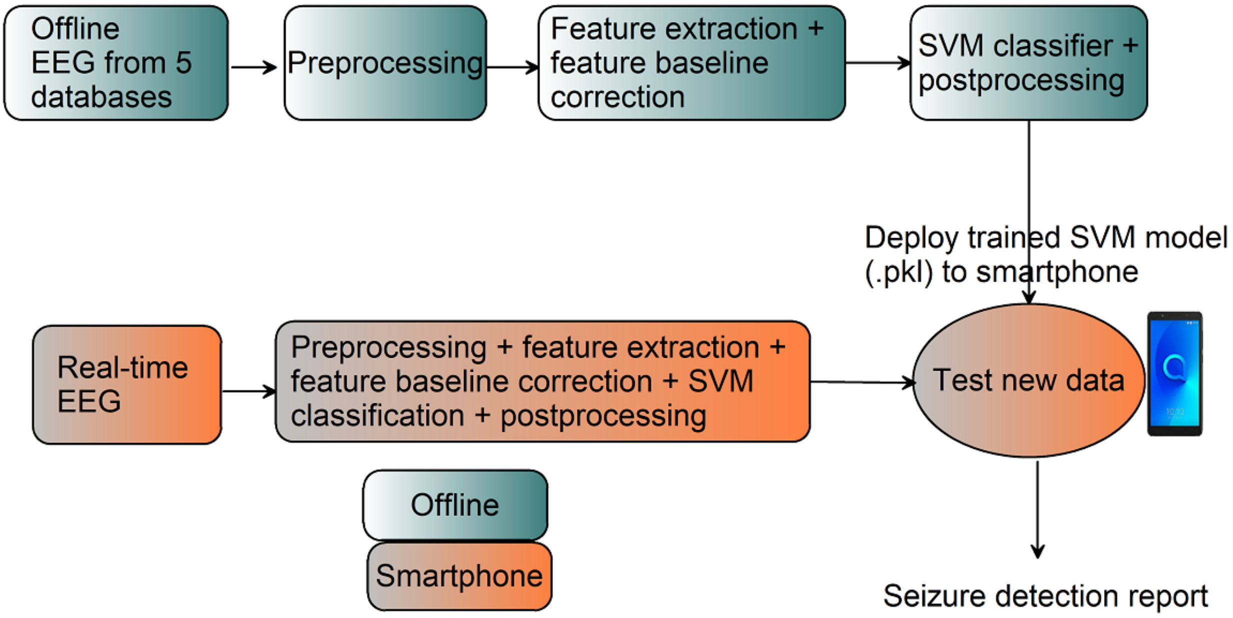
According to the International League Against Epilepsy, epileptic seizures are characterized by an unpredictable occurrence pattern and transient dysfunctions of the central nervous system due to excessive and synchronous abnormal neuronal activity in the cortex [1]. Electroencephalography (EEG) can be used to determine the epileptogenic zone or to monitor patients in the intensive care unit for seizures or monitor seizures for therapy adjustment. EEG signals are collected over a period of time and analyzed to detect seizure events. Today, almost everyone uses smartphones, and smartphone apps are being used to solve real-world human challenges including health-related issues. Regarding the remote monitoring of patients with epilepsy, there is a need to develop an efficient smartphone app that processes long-term EEG recordings for seizure detection. Therefore, the goal of this paper was to develop and evaluate the feasibility of a mobile app for the remote monitoring of patients with epilepsy.

In this context, an automatic mobile phone–based approach for epileptic seizure detection was proposed by Menshawy et al [2] using time, frequency, entropy, and discrete wavelet transform–based features with k-means clustering. EEG signals recorded from the EEG headset were stored in smartphones and transmitted to a server. The preprocessing, feature extraction, feature normalization, feature selection, and classification model of EEG signals were performed on a cloud server. The results were sent to the smartphones of patients and physicians via a backend server. Based on the classification results, caretakers were notified to take appropriate action. This study faced limitations in terms of memory as the complete EEG signal had to be sent to the server. Additionally, this approach was computationally expensive due to the use of a large number of features. McKenzie et al [3] assessed the ability of Smartphone Brain Scanner-2 to detect epileptiform abnormalities using an Android tablet that was wirelessly connected to a 14-electrode EasyCap headset. An Android-based smartphone app for monitoring patients with epilepsy was proposed using subband features and a support vector machine (SVM) classifier [4]. mHealth (mobile health) has been proposed to detect generalized tonic-clonic seizures, whereby an alarm is triggered for timely interventions resulting in a possibly reduced risk of sudden unexpected death in epilepsy [5].

Kiral-Kornek et al [6] proposed a mobile system–based epileptic seizure prediction using big data and deep learning using intracranial EEG signals. Typical statistics like seizures per month, average sensitivity, and average warning time were reported. Moreover, other studies have proposed a cloud-based alert system using advanced statistics [7] and have explored seizure prediction through deep learning techniques for EEG big data [8,9]. Some studies [2,6-14] have used cloud computing for EEG analysis and seizure detection. Additionally, a few mobile devices, namely SmartWatch, Embrace Watch, Brain Sentinel, and EpiWatch App, have been developed for seizure detection to alert caretakers and to prevent sudden unexpected death due to epilepsy [15].

Our study focuses on harnessing smartphone capabilities to implement the entire seizure detection model, which eliminates the need for cloud technology. The Sezect (Seizure Detect) app, our mobile phone–based seizure detection model, provides information such as the number of channels, sampling frequency, EEG signal duration, seizure frequency per channel, and seizure-affected channels. Further, the app was developed using open-source software, allowing researchers public access and the ability to replicate the process. Therefore, the proposed approach could be a valuable tool for the remote monitoring of patients with epilepsy. Figure 1 shows a block diagram of the proposed smartphone-based monitoring approach for patients with epilepsy.

Figure 1. Block diagram of the proposed smartphone-based monitoring approach for patients with epilepsy. EEG: electroencephalography; SVM: support vector machine.


Offline EEG from 5 databases
Preprocessing
Feature extraction + feature baseline correction
SVM classifier + postprocessing
Deploy trained SVM model (.pkl) to smartphone
Real-time EEG
Preprocessing + feature extraction + feature baseline correction + SVM classification + postprocessing
Test new data
Offline
Smartphone
Seizure detection report

Methods

Clinical EEG Recordings

In order to deploy the seizure detection model via smartphone, the cross-database model in our previous study was developed using EEG recordings from Ramaiah Medical College and Hospitals (RMCH) (Bengaluru, India), Children's Hospital Boston-Massachusetts Institute of Technology (CHB-MIT) (Boston, MA), Temple University Hospitals (TUH) (Philadelphia, PA), Maastricht University Medical Centre (MUMC) (Maastricht, Netherlands), and the University of Bonn (UBonn) (Bonn, Germany) [16]. The same cross-database model was implemented on a smartphone and validated using 20 new patients' EEG recordings collected from the RMCH and MUMC databases. EEG recordings with a total duration of 13 hours were tested via smartphone.

Ethical Considerations

The 3 EEG recordings used in our study, namely from CHB-MIT, TUH, and UBonn, are available publicly. Ethical committee approval was sought for the RMCH and MUMC EEG recordings before use in this study.

Chaquopy

Chaquopy is the Python software development kit for Android [17], which allows reuse of existing Python code on Android and takes advantage of Python Package Index packages including *numpy*, *sci-kit learn*, *scipy*, and others. The GitHub repository provides more details on how to use Chaquopy [18]. The *chaquopy-console* template was used to run the seizure detection Python code on the app.

Seizure Detection Model

The methods followed in this study were introduced by us in our previous studies [16,19-21]. The optimized cross-database seizure detection model was built in our previous study [16]. Two features, the successive decomposition index [19] and matrix determinant [20], were extracted from all 5 databases and their baseline was updated using adaptive median feature baseline correction [21]. The features were classified using the SVM classifier via the leave-one-database-out cross-validation method and a postprocessing technique was implemented by applying a 10-tap moving average filter to the classifier output to reduce false detections. This model was then coded in Python and exported into a pickle file for smartphones to test the new EEG recordings.

Mobile Phone–Based Seizure Monitoring

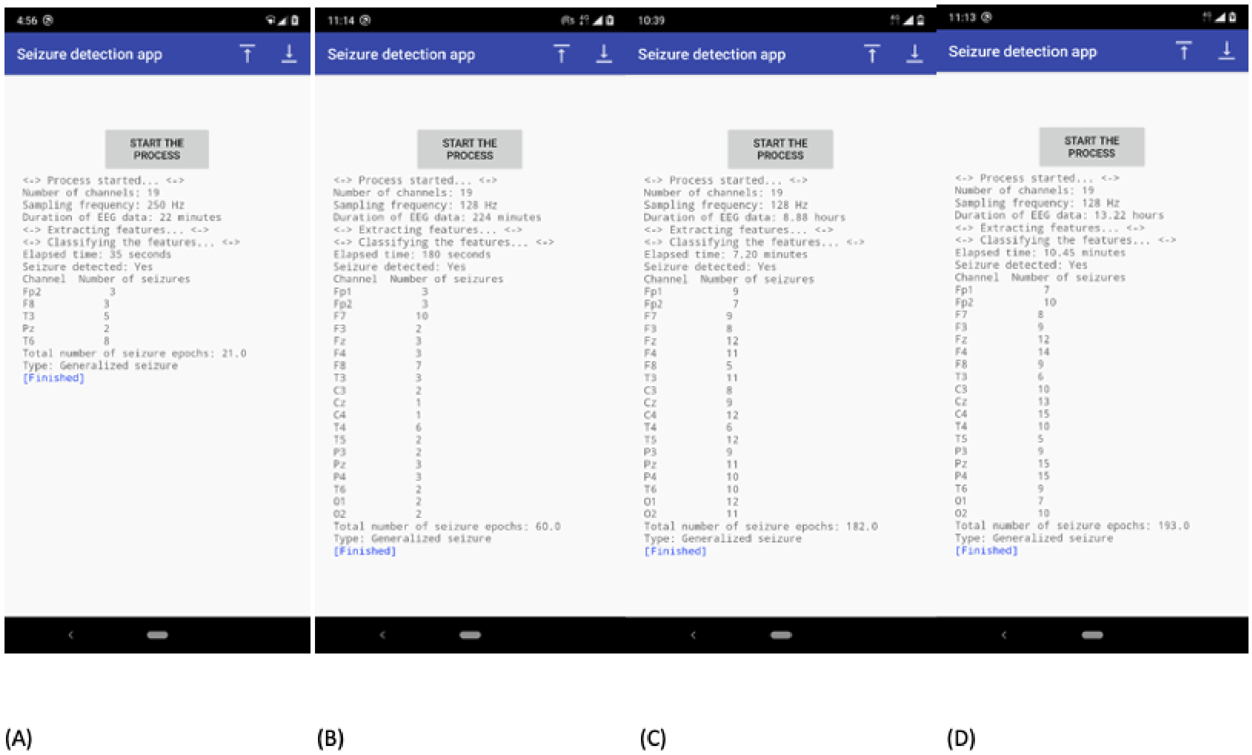
To demonstrate proof of concept, the Sezect Android app was tested for epileptic seizure detection using EEG signals. It is important to investigate how different versions or models of smartphones perform in processing EEG signals, which will be useful to know to make the proposed method scalable. Therefore, we tested the proposed algorithm on the following mobile phones: Nokia, Moto X Play, and Redmi Note 4. Overall, 20 new EEG recordings from patients with epilepsy from both MUMC and RMCH were used to evaluate the efficiency of these smartphones for seizure detection. Using *joblib* from the *sklearnexternals* library, the trained SVM model was dumped into a .pkl file and loaded into the Sezect app to test the recordings.

Results

The Sezect app was tested on 3 Android mobile phones with the following configurations: (1) Nokia 8.1 (Android 10), (2) Moto X Play (Android 8), and (3) Redmi Note 4 (Android 10). Screenshots of the Sezect app results using Nokia are shown in Figure 2, and a video of running the app is available online [22]. As shown in Figure 2, the app pulls information such as the number of channels, sampling frequency, and duration of the complete EEG data file. Further, it displays the elapsed time required to process the complete EEG file, the number of seizure events detected per channel, and the total number of seizure epochs (each epoch length is 10 seconds).

Figure 3 illustrates the time taken by various smartphones to process EEG recordings of different durations. The processing time of the Sezect app shows that a mobile platform is capable of handling large amounts of EEG data and perform feature calculation and classification. The various mobile phones did not differ substantially in processing time, which indicates a range of phone models can be used for implementation. Further, the robustness and scalability of the app was examined using various hardware configurations for all 3 smartphones. We observed a sensitivity of 93.5%, a specificity of 97.5%, and a false detection rate of 1.5 per hour for new EEG recordings using the Sezect app. The results suggest that our proposed seizure detection algorithm could be a valuable asset to remotely monitor patients with epilepsy using smartphone apps.

Figure 2. Screenshots of the Sezect app results: (A) the Maastricht University Medical Centre database with 22 minutes of electroencephalography (EEG) data and the Ramaiah Medical College and Hospitals database with (B) 3.73 hours, (C) 8.88 hours, and (D) 13.22 hours of EEG data.



(A)

```
<-> Process started... <->
Number of channels: 19
Sampling frequency: 250 Hz
Duration of EEG data: 22 minutes
<-> Extracting features... <->
<-> Classifying the features... <->
Elapsed time: 35 seconds
Seizure detected: Yes
Channel  Number of seizures
Fp2          3
F8          3
T3          5
Pz          2
T6          8
Total number of seizure epochs: 21.0
Type: Generalized seizure
[Finished]
```

(B)

```
<-> Process started... <->
Number of channels: 19
Sampling frequency: 128 Hz
Duration of EEG data: 224 minutes
<-> Extracting features... <->
<-> Classifying the features... <->
Elapsed time: 180 seconds
Seizure detected: Yes
Channel  Number of seizures
Fp1          3
Fp2          3
F7          10
F3          2
Fz          3
F4          3
F8          7
T3          3
C3          2
Cz          1
C4          1
T4          6
T5          2
P3          2
Pz          3
P4          3
T6          2
O1          2
O2          2
Total number of seizure epochs: 60.0
Type: Generalized seizure
[Finished]
```

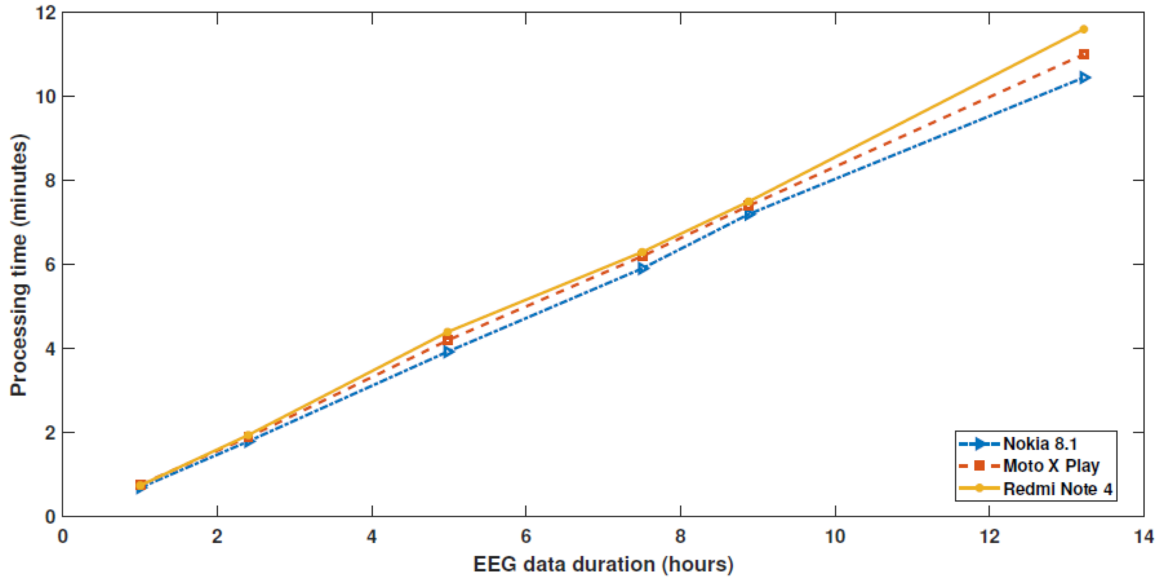
(C)

```
<-> Process started... <->
Number of channels: 19
Sampling frequency: 128 Hz
Duration of EEG data: 8.88 hours
<-> Extracting features... <->
<-> Classifying the features... <->
Elapsed time: 7.20 minutes
Seizure detected: Yes
Channel  Number of seizures
Fp1          9
Fp2          7
F7          9
F3          8
Fz          12
F4          11
F8          5
T3          11
C3          8
Cz          9
C4          12
T4          6
T5          12
P3          9
Pz          11
P4          10
T6          10
O1          12
O2          11
Total number of seizure epochs: 182.0
Type: Generalized seizure
[Finished]
```

(D)

```
<-> Process started... <->
Number of channels: 19
Sampling frequency: 128 Hz
Duration of EEG data: 13.22 hours
<-> Extracting features... <->
<-> Classifying the features... <->
Elapsed time: 10.45 minutes
Seizure detected: Yes
Channel  Number of seizures
Fp1          7
Fp2          10
F7          8
F3          9
Fz          12
F4          14
F8          9
T3          6
C3          10
Cz          13
C4          15
T4          10
T5          5
P3          9
Pz          15
P4          15
T6          9
O1          7
O2          10
Total number of seizure epochs: 193.0
Type: Generalized seizure
[Finished]
```

Figure 3. The processing time required to analyze and classify various durations of EEG signals on different smartphones. EEG: electroencephalography.



Discussion

Comparison With State-of-the-Art Studies

Few studies have used cloud technology, the internet of things, and smartphones to analyze EEGs and detect seizure epochs. Menshawy et al [2] used server-based processing for data preprocessing, feature engineering, and classification; subsequently, generated reports were sent to doctors, and caretakers were alerted upon detecting seizures. Cloud computing was effectively used in some studies [4,10,12,14] to perform feature engineering and classification using cloud technology. Moreover, mobile devices like SmartWatch, Embrace Watch, Brain Sentinel, and the EpiWatch App are in use, designed to detect specific types of seizures [15]. However, the proposed Sezect app was built using the cross-database model from 5 EEG databases and has been found to be effective in terms of computational time when tested on 3 different smartphones. Physicians and nurses working in rural areas can record EEG data and validate it using the Sezect app.

Contributions

The following is a summary of our contributions:

1. We developed the Sezect app for Android using open-source software to remotely monitor patients with epilepsy. The app is made available as open-source software to improve the reproducibility of our results. The source code for the Sezect app can be found online [23].
2. The feasibility of smartphones for handling large EEG recordings was determined using the Sezect app. Further, we examined the time complexity by assessing the elapsed time of the mobile app across various EEG durations.
3. Running all tasks on the cloud demands substantial memory and can be costly. This study's major contribution lies in demonstrating the feasibility of automated seizure detection via smartphones, eliminating the involvement of cloud infrastructure.

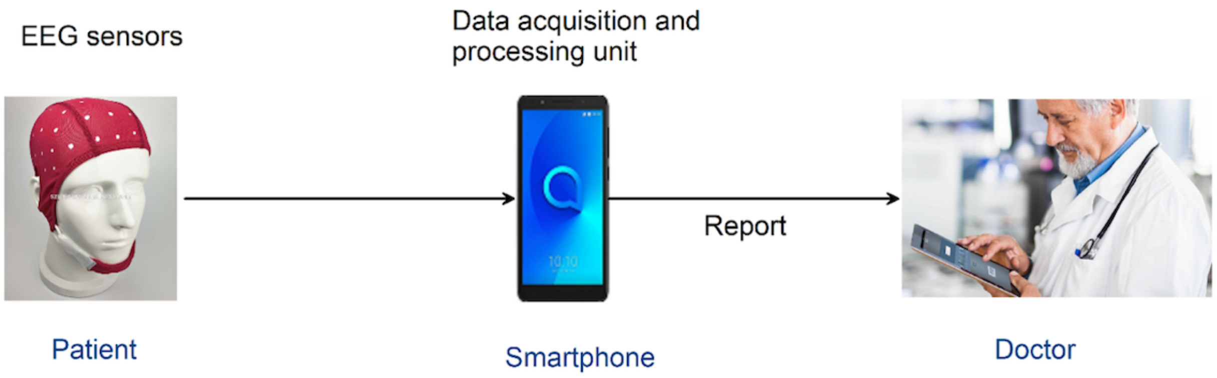
Clinical Significance

Remote monitoring using smartphone apps will be useful to monitor patients with epilepsy by analyzing EEG signals collected over a period of time. Smartphones can serve multiple uses in health care to improve the quality of patients' lives. The advanced technology of smartphones can be applied to solve the workload burden of experts.

Future Directions

This study presented a proof of concept for a low-cost mHealth solution aimed at the automated detection of epileptic seizures for remote monitoring. Figure 4 illustrates the architectural scope for a future remote monitoring system. In such a system, a wireless EEG headset will be provided to the patient and continuous real-time EEG signals will be recorded and stored on smartphones. A cross-database classification model within the smartphone will analyze EEG signals, generating a report sent directly to the relevant physician for further action.

Figure 4. A future scope architecture for real-time smartphone-based seizure detection and the remote monitoring of patients with epilepsy. EEG: electroencephalography.



Limitations

In our current implementation, we observed a slightly elevated false detection rate, which needs to be addressed in the future.

Conclusion

The feasibility of a mobile phone–based app for the remote monitoring of patients with epilepsy using a database-independent optimized algorithm was demonstrated. The app is open source, allowing researchers to reproduce it according to their specific needs. It was tested using 3 different types of smartphones. The results suggest that smartphones are capable of handling large amounts of EEG data for feature calculation and classification.

Acknowledgments

The authors are grateful to the Department of Clinical Neurophysiology, Maastricht University Medical Centre (Maastricht, The Netherlands), and Ramaiah Medical College and Hospitals (Bengaluru, India) for providing the EEG recordings. The authors are also grateful to Dr Ali Shoeb from Children's Hospital Boston-Massachusetts Institute of Technology, Dr RG Andrzejak from the University of Bonn, and the team at Temple University Hospitals for permitting the use of their database for our study.

This research did not receive any specific grants from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability

The EEG data used in this study are intended solely for research purposes and cannot be made available to the public.

Authors' Contributions

SR and NS formulated the problem statement. SR was involved in writing codes, simulations, and drafting the paper. NS, EDG, DMWH, YT, and SVR validated the results and reviewed the manuscript. PK and ASH were involved in clinical validation, clinical discussion, and the review process.

Conflicts of Interest

PK is the editor-in-chief of *JMIR Neurotechnology*. The other authors declare no conflicts of interest.

References

1. Fisher RS, Cross JH, French JA, Higurashi N, Hirsch E, Jansen FE, et al. Operational classification of seizure types by the International League Against Epilepsy: position paper of the ILAE Commission for Classification and Terminology. Epilepsia 2017 Apr 08;58(4):522-530 [FREE Full text] [doi: 10.1111/epi.13670] [Medline: 28276060]
2. EL Menshawy M, Benharref A, Serhani M. An automatic mobile-health based approach for EEG epileptic seizures detection. Expert Syst Appl 2015 Nov;42(20):7157-7174. [doi: 10.1016/j.eswa.2015.04.068]
3. McKenzie ED, Lim ASP, Leung ECW, Cole AJ, Lam AD, Eloyan A, et al. Validation of a smartphone-based EEG among people with epilepsy: a prospective study. Sci Rep 2017 Apr 03;7(1):45567 [FREE Full text] [doi: 10.1038/srep45567] [Medline: 28367974]
4. Lasefr Z, Reddy RR, Elleithy K. Smart phone application development for monitoring epilepsy seizure detection based on EEG signal classification. 2017 Presented at: IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference; Oct 19-21; New York, NY p. 83-87. [doi: 10.1109/uemcon.2017.8248992]
5. Ryvlin P, Beniczky S. Seizure detection and mobile health devices in epilepsy: update and future developments. Epilepsia 2018 Jun 05;59 Suppl 1(S1):7-8. [doi: 10.1111/epi.14088] [Medline: 29873830]
6. Kiral-Kornek I, Roy S, Nurse E, Mashford B, Karoly P, Carroll T, et al. Epileptic seizure prediction using big data and deep learning: toward a mobile system. EBioMedicine 2018 Jan;27:103-111 [FREE Full text] [doi: 10.1016/j.ebiom.2017.11.032] [Medline: 29262989]
7. Baldassano S, Zhao X, Brinkmann B, Kremen V, Bernabei J, Cook M, et al. Cloud computing for seizure detection in implanted neural devices. J Neural Eng 2019 Apr 04;16(2):026016 [FREE Full text] [doi: 10.1088/1741-2552/aaf92e] [Medline: 30560812]
8. Hosseini M, Pompili D, Elisevich K, Soltanian-Zadeh H. Optimized deep learning for EEG big data and seizure prediction BCI via Internet of Things. IEEE Trans Big Data 2017 Dec 1;3(4):392-404. [doi: 10.1109/tbdata.2017.2769670]
9. Hosseini MP, Soltanian-Zadeh H, Elisevich K, Pompili D. Cloud-based deep learning of big EEG data for epileptic seizure prediction. 2016 Presented at: IEEE Global Conference on Signal and Information Processing; Dec 7-9; Washington DC p. 1151-1155. [doi: 10.1109/globalsip.2016.7906022]
10. Sareen S, Sood SK, Gupta SK. An automatic prediction of epileptic seizures using cloud computing and wireless sensor networks. J Med Syst 2016 Nov 15;40(11):226. [doi: 10.1007/s10916-016-0579-1] [Medline: 27628727]
11. Vergara PM, de la Cal E, Villar JR, González VM, Sedano J. An IoT platform for epilepsy monitoring and supervising. J Sensors 2017;2017:1-18. [doi: 10.1155/2017/6043069]
12. Escobar Cruz N, Solarte J, Gonzalez-Vargas A. Automated epileptic seizure detection system based on a wearable prototypecloud computing to assist people with epilepsy. In: Figueroa-García J, Villegas J, Orozco-Arroyave J, Maya Duque P, editors. Applied Computer Sciences in Engineering. WEA 2018. Communications in Computer and Information Science, vol 916. Cham, Switzerland: Springer; 2018.
13. Zhang Z, Wen T, Huang W, Wang M, Li C. Automatic epileptic seizure detection in EEGs using MF-DFA, SVM based on cloud computing. XST 2017 Mar 21;25(2):261-272. [doi: 10.3233/xst-17258]
14. Marquez A, Dunn M, Ciriaco J, Farahmand F. iSeiz: a low-cost real-time seizure detection system utilizing cloud computing. 2017 Presented at: IEEE Global Humanitarian Technology Conference; Oct 19-22; San Jose, CA p. 1-7. [doi: 10.1109/ghtc.2017.8239249]
15. Greb E. Mobile devices may provide accurate seizure detection and help prevent SUDEP. Neurology Reviews 2017;25(2):28-29 [FREE Full text]
16. Raghu S, Sriraam N, Gommer ED, Hilkman DM, Temel Y, Rao SV, et al. Cross-database evaluation of EEG based epileptic seizures detection driven by adaptive median feature baseline correction. Clin Neurophysiol 2020 Jul;131(7):1567-1578. [doi: 10.1016/j.clinph.2020.03.033] [Medline: 32417698]
17. Chaquopy: Python SDK for Android. 2012. URL: https://chaquo.com/chaquopy/ [accessed 2019-05-05]
18. Chaquo/chaquopy. GitHub. URL: https://github.com/chaquo/chaquopy [accessed 2023-12-06]
19. Raghu S, Sriraam N, Vasudeva Rao S, Hegde AS, Kubben PL. Automated detection of epileptic seizures using successive decomposition index and support vector machine classifier in long-term EEG. Neural Comput & Applic 2019 Jul 31;32(13):8965-8984. [doi: 10.1007/s00521-019-04389-1]
20. Raghu S, Sriraam N, Hegde AS, Kubben PL. A novel approach for classification of epileptic seizures using matrix determinant. Expert Syst Appl 2019 Aug;127:323-341. [doi: 10.1016/j.eswa.2019.03.021]

21. Raghu S, Sriraam N, Gommer ED, Hilkman DM, Temel Y, Rao SV, et al. Adaptive median feature baseline correction for improving recognition of epileptic seizures in ICU EEG. Neurocomputing 2020 Sep;407:385-398. [doi: 10.1016/j.neucom.2020.04.144]
22. Raghu S. Video of Sezect Android app. Zenodo. 2019 Dec 30. URL: https://doi.org/10.5281/zenodo.3595429 [accessed 2023-11-28]
23. Raghu S. The proof of concept for epilepsy patients monitoring using Sezect Android app. Zenodo. 2019 Dec 24. URL: http://doi.org/10.5281/zenodo.3592415 [accessed 2019-11-28]

Abbreviations

CHB-MIT: Children's Hospital Boston-Massachusetts Institute of Technology
EEG: electroencephalography
mHealth: mobile health
MUMC: Maastricht University Medical Centre
RMCH: Ramaiah Medical College and Hospitals
Sezect: Seizure Detect
SVM: support vector machine
TUH: Temple University Hospitals
UBonn: University of Bonn

Edited by T de Azevedo Cardoso; submitted 08.07.23; peer-reviewed by T Yano, P Faroque; comments to author 04.10.23; revised version received 14.11.23; accepted 21.11.23; published 19.12.23.

Please cite as:
Sriraam N, Raghu S, Gommer ED, Hilkman DMW, Temel Y, Vasudeva Rao S, Hegde AS, L Kubben P
Application of a Low-Cost mHealth Solution for the Remote Monitoring of Patients With Epilepsy: Algorithm Development and Validation
JMIR Neurotech 2023;2:e50660
URL: https://neuro.jmir.org/2023/1/e50660
doi:10.2196/50660
PMID:

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org