JMIR Biomedical Engineering

Engineering for health technologies, medical devices, and innovative medical treatments and procedures
Volume 7 (2022), Issue 2   ISSN 2561-3278   Editor in Chief: Syed A. A. Rizvi, MD, PhD, MBA, MPH, BSN

Contents

Review

Viewpoint

Original Papers

Review

Detection of Suicide Risk Using Vocal Characteristics: Systematic Review

Ravi Iyer[1], BSc, MSc; Denny Meyer[1], PhD

Centre for Mental Health, Swinburne University of Technology, Hawthorn, Australia

Corresponding Author:
Ravi Iyer, BSc, MSc
Centre for Mental Health
Swinburne University of Technology
34 Wakefield Street
Hawthorn, 3122
Australia
Phone: 61 456565575
Email: raviiyer@swin.edu.au

Abstract

Background: In an age when telehealth services are increasingly being used for forward triage, there is a need for accurate suicide risk detection. Vocal characteristics analyzed using artificial intelligence are now proving capable of detecting suicide risk with accuracies superior to traditional survey-based approaches, suggesting an efficient and economical approach to ensuring ongoing patient safety.

Objective: This systematic review aimed to identify which vocal characteristics perform best at differentiating between patients with an elevated risk of suicide in comparison with other cohorts and identify the methodological specifications of the systems used to derive each feature and the accuracies of classification that result.

Methods: A search of MEDLINE via Ovid, Scopus, Computers and Applied Science Complete, CADTH, Web of Science, ProQuest Dissertations and Theses A&I, Australian Policy Online, and Mednar was conducted between 1995 and 2020 and updated in 2021. The inclusion criteria were human participants with no language, age, or setting restrictions applied; randomized controlled studies, observational cohort studies, and theses; studies that used some measure of vocal quality; and individuals assessed as being at high risk of suicide compared with other individuals at lower risk using a validated measure of suicide risk. Risk of bias was assessed using the Risk of Bias in Non-randomized Studies tool. A random-effects model meta-analysis was used wherever mean measures of vocal quality were reported.

Results: The search yielded 1074 unique citations, of which 30 (2.79%) were screened via full text. A total of 21 studies involving 1734 participants met all inclusion criteria. Most studies (15/21, 71%) sourced participants via either the Vanderbilt II database of recordings (8/21, 38%) or the Silverman and Silverman perceptual study recording database (7/21, 33%). Candidate vocal characteristics that performed best at differentiating between high risk of suicide and comparison cohorts included timing patterns of speech (median accuracy 95%), power spectral density sub-bands (median accuracy 90.3%), and mel-frequency cepstral coefficients (median accuracy 80%). A random-effects meta-analysis was used to compare 22 characteristics nested within 14% (3/21) of the studies, which demonstrated significant standardized mean differences for frequencies within the first and second formants (standardized mean difference ranged between −1.07 and −2.56) and jitter values (standardized mean difference=1.47). In 43% (9/21) of the studies, risk of bias was assessed as moderate, whereas in the remaining studies (12/21, 57%), the risk of bias was assessed as high.

Conclusions: Although several key methodological issues prevailed among the studies reviewed, there is promise in the use of vocal characteristics to detect elevations in suicide risk, particularly in novel settings such as telehealth or conversational agents.

Trial Registration: PROSPERO International Prospective Register of Systematic Reviews CRD420200167413; https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42020167413

(JMIR Biomed Eng 2022;7(2):e42386) doi:10.2196/42386

KEYWORDS

voice; suicide; suicidal; biological signal processing; machine learning; systematic review; review methodology; risk; speech; mental health

Introduction

Background

Telehealth alternatives may soon replace in-person visits to providers of primary health care [1]. Telehealth is effective in reducing the severity of mental illness [2], leading the Australian government to commit to universal access to telehealth care alternatives [3].

The potential utility of telehealth services to the community is undeniable, being ideally suited to reach sectors of the population that have historically faced barriers to access. In particular, rural and remote communities face unique challenges in terms of both economic disparity and location [4]. Telehealth services are also appealing to health care consumers of a younger age (19-44 years) [5]. These cohorts substantially overlap with those most at risk of suicide [6,7].

Telehealth is also being used in other, more novel ways. Primary health care providers in the United States are increasingly using *forward triage*, where patients are assessed before arrival via telehealth means and often via a conversational agent [8]. However, this may prove challenging when mental health is the main presenting issue as suicidality is a feature of most mental health disorders [9]. Thus, the transition from in-person provision of health care raises important ethical considerations. For example, how can escalation in suicide risk be accurately and efficiently assessed in the absence of in-person cues?

In >50 years of research, traditional methods of suicide risk assessment (ie, surveys) have yielded little more than chance accuracy in identifying elevated suicide risk [10]. Franklin et al [10] suggested that suicide risk assessment would benefit from the use of risk algorithms that can assess multiple predictors simultaneously. However, they did not consider the use of biological markers in their review. Such markers do not rely on patient testimony and may prove more accurate in the assessment of suicide risk [11].

Suicide-related biological marker research has focused mainly on identifying neurobiological changes associated with elevated risk. However, the downstream effects of these neurobiological changes may also be apparent and remain underresearched. In particular, changes in speech production and articulation—the subject of this review—have been associated with elevated suicide risk, as indicated in this section. There is an identified need to leverage these novel technologies in the provision of real-time adaptive personalization of counseling content to match consumer emotions [12]. This is consistent with the recent recommendations of Balcombe and De Leo [13], who also argue for the real-time tracking of consumer emotions via machine learning–trained predictive models that can assist in delivering more timely and efficient mental health care support at scale.

In their review of vocal characteristics used to identify suicide risk, Cummins et al [14] found that many characteristics could prove viable in the detection and differentiation of suicide risk presentations. They identified 4 types of vocal characteristics used for this purpose: prosodic (long-term changes in rhythm, stress, and intonation), voice production, formant (changes in vocal tract properties), and frequency (pitch). Cummins et al [14] noted that the speech of individuals at high risk of suicide is often distinguished by a hollow, toneless, and monotonous quality or by a breathy tone, which corresponds to a marked change in spectral slope (accuracies of 90% when using this variable to predict suicide risk) [15]. Cummins et al [14] also noted that the second formant bandwidth and power spectral densities between 0 and 1000 Hz are promising candidates for further research (accuracies of 90% were obtained using a combination of these features).

Homan et al [16] recently reviewed the use of both voice signals and text-based data to predict suicide risk. The findings of Cummins et al [14] were supported by those of Homan et al [16], who also suggested pause length and jitter (the timing of the glottal pulse) as additional candidates. However, the authors did not discuss accuracy of prediction or the methodological specifications informing the systems of classification used. However, both Cummins et al [14] and Homan et al [16] agree that, despite prevailing methodological issues, namely, small sample sizes, lack of control of covariates, and validity of ground truth, there is evidence that suicide risk does alter the human voice in substantive and important ways and may be predictive of elevated suicide risk. Other authors have also noted the equivocality of current findings and the need for further confirmatory research [17].

Objectives

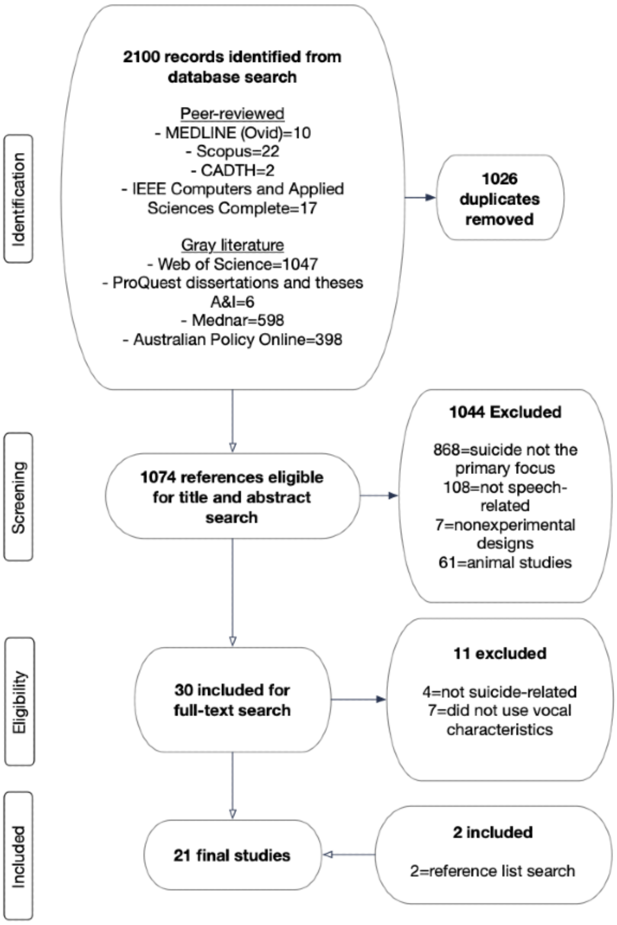
Thus, the primary objective of this systematic review was to assess the accuracy of vocal characteristics in differentiating between individuals at risk of suicide and those who are not at risk. The secondary objective was to assess the methodological specifications used in these systems of classification.

Methods

Design

This systematic review was conducted using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) standards [18] (Figure 1) and checklist (Multimedia Appendix 1). The systematic review protocol was registered with PROSPERO on April 28, 2020 (registration CRD420200167413) [19]. The Population, Intervention, Comparator, Outcome, and Study Design framework defined the research questions and search terms. The research questions were as follows: Which vocal characteristics can differentiate between high and low risk of suicide among both adult and adolescent populations with a high level of accuracy? and What are the methodological specifications used to derive the vocal characteristics and inform the levels of accuracy obtained?

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.


Identification
2100 records identified from database search
Peer-reviewed
- MEDLINE (Ovid)=10
- Scopus=22
- CADTH=2
- IEEE Computers and Applied Sciences Complete=17
Gray literature
- Web of Science=1047
- ProQuest dissertations and theses A&I=6
- Mednar=598
- Australian Policy Online=398
1026 duplicates removed
Screening
1074 references eligible for title and abstract search
1044 Excluded
868=suicide not the primary focus
108=not speech-related
7=nonexperimental designs
61=animal studies
Eligibility
30 included for full-text search
11 excluded
4=not suicide-related
7=did not use vocal characteristics
Included
21 final studies
2 included
2=reference list search

Information Sources

MEDLINE via Ovid, Scopus, Computers and Applied Sciences Complete, and CADTH, in addition to the gray literature databases Web of Science, ProQuest Dissertations and Theses A&I, Australian Policy Online, and Mednar, were searched initially from January 1, 1990, to December 31, 2020, and updated in January 2022.

Search Strategy

Search strategies were developed using Medical Subject Headings and keyword string searches that included synonyms of “suicide,” “vocal,” and “algorithm” as separate blocks. A final block of various vocal characteristics was also added, informed by a preliminary survey of the literature. Gray literature was included to ensure a breadth of sources and that insights from unpublished resources might also be included (ie, theses). As a final step, the reference lists of all the included studies were reviewed to ensure that all possible studies were included. Refer to Multimedia Appendix 2 for all terms and search strategies used.

Inclusion and Exclusion Criteria

The participants were human, with no language or age restrictions applied. The focus of inquiry was single or multiple measures of vocal quality, which included measures of volume, pace, pitch, rate, rhythm, fluency, articulation, enunciation, and tone. The presence of suicidal ideation or recent behavior was considered the intervention, whereas the absence of such ideation or behavior was the comparator. The primary outcome was a validated measure of suicide risk, whereas no setting restrictions were applied. The study design included randomized controlled trials, cohort studies only, and other unpublished research (ie, theses).

We followed the International Classification of Diseases, 11th Revision, which defines *suicidal ideation* as thoughts, ideas, or ruminations about the possibility of ending one’s life; *suicide behavior* as concrete actions that are taken in preparation for fulfilling a wish to end one’s life; and *suicide attempt* as a specific episode of self-harming behavior undertaken with the conscious intention of ending one’s life.

Studies were excluded when they involved animal populations; were unrelated to either the evaluation of vocal quality or suicide risk; did not involve a comparison group; were single-case studies; or did not provide sufficient detail to establish all the Population, Intervention, Comparator, Outcome, and Study Design criteria.

Selection and Data Collection Process

Both authors independently reviewed the title, abstract (step 1), and full text (step 2) of each publication identified in accordance with the inclusion and exclusion criteria. NVivo (version 12; QSR International) [20] was used to classify each publication for inclusion (green), in doubt (amber), and exclusion (red), with any in-doubt publications discussed further by the authors before consensus. Each publication was also coded to provide a rationale for exclusion (ie, 1=suicide not the primary focus, 2=non–speech-related, 3=animal study, and 4=no comparison between groups).

Data Extraction and Quality Assessment

Information was extracted from the included studies according to the following five categories: (1) participant recruitment and characteristics, (2) preprocessing methodological considerations, (3) vocal characteristics, (4) accuracy, and (5) algorithmic approach to classification.

The included studies were also assessed for quality of evidence by RI, confirmed by DM using the Oxford Centre for Evidence-Based Medicine Levels of Evidence. Each study was rated with a score of 1 to 5, where randomized controlled trials typically scored higher (score=2) than nonrandomized studies (score=3). Disagreements were resolved through discussion.

Risk-of-Bias Assessment

RI assessed the methodological quality of the final studies using the Risk of Bias in Non-randomized Studies tool developed by the Cochrane Collaboration [21]. The Risk of Bias in Non-randomized Studies tool involves 3 stages of assessment, including specification of the research question (stage 1) and specification of the effect of interest, result to assess, identification of confounders and cointerventions, risk-of-bias judgment for each domain, and an overall risk-of-bias determination for each study (stage 2). This is then synthesized as an overall risk-of-bias assessment for all studies (stage 3). The risk-of-bias domains include confounding, selection of participants, classification of interventions, deviations from planned interventions, missing data, outcome measurement, reporting of results, and overall bias. The risk of bias was assessed as low, moderate, or high. The risk-of-bias assessments are available in Multimedia Appendix 3 [15,22-41].

Synthesis Methods

The included studies were heterogeneous in terms of the vocal characteristics assessed and reporting, with classification accuracies included in some studies and mean outcome measures included in others. A narrative synthesis was used to organize the information from the included studies where mean outcome measures were not reported. The guidelines of Rodgers et al [42] were applied, which included a preliminary analysis and exploration of relationships followed by the assessment of the robustness of the synthesis.

Wherever possible, data were presented in tabular form, with information broadly organized around study and participant characteristics, followed by the 2 study questions: classification accuracy of suicide risk using vocal characteristics in the first section and methodological steps taken in the second section.

Where mean outcome measures were reported, a random-effects model meta-analysis was conducted to synthesize the available information, although multiple vocal characteristics were typically reported in a small number of studies. Using the R package *metafor* (version 3.8-1; R Foundation for Statistical Computing), standardized mean differences were derived from the mean outcome measures reported. Standardized mean differences for each vocal characteristic were then illustrated using a forest plot. All data used in this systematic review are available in Multimedia Appendix 4.

Results

Figure 1 illustrates the number of studies that were initially identified, screened, deemed eligible, and included in the final analysis.

Summary of the Included Studies

A total of 1074 studies were initially identified. After careful screening, 21 studies from 4 countries were found to comply with all inclusion and exclusion criteria. These studies are summarized in Table 1. The included studies featured 1734 participants overall, with 14% (3/21) of the studies [15,22,23] involving adolescent populations only. The publications by Campbell [24], Sanadi [25], and Sinha [26] were theses, whereas the remaining studies (18/21, 86%) were peer-reviewed journal articles. Most studies (11/21, 52%) were observational in nature, and most studies used participant recordings from either the Vanderbilt II database (8/21, 38%) or the Silverman and Silverman perceptual study (7/21, 33%). These data sources are summarized in Table 2. The number of studies published by year is illustrated in Figure 2 and can be seen to increase slightly from 2006 onward.

Table 1. Sample characteristics of the included studies (N=21).

Author, year (country)	Participants, N	Design	Sample	Assessment measure	Participant age (years)
Anunvrapong and Yingthawornthuk [27], 2014 (Thailand)	30 female	Observational	Psychiatric inpatients	Psychiatric interview	—[a]
Belouali et al [28], 2021 (United States)	124	Longitudinal	Veterans	Patient Health Questionnaire-9	—
Campbell [24], 1995 (United States)	3	Observational	Telephone call recordings	Clinician-rated	—
Figueroa Saavedra et al [29], 2020 (Chile)	100 (60 female and 40 male)	Cross-sectional	University students	Okasha Suicidality Scale	18-19
France et al [30], 2000 (United States)	115 (38 female and 77 male)	Observational	Psychiatric inpatients	Beck Depression Inventory and Hamilton Depression Rating Scale	25-65
Keskinpala et al [31], 2007 (United States)	169 (92 female and 77 male)	Observational	Psychiatric inpatients	Clinician-rated	—
Nik Hashim et al [32], 2015 (Malaysia)	89 (54 female and 35 male)	Observational	Psychiatric inpatients	Beck Depression Inventory-II, Hamilton Depression Rating Scale, Mini International Neuropsychiatric Interview, and Pierce Suicide Intent Scale	—
Nik Hashim et al [33], 2015 (Malaysia)	126	Controlled study	Psychiatric inpatients	Hamilton Depression Rating Scale	22-62 (mean 42.6, SD 10.2)
Ozdas et al [34], 2000 (United States)	20 male	Controlled study	Psychiatric inpatients	Hamilton Depression Rating Scale	25-65
Ozdas et al [35], 2004 (United States)	30 male	Controlled study	Psychiatric inpatients	Clinician-rated	25-65
Ozdas et al [36], 2004 (United States)	30 male	Controlled study	Psychiatric inpatients	Clinician-rated	25-65
Pestian et al [37], 2017 (United States)	379	Controlled study	Psychiatric inpatients or outpatients	Clinician-rated	Adolescent
Sanadi [25], 2011 (United States)	60	Observational	Psychiatric inpatients	Clinician-rated	—
Scherer et al [22], 2013 (United States)	381	Controlled	Databases of recordings	Patient Health Questionnaire-9 and Beck Depression Inventory	Adult (mean 44.7, SD 12.37) and adolescent (13-17)
Scherer et al [15], 2015 (United States)	60	Controlled	Psychiatric inpatients	Columbia Suicide Severity Rating Scale, Suicidal Ideation Questionnaire-Junior, and Ubiquitous Questionnaire	13-17
Sinha [26], 2013 (United States)	17	Observational	Psychiatric inpatients	Clinician-rated	25-65
Subari et al [38], 2010 (Malaysia)	30	Observational	Psychiatric inpatients	Clinician-rated	25-65
Venek et al [23], 2017 (United States)	60	Controlled	Psychiatric inpatients	Columbia Suicide Severity Rating Scale, Suicidal Ideation Questionnaire-Junior, and Ubiquitous Questionnaire	13-17 (mean 15.47, SD 1.5)
Yingthawornsuk et al [39], 2006 (United States)	32 male	Observational	Psychiatric inpatients	Beck Depression Inventory-II	25-65
Yingthawornsuk et al [40], 2007 (United States)	20 female	Observational	Psychiatric inpatients	Beck Depression Inventory-II	25-65
Yingthawornsuk and Shiavi [41], 2008 (Thailand)	25 male	Observational	Psychiatric inpatients	Beck Depression Inventory-II	25-65

[a] Not available.

Table 2. Sources of study participants.

Participant source, year	Details	Studies
Vanderbilt II database, 1993 [43]	Database of recordings of interviews with individuals responding to an advertisement for low-cost psychotherapy; participants met DSM-IV[a] criteria for major depression	• Anunvrapong and Yingthawornthuk [27] • France et al [30] • Ozdas et al [34-36] • Subari et al [38]
Cognitive behavioral therapy and psychopharmacology study, 1992 [44]	Database of recordings of psychotherapy sessions comparing the effects of cognitive behavioral therapy with psychopharmacological interventions	• France et al [30] • Ozdas et al [34,36]
Silverman and Silverman perceptual study, nd[b] [45]	Database of recordings of psychotherapy sessions and suicide notes of patients who had attempted or completed suicide within hours to weeks of the recordings	• France et al [30] • Ozdas et al [34-36] • Subari et al [38]
Vanderbilt University Hospital emergency department	Study involving recordings of Vanderbilt University Hospital emergency department inpatient admissions	• Nik Hashim et al [32] • Yingthawornsuk et al [39,40] • Yingthawornsuk and Shiavi [41]
Cincinnati Children's Hospital interview corpus	60 adolescents enrolled in a prospective study, 30 presenting to the emergency department with suicidal ideation and behaviors versus 30 controls presenting with orthopedic injuries	• Scherer et al [15,22] • Venek et al [23] • Pestian et al [37]
DAIC[c], 2014 [46]	Database of 621 recordings of distressed and nondistressed individuals diagnosed with anxiety, depression, and posttraumatic stress disorder; interviews were conducted in person and via an autonomous agent	• Scherer et al [22]
AVEC[d]	Database of 292 audiovisual recordings of interviews with participants with depression conducted via an autonomous agent	• Scherer et al [22]
Temuco data set	Database of recordings of interviews with 60 first-year university students from the Faculty of Health Sciences of the Autonomous University of Chile, Temuco	• Figueroa Saavedra et al [29]
Washington DC[e] Veterans Affairs Medical Center trial	Large ongoing prospective trial of veterans diagnosed with Gulf War syndrome	• Belouali et al [28]

[a] DSM-IV: Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition.

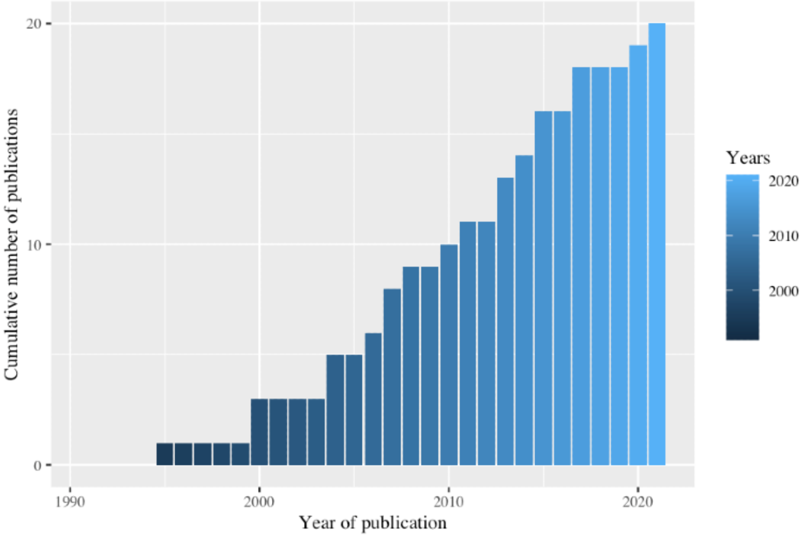
[b] nd: no date.

[c] DAIC: Distress Assessment Interview Corpus.

[d] AVEC: Audio-Visual Depression Corpus.

[e] DC: District of Columbia.

Figure 2. Cumulative number of publications by year.



What Are the Voice Signal Characteristics That Distinguish Elevated Suicide Risk From Other Cohorts?

Most studies (8/21, 38%) used frequency-based characteristics to differentiate participants at high risk of suicide from depressed and healthy cohorts, whereas 33% (7/21) of the studies used power spectral densities, 29% (6/21) used mel-frequency cepstral coefficients, 24% (5/21) used glottal cycle characteristics, and 14% (3/21) used timing patterns of speech. The highest median level of accuracy was attained using timing patterns of speech (85.5%), followed by power spectral densities (81.5%). Both the minimum and maximum levels of accuracy resulted from the use of power spectral densities (30.1% and 98.1%, respectively). In total, 19% (4/21) of the studies used vocal characteristics from a mixture of categories. For those studies that reported the levels of classification accuracy (15/21, 71%), median accuracies are reported in Table 3.

Table 3. Median classification accuracies of the voice signal characteristics selected in each study.

Primary feature	Studies	Accuracy (%), range	Accuracy (%), median
Frequency-based	• Campbell [24] • Francea et al [30] • Ozdas et al [34] • Beloualia et al [28] • Pestiana et al [37] • Figueroa Saavedra et al [29] • Scherera et al [22] • Sinha [26] • Venek[a] et al [23]	61.0-85.0	77.3
Power spectral densities	• France et al [30] • Nik Hashima et al [32] • Keskinpala et al [31] • Sanadi [25] • Yingthawornsuk et al [39] • Yingthawornsuk et al [40] • Yingthawornsuk and Shiavi [41]	30.1-98.1	81.5
Mel-frequency cepstral coefficients	• Belouali et al [28] • Nik Hashim et al [32] • Keskinpala et al [31] • Ozdas et al [35] • Subari et al [38] • Yingthawornsuk et al [40]	60.0-90.0	78.3
Glottal cycle characteristics	• Belouali et al [28] • Ozdasa et al [35] • Pestian et al [37] • Scherera et al [15] • Venek et al [23]	60.0-85.0	78.9
Timing patterns of speech	• Nik Hashim et al [32] • Nik Hashim et al [33] • Scherer et al [22]	66.0-100.0	85.5

[a] Combined with other voice biometrics.

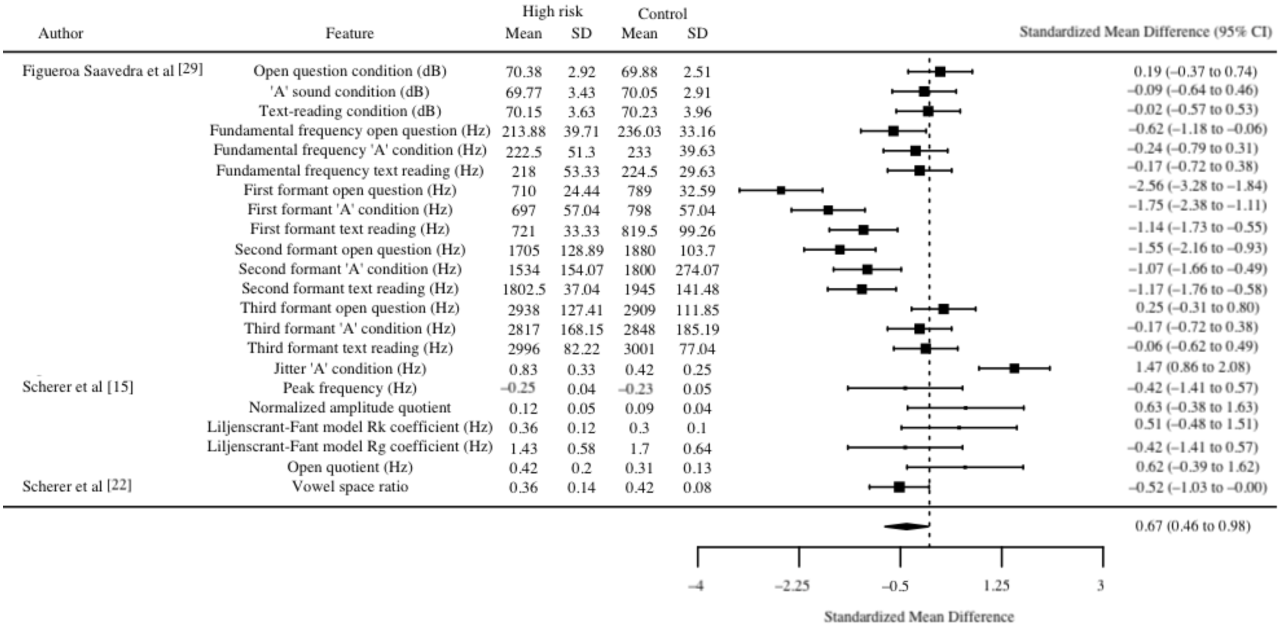
A Comparison of 22 Measures for Identifying High Risk of Suicide

A random-effects model meta-analysis was used to compare 22 different measures nested within 14% (3/21) of the publications. The included studies [15,22,29] involved 80 participants and 22 different vocal characteristics. The standardized mean difference for each vocal characteristic is summarized in the forest plot in Figure 3.

Positive standardized mean differences suggest higher levels of the vocal characteristic in the high–suicide-risk cohort, whereas negative standardized mean differences suggest that higher levels of the characteristic are found instead in the low-risk group.

A subgroup formed by the frequencies of the first and second formants applied to each of the 3 conditions reported by Figueroa Saavedra et al [29] was significant in differentiating between participants with and without elevated suicide risk. These significant negative standardized mean differences suggest that those at high risk of suicide feature lower vocal tract resonance frequencies, specifically in the lower formant regions rather than in the higher formant regions (ie, above the second formant of frequencies). Also of note were jitter values in the held "A" vowel condition in the study by Figueroa Saavedra et al [29]. Jitter in this condition yielded significant positive differentiation for participants with and without high suicide risk, suggesting that those at higher risk of suicide exhibited higher levels of roughness or hoarseness of articulation.

Figure 3. Random-effects meta-analysis forest plot illustrating 22 vocal characteristics in 3 studies.



Author	Feature	High risk Mean	High risk SD	Control Mean	Control SD	Standardized Mean Difference (95% CI)
Figueroa Saavedra et al [29]	Open question condition (dB)	70.38	2.92	69.88	2.51	0.19 (–0.37 to 0.74)
	'A' sound condition (dB)	69.77	3.43	70.05	2.91	–0.09 (–0.64 to 0.46)
	Text-reading condition (dB)	70.15	3.63	70.23	3.96	–0.02 (–0.57 to 0.53)
	Fundamental frequency open question (Hz)	213.88	39.71	236.03	33.16	–0.62 (–1.18 to –0.06)
	Fundamental frequency 'A' condition (Hz)	222.5	51.3	233	39.63	–0.24 (–0.79 to 0.31)
	Fundamental frequency text reading (Hz)	218	53.33	224.5	29.63	–0.17 (–0.72 to 0.38)
	First formant open question (Hz)	710	24.44	789	32.59	–2.56 (–3.28 to –1.84)
	First formant 'A' condition (Hz)	697	57.04	798	57.04	–1.75 (–2.38 to –1.11)
	First formant text reading (Hz)	721	33.33	819.5	99.26	–1.14 (–1.73 to –0.55)
	Second formant open question (Hz)	1705	128.89	1880	103.7	–1.55 (–2.16 to –0.93)
	Second formant 'A' condition (Hz)	1534	154.07	1800	274.07	–1.07 (–1.66 to –0.49)
	Second formant text reading (Hz)	1802.5	37.04	1945	141.48	–1.17 (–1.76 to –0.58)
	Third formant open question (Hz)	2938	127.41	2909	111.85	0.25 (–0.31 to 0.80)
	Third formant 'A' condition (Hz)	2817	168.15	2848	185.19	–0.17 (–0.72 to 0.38)
	Third formant text reading (Hz)	2996	82.22	3001	77.04	–0.06 (–0.62 to 0.49)
	Jitter 'A' condition (Hz)	0.83	0.33	0.42	0.25	1.47 (0.86 to 2.08)
Scherer et al [15]	Peak frequency (Hz)	–0.25	0.04	–0.23	0.05	–0.42 (–1.41 to 0.57)
	Normalized amplitude quotient	0.12	0.05	0.09	0.04	0.63 (–0.38 to 1.63)
	Liljenscrant-Fant model Rk coefficient (Hz)	0.36	0.12	0.3	0.1	0.51 (–0.48 to 1.51)
	Liljenscrant-Fant model Rg coefficient (Hz)	1.43	0.58	1.7	0.64	–0.42 (–1.41 to 0.57)
	Open quotient (Hz)	0.42	0.2	0.31	0.13	0.62 (–0.39 to 1.62)
Scherer et al [22]	Vowel space ratio	0.36	0.14	0.42	0.08	–0.52 (–1.03 to –0.00)
						0.67 (0.46 to 0.98)

What Are the Methodological Specifications Used?

Preprocessing is an important stage that occurs before the classification of vocal characteristics. This involves modifications to the voice signal to ensure greater precision in isolating its specific characteristics. Multimedia Appendix 5 illustrates an ideal preprocessing workflow.

The reviewed studies used a range of software for preprocessing and analysis of the vocal characteristics, including Microsound Editor [30,34,35,38,47] to identify and remove silence segments, MATLAB [22,23,25,26,30,39,48], COVAREP [22,23,37,49], and Praat [29,50] to facilitate subsequent analyses.

Most studies (11/21, 52%) [24,27,30,31,34-36,38-41] first converted the signal from analog to digital using a 16-bit recording at a sampling rate of 10 kHz. However, since 2010, recordings were sampled at higher rates. Venek et al [23] and Scherer et al [15,22], for instance, sampled speech at 16 kHz, whereas the remaining studies (19/21, 90%) [25,26,29,32,51] sampled it at 44.1 kHz.

All studies (21/21, 100%) then used a band-pass antialiasing filter to restrict the digital signal to a frequency range of 0 to 5000 Hz. Campbell [24] was the only author to analyze recordings sourced from telephone calls; thus, a band-pass filter restricting the frequency range to between 300 and 3000 Hz was automatically applied. After filtering, the signal was normalized [25,26,30,35,36,39-41] and detrended [25,30,39-41] to facilitate comparison between speakers and isolate the variable signal components, respectively.

Following these steps, the voice signal was differentiated between voiced and unvoiced types in 14% (3/21) of the studies. Subari et al [38] categorized voiced segments by the presence of cepstral peaks, Ozdas et al [34] differentiated between voiced and unvoiced signals using a discrete wavelet transform, whereas Sinha [26] adapted this approach to include 5 band-pass filters instead that selectively identified signal energies corresponding to each subband.

As noted, power spectral densities were investigated in several studies (6/21, 29%) [26,30,31,39-41]. Power spectral densities are derived from short-windowed segments of the voice signal. All studies (21/21, 100%) applied nonoverlapping Hamming windows to filter each 40- or 51.2-millisecond signal segment. A total of 10% (2/21) of the studies [38] used linear predictive coding applied to 15- and 25.6-millisecond segment durations to derive the first 3 formants and bandwidths.

Finally, most studies (7/21, 33%) used quadratic discriminant analysis to classify the voice signals of participants at high risk of suicide from other cohorts, and 57% (12/21) of the studies used either maximum likelihood, linear discriminant analysis, or support vector machines. As demonstrated in Table 4, the highest median level of accuracy was obtained using quadratic discriminant analysis. Both the minimum and maximum levels of accuracy were also recorded using quadratic discriminant analysis (21.4% and 100%, respectively). The median levels of classification accuracy are summarized in Table 4.

Table 4. Median accuracies achieved by classification algorithm.

Algorithm	Studies	Classification accuracy (%)	
		Range	Median
Maximum likelihood	- Ozdas et al [34] - Ozdas et al [35] - Ozdas et al [36] - Subari et al [38]	60.0-85.0	80.0
Linear discriminant analysis	- France et al [30] - Nik Hashim et al [51] - Sanadi [25] - Sinha [26]	30.1-98.1	79.7
Quadratic discriminant analysis	- Nik Hashim et al [51] - Keskinpala et al [31] - Sanadi [25] - Sinha [26] - Yingthawornsuk et al [39] - Yingthawornsuk et al [40] - Yingthawornsuk and Shiavi [41]	21.4-100.0	85.4
Hierarchical mixed model	- Scherer et al [15]	69.0-81.0	75.0
Support vector machine	- Pestian et al [37] - Scherer et al [15] - Venek et al [23] - Belouali et al [28]	61.0-85.0	75.9

Risk of Bias and Quality Assessment

The risk of bias was assessed to explore the variability in the quality of the included publications. None of the included studies were at low risk of bias. In 43% (9/21) of the studies, the risk of bias was moderate [15,22,23,26,28-30,32,36-41], whereas in the remaining 57% (12/21), the risk of bias was assessed as high [24,25,27,31,33-35]. The main sources of bias were confounding factors, selection of participants, and selective reporting of results. The quality of evidence of most (18/21, 86%) of the included studies [15,22,23,27-37,39-41] was assessed with a rating of 3, whereas 14% (3/21) of the studies [17,24,52] were assessed at the lowest rating of 5.

Discussion

Principal Findings

A systematic review (1995-2021) was undertaken with 2 objectives in mind: to identify which vocal characteristics could accurately differentiate between individuals at high risk of suicide compared with those at lower risk and identify the methodological specifications that inform the derivation of each vocal characteristic and the classification accuracies that result.

A number of vocal characteristics were found to differentiate between high risk of suicide and comparison cohorts with a high level of accuracy. Of note were the median accuracies obtained using the timing patterns of speech (median accuracy 85.5%) and power spectral densities (median accuracy 81.5%). Furthermore, a random-effects meta-analysis that included 22 vocal characteristics from 14% (3/21) of the studies revealed that frequencies within the lower formants (1 and 2) and jitter provided significant standardized mean differences between high- and low–suicide-risk signals, suggesting that participants at high risk of suicide may have lower vocal tract resonance frequencies while speaking with greater roughness of speech.

These results are broadly consistent with several recent investigations that have also found significant increases in lower formant frequencies under stressful conditions, suggesting a reduction in articulatory clarity [17]; decreases in the quantity of speech among those at high risk of suicide [53]; and changes in jitter, a measure of cycle-to-cycle variation in the fundamental frequency that decreases under anxiety-producing conditions [17].

The findings of this study that mel-frequency cepstral coefficients characterize muscle tension and control of the vocal tract, a measure particularly sensitive to changes in stress, are also supported by previous research [54]. However, although increases in root mean squared amplitude or loudness have been found among speakers at high risk of suicide in more recent studies [55], the use of this variable was limited to only select studies reviewed (2/21, 10%) [30,37].

This review also aimed to identify the system specifications used to derive each vocal characteristic and their levels of accuracy. All studies (21/21, 100%) were found to adopt a similar workflow of preprocessing steps that broadly included (1) conversion from analog to digital signals, (2) band-pass filtering, (3) normalization and detrending, (4) differentiation between voiced and unvoiced signals, (5) removal of silent passages, and (6) signal segmentation before classification.

Although most studies (12/21, 57%) used band-pass filtering to remove frequencies >5000 Hz, only Campbell [24] used recordings sourced from telephone calls. As noted by the author, these sources automatically filter signals to between 300 and 3000 Hz. Although noise is reduced, this approach also removes the fundamental frequency from the signal signature but is known to overestimate the first formant frequency values by as much as 13% [56]. These may be important considerations for future studies that aim to source data from more novel settings such as telephone helplines or conversational agents.

Most studies (14/21, 67%) discussed the use of signal normalization to ensure a comparison between speakers. However, only in the study by Subari et al [38] was the effect on the overall accuracy of different forms of normalization investigated. Both a maximum likelihood–derived warping factor and normalization based on median third formant values were optimized based on the levels of classification accuracy. The authors noted that the formant-derived approach is preferable given its lower computational load and consideration of the sex of the participant. Normalization is a crucial consideration when the vocal tract of speakers can differ by as much as 7 cm [57].

Detrending was discussed in some studies (5/21, 24%). It is presumed that, by removing the mean signal, the authors sought to reveal the nonstationary signal components that might better differentiate one speaker from another. Although voice signals are known to be stationary only over short time frames (<40 ms) [58], several studies (8/21, 38%) used signal segmentation in excess of 50 milliseconds. The strategy of mean removal with potentially nonstationary signals risks dampening low-frequency sounds while attenuating high-frequency sounds and can also introduce secondary artifacts that may cloud ongoing analyses [58]. The preference in several studies (5/21, 24%) to detrend via discrete wavelet transform overcomes many of the aforementioned issues and appears well suited to the analysis of nonstationary signals at longer time frames of capture.

Given the differences in the frequency and amplitude spectra between voiced and unvoiced segments of speech, it is unsurprising that most of the reviewed publications (11/21, 52%) opted to differentiate between these signal types before classification. In total, 3 different approaches to voiced or unvoiced signal differentiation were used: approximation of voiced signals via the presence of cepstral peaks; frequency mapping via discrete wavelet transform using the highest-scale wavelets (2^5) to categorize voiced signals; and selective band-pass filtering with frequencies >2500 Hz categorized as unvoiced, whereas signals between 320 and 2499 Hz were categorized as voiced. Further investigation is required to determine which approach best optimizes accuracies of classification; however, selective band-pass filtering has the advantage of not altering the signal in any way.

Regarding those studies that analyzed power spectral densities (9/21, 43%), nonoverlapping Hamming windows were the preferred approach to derive the short–time frame Fourier transform, converting the signal from the time to the frequency domain. This nonstandard approach has the effect of capturing frequencies at regular intervals corresponding to window width while introducing high frequencies as each window tapers into the next. The standard approach that incorporates overlapping windows, smoothing the effects at the tails, seems preferable to the nonstandard approach commonly used [59].

In only a minority of the reviewed studies (3/21, 14%) was supervised machine learning used (support vector machine). However, the highest median levels of accuracy were obtained not when these advanced forms of classification were used but rather when unsupervised quadratic discriminant analysis was used. Contrary to the much touted superiority of supervised machine learning methods, our findings suggest that higher levels of accuracy were instead obtained using less complex classification approaches. However, further investigation using more sophisticated approaches such as neural networks is clearly warranted.

In only the study by Scherer et al [15] was a mixed-effects model used. This approach might better capture the correlated structure of voice signal segments and better account for intraspeaker variance than other approaches.

Future Directions and Implications for Practice

Suicide risk has been treated as static, stable, and invariant following initial assessment. Recent studies demonstrate that suicide risk can, in fact, change dramatically over time, suggesting that future studies might use insights from ecological momentary assessment [52,60]. Alternatively, future studies could adopt the approach of Campbell [24] by using trained personnel to assess the changing level of suicide risk across time within each recording as well as between recordings. Such approaches might better reflect the real-time change in suicidality, acknowledging that individuals frequently cycle in and out of risk.

In the reviewed literature, risk assessments were typically performed by coauthors (eg, Silverman in the study by Campbell [24] or Salomon in the study by Sinha [26]), an approach that may bias the objective assessment of risk. Future research might use multiple assessors of suicide risk where measures of interrater reliability can be analyzed and possible biases can be isolated and addressed.

Our analysis of the preprocessing workflows suggests that greater transparency regarding methodological considerations is urgently required. The reviewed publications were clearly aimed at an informed and knowledgeable engineering readership, and it was common to refer to complex methods using technical terms (eg, windowing). It would assist in reproducibility to understand preprocessing decisions in greater detail (ie, window type).

Certain vocal characteristics have been proven to more accurately differentiate between high and low risk of suicide. In particular, the timing patterns of speech and the vowel space occupied by different speaker articulations hold considerable promise. Also of note are insights derived from power spectral densities and frequency-related categories, underused methods such as the Liljencrants-Fant model of glottal flow, and mel-frequency cepstral coefficients. However, as demonstrated by Pestian et al [37] and Venek et al [23], the future of research in this field leans toward combining multiple features within high-powered machine learning algorithms such as support vector machine, although it should be noted that lower-powered approaches appeared in this review to yield greater levels of accuracy (ie, quadratic discriminant analysis).

The power of these advanced machine learning algorithms can only be used with an adequately powered sample. Cummins et al [14] called for greater collaboration between research teams to address this ongoing problem. An alternative approach might be to secure greater industry partnerships and look to novel settings with high call volumes such as telemental health. Given that a recent review found poor support for conventional suicide screening methods [10], there is a clear case for incorporating voice signal–informed analysis into existing telehealth and other e-services, in particular suicide helplines. These settings typically have large call volumes that increasingly feature elevated risk of suicide, particularly in the COVID-19 era. However, such collaborations also raise other ethical issues such as how best to safeguard callers' rights to privacy and secure consent.

Limitations

There are some limitations to this review. Of note was the lack of specificity in the definitions of high risk of suicide. In only 43% (9/21) of the studies analyzed, the high-risk cohorts were truly reflective of imminent risk of suicide. These studies used recordings of participants sourced from the Silverman database of suicide notes left by patients who had either attempted or completed suicide or, alternatively, from the Cincinnati Children's Hospital interview corpus, where participants were recruited and interviewed immediately following presentation at emergency departments with acute suicidality. In the remaining studies (12/21, 57%), participants were assigned to the high-risk cohort based on cutoff scores on diverse psychometric tests, including the Beck Depression Inventory and Hamilton Depression Rating Scale. Several large-scale reviews [10,61-63] attest to the low precision and recall of suicide rating scales, suggesting that participants assigned to cohorts with high risk of suicide in these studies may also feature a proportion of false positives and, conversely, a proportion of false negatives in the control groups.

There was also a diversity of vocal characteristics trialed. Across the reviewed publications, it was challenging to find the same set of features replicated in other samples. It was more common to find a single characteristic combined with others to optimize discrimination between levels of suicide risk. This presented difficulties in determining which features might be reliably accurate across different settings.

Except for a notable large-scale multicenter trial [37], most of the reviewed studies (17/21, 81%) involved small samples, typically <60 participants, sometimes divided between 3 comparison groups [38,41]. As highlighted by Button et al [64], small-sample research is plagued by a number of issues, namely, reduced power, low reproducibility of results, and reductions in the likelihood that the results obtained reflect a true effect. Undoubtedly, these problems are amplified when the sample size of the comparison groups is <10, as was the case in several of the studies reviewed (8/21, 38%) [24,26,34-36,38,40,41]. Also of consideration were the high proportion of studies involving psychiatric inpatients (17/21, 81%) and the homogeneity of the databases from which participants were recruited, further limiting generalizability.

Also of note were the controlled conditions within which the trials took place. It was common for participants to be invited into a room away from extraneous noise and asked to read prescripted text such as the rainbow passage (eg, Yingthawornsuk et al [39]) or to articulate prolonged vowel sounds (eg, Scherer et al [22]). Although these research protocols increase the likelihood of uncovering candidate vocal characteristics, they also reduce the generalizability of the research findings to other settings, particularly to telephony and other novel eHealth-based applications where these controls are impractical to implement and noise is the rule rather than the exception. Except for the study by Campbell [24], no studies sourced participants from these more ecologically valid settings.

There was a paucity of detail relating to specific preprocessing elements supported by the high risk of bias prevailing among the reviewed studies. One of the cardinal requirements for cumulative science is that methodologies are replicable [65]. Publications are increasingly restricting the number of allowable words. Prospective authors might be tempted to limit descriptions of the methodology in favor of the results and discussion. However, most publications also allow for appendixes that can provide supplementary information relating to the methodology.

However, there were also notable strengths to this review. This review has expanded upon the findings of Cummins et al [14] and Homan et al [16] in important ways. We have updated the research findings to 2021, and based on the accuracy of discrimination between levels of suicide risk, we were able to identify a number of promising candidate vocal characteristics that warrant further investigation. We were also able to identify and discuss a number of preprocessing steps used before the classification of voice signals.

Conclusions

The data indicate that several characteristics successfully differentiate between individuals at high and low risk of suicide. An analysis of power spectral density subbands yielded high accuracies of discrimination between comparison groups (eg, 90.3% accuracy in the study by Yingthawornsuk et al [40]); however, the studies that used power spectral densities disagreed on whether the lower- or higher-frequency subbands were of key importance and also disagreed on the need for single or combined feature analyses. Second, several studies (4/21, 19%) found higher formant frequencies and a narrowing of bandwidth among those at elevated risk of suicide [22,23,29,30]. Higher levels of predictive accuracy were found when formant features were combined with other features (eg, 80% accuracy in the study by France et al [30]). Third, Nik Hashim et al [32,51] and Scherer et al [22] found that the timing patterns of speech in speakers at elevated risk of suicide differed in a number of important ways from those of speakers at low risk of suicide. In particular, pauses were protracted, whereas certain vowel sounds were held for longer periods among those at elevated risk of suicide. Fourth, both the study by Anunvrapong and Yingthawornthuk [27] and the study by Ozdas et al [35] found that the analysis of mel-frequency cepstral coefficients—which attempts to mimic the energy spectrum of human hearing—successfully differentiated between speakers at high and low risk of suicide. However, a reduced filter bank (first 4 frequencies) yielded greater accuracies. Finally, both Scherer et al [22] and Venek et al [23] found that certain coefficients of the Liljencrants-Fant model of glottal flow significantly differentiated between high and low risk of suicide, suggesting that those at high risk of suicide often speak in breathier tones. This was particularly apparent among adolescents.

Although this systematic review revealed a number of limitations in the current literature in this field, the level of accuracy achieved is promising, suggesting that future research, particularly in more novel areas of telemental health, holds considerable promise for the detection and prevention of suicide in the community.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 31 KB - biomedeng_v7i2e42386_app1.docx]

Multimedia Appendix 2
Representative search strategies. STFT: short–time frame Fourier transform.
[PNG File , 104 KB - biomedeng_v7i2e42386_app2.png]

Multimedia Appendix 3
Risk-of-bias assessments.
[DOCX File , 23 KB - biomedeng_v7i2e42386_app3.docx]

Multimedia Appendix 4
Review data.
[XLSX File (Microsoft Excel File), 1504 KB - biomedeng_v7i2e42386_app4.xlsx]

Multimedia Appendix 5
Ideal preprocessing workflow.
[DOCX File , 14 KB - biomedeng_v7i2e42386_app5.docx]

References

1. Duffy S, Lee TH. In-person health care as option B. N Engl J Med 2018 Jan 11;378(2):104-106. [doi: 10.1056/NEJMp1710735] [Medline: 29320653]
2. Lawes-Wickwar S, McBain H, Mulligan K. Application and effectiveness of telehealth to support severe mental illness management: systematic review. JMIR Ment Health 2018 Nov 21;5(4):e62 [FREE Full text] [doi: 10.2196/mental.8816] [Medline: 30463836]
3. Hunt G. Doorstop interview on 27 November 2020. Ministers, Department of health and aged care. 2020 Nov 27. URL: https://www.health.gov.au/ministers/the-hon-greg-hunt-mp/media/doorstop-interview-on-27-november-2020 [accessed 2021-02-09]
4. Pfender E. Mental health and COVID-19: implications for the future of telehealth. J Patient Exp 2020 Aug;7(4):433-435 [FREE Full text] [doi: 10.1177/2374373520948436] [Medline: 33062854]
5. Jaffe DH, Lee L, Huynh S, Haskell TP. Health inequalities in the use of telehealth in the United States in the lens of COVID-19. Popul Health Manag 2020 Oct;23(5):368-377. [doi: 10.1089/pop.2020.0186] [Medline: 32816644]
6. World Health Organisation. Suicide in the world: Global health estimates. World Health Organisation. 2019. URL: https://apps.who.int/iris/bitstream/handle/10665/326948/WHO-MSD-MER-19.3-eng.pdf [accessed 2021-02-12]
7. Crnek-Georgeson KT, Wilson LA, Page A. Factors influencing suicide in older rural males: a review of Australian studies. Rural Remote Health 2017;17(4):4020-4024 [FREE Full text] [doi: 10.22605/RRH4020] [Medline: 29031280]
8. 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]
9. Turecki G, Brent DA. Suicide and suicidal behaviour. Lancet 2016;387(10024):1227-1239 [FREE Full text] [doi: 10.1016/S0140-6736(15)00234-2] [Medline: 26385066]
10. Franklin JC, Ribeiro JD, Fox KR, Bentley KH, Kleiman EM, Huang X, et al. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. Psychol Bull 2017;143(2):187-232. [doi: 10.1037/bul0000084] [Medline: 27841450]

11. Sudol K, Mann JJ. Biomarkers of suicide attempt behavior: towards a biological model of risk. Curr Psychiatry Rep 2017;19(6):31. [doi: 10.1007/s11920-017-0781-y] [Medline: 28470485]
12. Qu C, Sas C, Daudén Roquet C, Doherty G. Correction: functionality of top-rated mobile apps for depression: systematic search and evaluation. JMIR Ment Health 2020 Feb 21;7(2):e18042 [FREE Full text] [doi: 10.2196/18042] [Medline: 32130145]
13. Balcombe L, De Leo D. An integrated blueprint for digital mental health services amidst COVID-19. JMIR Ment Health 2020 Jul 22;7(7):e21718 [FREE Full text] [doi: 10.2196/21718] [Medline: 32668402]
14. Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF. A review of depression and suicide risk assessment using speech analysis. Speech Commun 2015;71:10-49. [doi: 10.1016/j.specom.2015.03.004]
15. Scherer S, Pestian J, Morency LP. Investigating the speech characteristics of suicidal adolescents. In: Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. 2013 Oct 21 Presented at: 2013 IEEE International Conference on Acoustics, Speech Signal Process; 2013; Vancouver. [doi: 10.1109/icassp.2013.6637740]
16. Homan S, Gabi M, Klee N, Bachmann S, Moser A, Duri' M, et al. Linguistic features of suicidal thoughts and behaviors: a systematic review. Clin Psychol Rev 2022 Jul;95:102161 [FREE Full text] [doi: 10.1016/j.cpr.2022.102161] [Medline: 35636131]
17. Teferra BG, Borwein S, DeSouza DD, Simpson W, Rheault L, Rose J. Acoustic and linguistic features of impromptu speech and their association with anxiety: validation study. JMIR Ment Health 2022 Jul 08;9(7):e36828 [FREE Full text] [doi: 10.2196/36828] [Medline: 35802401]
18. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med 2009 Jul 21;6(7):e1000100 [FREE Full text] [doi: 10.1371/journal.pmed.1000100] [Medline: 19621070]
19. Iyer R, Meyer D, Nedeljkovic M, Seabrook L. Speech biomarkers of suicide risk: a systematic review. PROSPERO. 2020. URL: https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42020167413 [accessed 2020-01-31]
20. NVIVO. NVIVO QSR International. URL: https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home [accessed 2020-02-01]
21. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016 Oct 12;355:i4919 [FREE Full text] [doi: 10.1136/bmj.i4919] [Medline: 27733354]
22. Scherer S, Morency L, Gratch J, Pestian J. Reduced vowel space is a robust indicator of psychological distress: a cross-corpus analysis. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing.: IEEE; 2015 Presented at: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2015; Brisbane. [doi: 10.1109/ICASSP.2015.7178880]
23. Venek V, Scherer S, Morency L, Rizzo AS, Pestian J. Adolescent suicidal risk assessment in clinician-patient interaction. IEEE Trans Affective Comput 2017 Apr 1;8(2):204-215. [doi: 10.1109/TAFFC.2016.2518665]
24. Campbell L. Statistical Characteristics of Fundamental Frequency Distributions in the Speech of Suicidal Patients. Nashville: Vanderbilt University; 1995.
25. Sanadi W, Hasan WA. Acoustic analysis of speech based on power spectral density features in detecting suicidal risk among female patients dissertation. Vanderbilt University Institutional Repository. 2011. URL: https://etd.library.vanderbilt.edu/etd-03252011-142343 [accessed 2022-12-09]
26. Sinha A. Human and machine recognition of the vocal characteristics of suicide. Vanderbilt University. 2013 Dec. URL: https://ir.vanderbilt.edu/bitstream/handle/1803/15036/SINHA.pdf?sequence=1 [accessed 2022-12-09]
27. Anunvrapong P, Yingthawornthuk T. Characterisation MCFF in depressed speech sample as assessment of suicidal risk. In: Proceedings of the 2014 International Conference on Advanced Computational Technologies & Creative Media (ICACTCM). 2014 Presented at: International Conference on Advanced Computational Technologies & Creative Media (ICACTCM?) Aug; 2014; Pattaya. [doi: 10.15242/IIE.E0814549]
28. Belouali A, Gupta S, Sourirajan V, Yu J, Allen N, Alaoui A, et al. Acoustic and language analysis of speech for suicidal ideation among US veterans. BioData Min 2021 Feb 02;14(1):11 [FREE Full text] [doi: 10.1186/s13040-021-00245-y] [Medline: 33531048]
29. Figueroa Saavedra C, Otzen Hernández T, Alarcón Godoy C, Ríos Pérez A, Frugone Salinas D, Lagos Hernández R. Association between suicidal ideation and acoustic parameters of university students' voice and speech: a pilot study. Logoped Phoniatr Vocol 2021 Jul;46(2):55-62. [doi: 10.1080/14015439.2020.1733075] [Medline: 32138570]
30. France DJ, Shiavi RG, Silverman S, Silverman M, Wilkes DM. Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Trans Biomed Eng 2000;47(7):829-837. [doi: 10.1109/10.846676] [Medline: 10916253]
31. Keskinpala H, Yingthawornsuk T, Wilkes D, Shiavi R, Salomon R. Screening for high risk suicidal states using mel-cepstral coefficients and energy in frequency bands. In: Proceedings of the 15th European Signal Processing Conference. 2007 Presented at: 2007 15th European Signal Processing Conference; Sep 03-07, 2007; Poznan URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7099204&isnumber=7067186 [doi: 10.21437/interspeech.2007-144]
32. Nik HN, Wilkes M, Salomon R. Timing patterns of speech as potential indicators of near-term suicidal risk. Int J Multidisciplinary Curr Res 2015;3:1102-1116 [FREE Full text]

33. Hashim NW, Wilkes M, Salomon R, Meggs J, France DJ. Evaluation of voice acoustics as predictors of clinical depression scores. J Voice 2017;31(2):256.e1-256.e6. [doi: 10.1016/j.jvoice.2016.06.006]
34. Ozdas A, Shiavi R, Silverman S, Silverman M, Wilkes D. Analysis of fundamental frequency for near term suicidal risk assessment. In: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics. 2000 Presented at: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics; Oct 08-11, 2000; Nashville, TN, USA. [doi: 10.1109/icsmc.2000.886379]
35. Ozdas A, Shiavi RG, Wilkes DM, Silverman MK, Silverman SE. Analysis of vocal tract characteristics for near-term suicidal risk assessment. Methods Inf Med 2018;43(01):36-38. [doi: 10.1055/s-0038-1633420]
36. Ozdas A, Shiavi RG, Silverman SE, Silverman MK, Wilkes DM. Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk. IEEE Trans Biomed Eng 2004;51(9):1530-1540. [doi: 10.1109/TBME.2004.827544] [Medline: 15376501]
37. Pestian JP, Sorter M, Connolly B, Bretonnel Cohen K, McCullumsmith C, Gee JT, et al. A machine learning approach to identifying the thought markers of suicidal subjects: a prospective multicenter trial. Suicide Life Threat Behav 2017;47(1):112-121. [doi: 10.1111/sltb.12312] [Medline: 27813129]
38. Subari K, Wilkes D, Shiavi R, Silverman S, Silverman M. Comparison of speaker normalization techniques for classification of emotionally disturbed subjects based on voice. In: Proceedings of the IEEE EMBS Conference on Biomedical Engineering and Sciences. 2010 Presented at: IEEE EMBS Conference on Biomedical Engineering and Sciences; Nov 30-Dec 2, 2010; Kuala Lumpur. [doi: 10.1109/iecbes.2010.5742248]
39. Yingthawornsuk T, Keskinpala K, France D. Objective estimation of suicidal risk using vocal output characteristics. In: Proceedings of the Interspeech - 9th International Conference on Spoken Language Processing. 2006 Presented at: Interspeech - 9th International Conference on Spoken Language Processing; Sep 17-21, 2006; Pittsburgh URL: https://www.isca-speech.org/archive/archive_papers/interspeech_2007/i07_0766.pdf [doi: 10.21437/interspeech.2006-231]
40. Yingthawornsuk T, Keskinpala K, Wilkes D, Shiavi R, Salomon R. Direct Acoustic Feature Using Iterative EM Algorithm and Spectral Energy for Classifying Suicidal Speech. In: Proceedings of the Interspeech - 8th International Conference on Spoken Language Processing. 2007 Presented at: Interspeech - 8th International Conference on Spoken Language Processing; Aug 27-31, 2007; Antwerp, Belgium. [doi: 10.21437/interspeech.2007-144]
41. Yingthawornsuk T, Shiavi R. Distinguishing depression and suicidal risk in men using GMM based frequency contents of affective vocal tract response. In: Proceedings of the International Conference on Control, Automation and Systems. 2008 Presented at: International Conference on Control, Automation and Systems; Oct 14-17, 2008; Seoul. [doi: 10.1109/iccas.2008.4694621]
42. Rodgers M, Sowden A, Petticrew M, Arai L, Roberts H, Britten N, et al. Testing methodological guidance on the conduct of narrative synthesis in systematic reviews. Evaluation 2009;15(1):49-73. [doi: 10.1177/1356389008097871]
43. Henry WP, Strupp HH, Butler SF, Schacht TE, Binder JL. Effects of training in time-limited dynamic psychotherapy: changes in therapist behavior. J Consulting Clin Psychol 1993;61(3):434-440. [doi: 10.1037/0022-006x.61.3.434]
44. Hollon SD, DeRubeis RJ, Evans MD, Wiemer MJ, Garvey MJ, Grove WM, et al. Cognitive therapy and pharmacotherapy for depression. Singly and in combination. Arch Gen Psychiatry 1992;49(10):774-781. [doi: 10.1001/archpsyc.1992.01820100018004] [Medline: 1417429]
45. Silverman M, Silverman S. From sound to silence: a preliminary investigation of the use of vocal parameters in the prediction of near-term suicidal risk. J Med Psychotherapy 2000:1-10 (forthcoming).
46. Gratch J, Arstein R, Lucas G, Stratou G. The distress analysis interview corpus of human and computer interviews. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation. 2014 Presented at: Ninth International Conference on Language Resources and Evaluation; May 26-31, 2014; Reykjavik. [doi: 10.1016/b978-0-12-410393-1.00011-9]
47. Microstudio. MTU Technology Unlimited (MTU). URL: http://www.mtu.com/upgrades/microstudio.htm [accessed 2020-03-12]
48. MATLAB version 7. The Mathworks Inc. URL: https://au.mathworks.com/products/matlab.html [accessed 2020-03-10]
49. Degottex G, Drugman T, Raitio T, Scherer S. COVAREP A collaborative voice analysis repository for speech technologies. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal processing (ICASSP). 2014 Presented at: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); May 04-09, 2014; Florence. [doi: 10.1109/icassp.2014.6853739]
50. Moltu C, Stefansen J, Svisdahl M, Veseth M. Negotiating the coresearcher mandate - service users' experiences of doing collaborative research on mental health. Disabil Rehabil 2012;34(19):1608-1616. [doi: 10.3109/09638288.2012.656792] [Medline: 22489612]
51. Nik HN, Wilkes M, Salomon R, Meggs J. Analysis of timing pattern of speech as possible indicator for near-term suicidal risk and depression in male patients. Int Proceed Comp Sci Inf Tech 2012;58:6.
52. Hallensleben N, Spangenberg L, Forkmann T, Rath D, Hegerl U, Kersting A, et al. Investigating the dynamics of suicidal ideation. Crisis 2018;39(1):65-69. [doi: 10.1027/0227-5910/a000464] [Medline: 28468557]

53. Galatzer-Levy I, Abbas A, Ries A, Homan S, Sels L, Koesmahargyo V, et al. Validation of visual and auditory digital markers of suicidality in acutely suicidal psychiatric inpatients: proof-of-concept study. J Med Internet Res 2021 Jun 03;23(6):e25199 [FREE Full text] [doi: 10.2196/25199] [Medline: 34081022]
54. König A, Riviere K, Linz N, Lindsay H, Elbaum J, Fabre R, et al. Measuring stress in health professionals over the phone using automatic speech analysis during the COVID-19 pandemic: observational pilot study. J Med Internet Res 2021 Apr 19;23(4):e24191 [FREE Full text] [doi: 10.2196/24191] [Medline: 33739930]
55. Iyer R, Nedeljkovic M, Meyer D. Using voice biomarkers to classify suicide risk in adult telehealth callers: retrospective observational study. JMIR Ment Health 2022 Aug 15;9(8):e39807 [FREE Full text] [doi: 10.2196/39807] [Medline: 35969444]
56. Künzel HJ. Beware of the 'telephone effect': the influence of telephone transmission on the measurement of formant frequencies. Forensic Linguistics 2001;8(1):80-99. [doi: 10.1558/sll.2001.8.1.80]
57. Lammert AC, Narayanan SS. On short-time estimation of vocal tract length from formant frequencies. PLoS One 2015;10(7):e0132193 [FREE Full text] [doi: 10.1371/journal.pone.0132193] [Medline: 26177102]
58. Rabiner L, Schafer R. Theory and Applications of Digital Speech Processing. Upper Saddle River: Pearson; 2011.
59. Benesty J, Sondhi M, Huang Y. Introduction to speech processing. In: Springer Handbook of Speech Processing. Berlin: Springer; 2008:1-4.
60. Kleiman EM, Turner BJ, Fedor S, Beale EE, Huffman JC, Nock MK. Examination of real-time fluctuations in suicidal ideation and its risk factors: results from two ecological momentary assessment studies. J Abnorm Psychol 2017 Aug;126(6):726-738. [doi: 10.1037/abn0000273] [Medline: 28481571]
61. Lotito M, Cook E. A review of suicide risk assessment instruments and approaches. Mental Health Clinician 2015;5:216-223. [doi: 10.9740/mhc.2015.09.216]
62. Runeson B, Odeberg J, Pettersson A, Edbom T, Jildevik Adamsson I, Waern M. Instruments for the assessment of suicide risk: a systematic review evaluating the certainty of the evidence. PLoS One 2017 Jul 19;12(7):e0180292 [FREE Full text] [doi: 10.1371/journal.pone.0180292] [Medline: 28723978]
63. Carter G, Milner A, McGill K, Pirkis J, Kapur N, Spittal MJ. Predicting suicidal behaviours using clinical instruments: systematic review and meta-analysis of positive predictive values for risk scales. Br J Psychiatry 2017 Jun 02;210(6):387-395. [doi: 10.1192/bjp.bp.116.182717] [Medline: 28302700]
64. Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 2013;14(5):365-376. [doi: 10.1038/nrn3475] [Medline: 23571845]
65. Asendorpf JB, Conner M, De Fruyt F, De Houwer J, Denissen JJ, Fiedler K, et al. Recommendations for increasing replicability in psychology. Eur J Pers 2020;27(2):108-119. [doi: 10.1002/per.1919]

Abbreviations

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Edited by A Mavragani; submitted 01.09.22; peer-reviewed by P Upadhyaya, A Teles; comments to author 27.10.22; revised version received 01.12.22; accepted 01.12.22; published 22.12.22.

Please cite as:
Iyer R, Meyer D
Detection of Suicide Risk Using Vocal Characteristics: Systematic Review
JMIR Biomed Eng 2022;7(2):e42386
URL: https://biomedeng.jmir.org/2022/2/e42386
doi:10.2196/42386
PMID:38875684

Viewpoint

A Bayesian Network Concept for Pain Assessment

Omowunmi Sadik[1*], PhD; J David Schaffer[2*], PhD; Walker Land[3*], BA, MSc; Huize Xue[4*], BSc; Idris Yazgan[5*], PhD; Korkut Kafesçiler[6*], MD; Mürvet Sungur[7], MD

[1]Department of Chemistry & Environmental Science, New Jersey Institute of Technology, Newark, NJ, United States

[2]Institute of Justice and Well-being, College of Community and Public Affairs, State University of New York-Binghamton, Binghamton, NY, United States

[3]Department of Bioengineering, State University of New York, Binghamton, NY, United States

[4]Department of Physics, New Jersey Institute of Technology, Newark, NJ, United States

[5]Department of Biology, Kastamonu University, Kastamonu, Turkey

[6]Anesthesiology Clinics, Manisa State Hospital, Yunus Emre-Manisa, Turkey

[7]Microbiology Clinics, Manisa Merkez Efendi State Hospital, Yunus Emre Manisa, Turkey

[*]these authors contributed equally

Corresponding Author:
Omowunmi Sadik, PhD
Department of Chemistry & Environmental Science
New Jersey Institute of Technology
161 Warren Street
University Heights
Newark, NJ, 07102
United States
Phone: 1 9735962833
Email: SADIK@NJIT.EDU

Abstract

In this study, we propose an approach that provides a useful data summary related to a patient's experience of pain. Because pain is a very important but subjective phenomenon that currently has no calibratable method for assessing it, we suggest an approach that uses calibratable biomarker sensors with the patient's self-assessment of perceived pain. We surmise that such an approach may only be able to clearly distinguish between cases in which the available evidence is consistent. However, this information may provide clinicians with valuable insights, and as research progresses into how biomarkers are related to pain, more specific insights may emerge regarding how specific evidence inconsistencies may point to particular pain causes. We provide a brief overview of pain science, including the types of pain, contemporary pain theories, pain, and pain assessment techniques. Next, we present novel approaches to pain sensor development, including an overview of research on pain-related biomarker sensors and artificial intelligence methods for summarizing the evidence. We then provide some illustrations of the implementation of our approach. Some specifics are presented in the Methods section of this paper. For example, in a set of 379 patients, we observed 80% evidence of consistency and 5 types of inconsistencies. Information regarding the gender and individual differences in cyclooxygenase-2 and inducible nitric oxide synthase data on reported pain could contribute to the inconsistency. Different causes of inconsistencies are also attributed to cultural or temporal variability of cyclooxygenase-2 and inducible nitric oxide synthase (as well as their serum variation and half-life), visual analog scale, and other tools. We emphasize that this presentation is illustrative. Much work remains to be done before implementing and testing this approach in a clinically meaningful context.

(JMIR Biomed Eng 2022;7(2):e35711) doi:10.2196/35711

KEYWORDS

pain science; pain biomarkers; novel biosensors; Bayesian network; artificial intelligence; AI evidential reasoning; pain self-report; probability of pain levels; cyclooxygenase-2; COX-2; inducible nitric oxide synthase; iNOS

Pain and the Scope of This Work

Pain is largely subjective yet critical to assess for clinicians to provide proper care. A major challenge (much discussed in the literature) is that, at this time, there is no objectively calibratable way to measure pain. This study aims to present a novel approach to address this problem. Rather than seeking a specific calibratable pain scale, we propose an approach that combines calibratable measures of pain-related biomarkers with patient self-assessments of perceived pain using artificial intelligence (AI) methods that can at least provide a view of the extent to which available evidence is consistent. We postulate that with an appropriate set of evidence, we may discover specific patterns of evidence that will provide valuable insights that clinicians may use to improve pain care. We present a modest illustration of an approach to this problem by using two pain-related biomarkers, cyclooxygenase-2 (COX-2) and inducible nitrous oxide synthase (iNOS), and a Bayesian network (BN) model. The reader should keep in mind that this illustration is only an example and is not meant to be *the* answer. Much work remains to be done and will require a dedicated team of pain experts, including those knowledgeable in pain care and neurology, with biochemical or molecular biology of pain mechanisms.

A Brief Overview of Pain

Pain Defined

According to the International Association for the Study of Pain, pain is an unpleasant sensory and emotional experience associated with, or resembling that is associated with, actual or potential tissue damage [1]. One common distinction is between acute and chronic pain. Acute pain is typically caused by the stimulation of peripheral nerve endings because of inflammation or trauma or interference with nerve pathways (neuropathic) because of the nerve being severed (for example, during surgery). Tissue healing generally results in the cessation of acute pain. Chronic pain arises in acute pain situations, generally when the acute pain is very intense or lasts for an extended period. It is important to remember that in most situations, acute pain serves as a critical protective mechanism in preventing further tissue injury. By reducing the risk of continued trauma, the tissue can heal more rapidly and the pain will subside. The main challenge in dealing with intensive acute pain is to prevent the overuse of strong opioids, morphine, codeine, and cocaine, leading to addiction through the euphoria created using these medications.

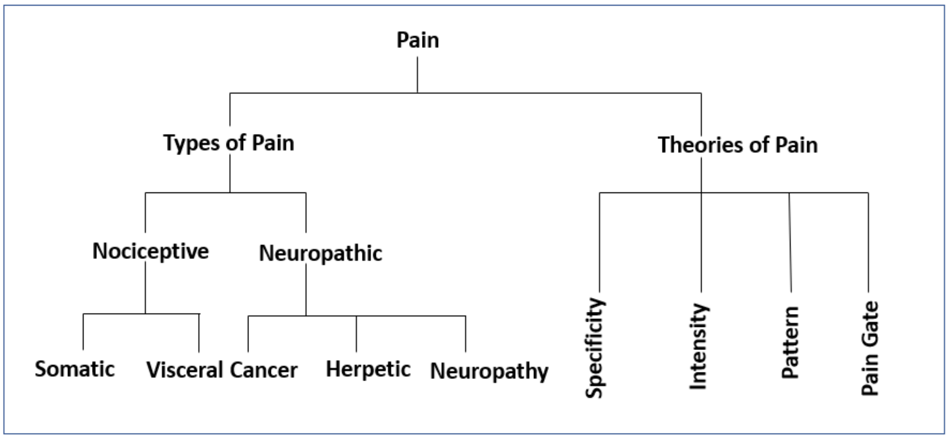
According to the National Health Interview Survey of 2019, a total of 50.2 million or approximately 20.5% of American adults experience chronic pain, with the most common examples being back pain and hip, knee, or foot pain [2]. Chronic low back pain affects a significant segment of the population. It is a heterogeneous disease that includes several causes of pain syndromes, latent molecular pathologies, genetic and psychological factors, and a history of injury. The Institute of Medicine has estimated that chronic pain affects approximately 100 million adults in the United States, with an estimated annual cost of up to US $635 billion [3].

Pain is perceived centrally and is strongly influenced by physical, physiological, and social or cultural factors. Another distinction often made is between pain caused by tissue damage (sometimes called inflammatory or nociceptive) and pain caused by nerve damage (neuropathic), where nerve signals may not be driven by local tissue damage.

Types of Pain

Depending on the quality, quantity, and duration, pain can be categorized into 2 major types: nociceptive and neuropathic pain. Distinguishing them is very important if proper treatment is to be achieved, because their causes and treatments are different. Figure 1 shows the classification of the major types of pain and contemporary theories of pain.

Figure 1. Classification of the major types of pain and contemporary pain theories.



Nociceptive Pain

Nociceptive pain can be attributed to tissue damage. Whole or undamaged neurons report damage, and pain is experienced [4]. It can be subdivided into somatic and visceral (gut) pain. This pain may be localized, constant, and often with an aching or pulsating quality. Nociceptive pain can be experienced as razor sharp, dull, or aching. Visceral pain is a subtype of nociceptive pain that involves the internal organs and tends to be episodic and poorly localized [5]. This type of pain is usually acute and is responsive to nonsteroidal anti-inflammatory drugs and opioids [6]. Examples of this type of pain include inflammation, burns, bruises, bone pain, and myofascial pain.

Neuropathic (Nerve) Pain

Neuropathic pain results from an injury or the malfunction of the peripheral or central nervous system [7,8]. This pain is often precipitated by an injury that may or may not involve actual damage to the nervous system [9]. Nerves can be permeated or compressed by tumors, suppressed by scar tissues, or inflamed by infections. This pain frequently involves burning, piercing, or electric shock qualities [9]. This type of pain may persist beyond the apparent healing of any damaged tissue. Neuropathic pain is often chronic and tends to have a less robust response to treatment with nonsteroidal anti-inflammatory drugs and opioids but may respond well to other drugs such as antiseizure and antidepressant medications [9]. Neuropathic problems tend to be irreversible, but partial improvement is often possible with proper treatment [7,8].

Examples include postherpetic neuralgia, nerve injury, cancer pain, phantom limb pain, entrapment neuropathy, and peripheral neuropathy (widespread nerve damage) [7,8]. Diabetes is among the many causes of peripheral neuropathy, but it can also be caused by chronic alcohol use, chemotherapy, vitamin deficiencies, and numerous other medical conditions, many of which may sometimes go undiagnosed [9].

Theories of Pain

Several theories of pain have been postulated for centuries to explain the mechanisms underlying pain perception [10-13]. The most modern theories include the specificity, intensity, pattern, and gate control theories of pain.

The specificity theory teaches that when specific nociceptive receptors in the periphery are stimulated, they transmit signals to the brain's pain center, which ultimately produces the perception of pain. This theory holds that the amount of pain is related to the amount of tissue damage. Assuming that the free nerve endings are the pain receptors, the theory has failed to find the pain receptors or the fibers specifically devoted to transmission, and it does not account for people who continue to experience pain long after the injury has healed.

The intensive (or summation) theory of pain asserts that pain is not a unique sensory experience but an emotion that occurs when a stimulus is stronger than usual. According to this theory, pain results from excessive stimulation of the sense of touch, with summation occurring in the dorsal horn cells. This explains why some form of summation must occur for subthreshold stimuli to become unbearable. The pattern theory of pain states that any somesthetic sensation occurs through a particular neural firing. This asserts that there are no specialized receptors. Pain occurs when the rate and pattern of sensory inputs exceed a threshold. The intensity evokes a pattern of impulses that are interpreted by the brain as pain.

The gate control theory claims that pain operates at the spinal level. It recognizes experimental evidence that supports the specificity and pattern theories. It carefully discusses the shortcomings of the specificity and pattern theories and attempts to bridge the gaps between the 2 dominant theories.

According to the pain gate control mechanism, Melzack and Wall (1965 [11]) accepted that there are nociceptors (pain fibers) and touch fibers. The pain gate mechanism was proposed as an alternative to the specificity theory of pain, which holds that pain is a specific modality with its own specialized sensors, neuronal pathways, and centers [14] and the pattern theory, which maintains that the stimulus intensity of nonspecific receptors and central summation are critical determinants of pain [15]. The pain gate control mechanism postulates that injury is transmitted from pain receptors to the central nervous system (CNS) via two types of nerve fibers: (1) small unmyelinated fibers (C-type) and (2) large myelin-containing fibers (A delta type), which transmit sharp, brief pain rapidly via the peripheral nerves through a gate mechanism. Larger-diameter nerve fibers pass through the same gate. If other subcutaneous stimuli are transmitted, the "gate" through which the pain impulse must travel is temporarily "blocked" by the other stimuli. The brain is unable to acknowledge pain impulses when transmitting other stimuli. When the gates are open, pain impulses flow freely.

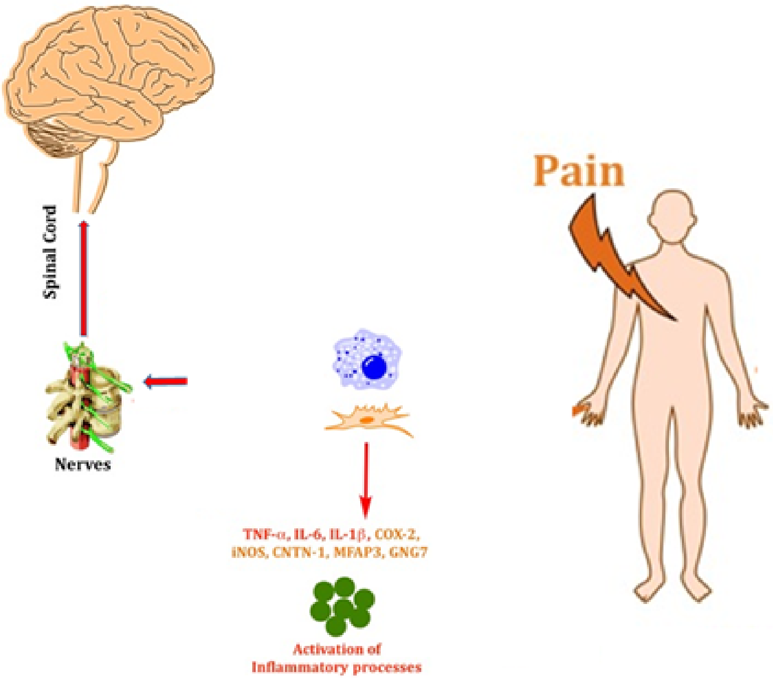
The theory, as originally propounded, states that the opening or closing of the "gate" depends on the relative activity of large-diameter (normal receptors) and small-diameter (pain receptors) fibers. It teaches that activity in large-diameter fibers tends to close the "gate," and activity in small-diameter fibers tends to open it [12]. Garrison and Foreman [16] support this theory, demonstrating that the cell activity of dorsal horn neurons decreases during transcutaneous electrical nerve stimulation (TENS) application to somatic receptive fields. Ultimately, this can potentially transmit noxious information to supraspinal levels. These findings support the "gate control theory of pain" in that less noxious information would be involved in the pain perception process. Garrison and Foreman [16] also showed that there is a differential effect in that more cells respond to conventional high-frequency, low-intensity TENS variables than they do to low-frequency, high-intensity ALTENS variables.

Biochemical Nature of Pain

Inflammation has been associated with pain. In a unifying theory of pain, the biochemical theory origin of pain asserts that regardless of the type of pain, whether acute pain or chronic pain as in arthritis, migraine, back or neck pain from herniated disks, complex regional pain syndrome or reflex sympathetic dystrophy pain, fibromyalgia, interstitial cystitis, neuropathic pain, or poststroke pain, the underlying basis is inflammation and the inflammatory response [17-19]. Therefore, irrespective of the characteristics of the pain, whether it is sharp, dull, aching, burning, stabbing, numbing, or tingling, it is asserted that all pain arises from inflammation and the inflammatory response. However, pain could be subjective, with pain reported without any tissue damage or underlying pathophysiological cause. Studies have shown that stress, anxiety, and other psychological factors may be responsible for the elevation in biomarkers.

According to the unifying theory of pain pioneered by Omoigui [17-19], the origin of all pain is inflammation and the inflammatory response. Biochemical mediators of inflammation include cytokines, neuropeptides, growth factors, and neurotransmitters. Irrespective of the type of pain, acute or chronic pain, peripheral or central pain, and nociceptive or neuropathic pain, the underlying origin is inflammation and the inflammatory response. The activation of pain receptors, transmission and modulation of pain signals, neuroplasticity, and central sensitization are all one continuum of inflammation and the inflammatory response. This theory proposes the reclassification and treatment of pain syndromes based on the inflammatory profile. Every pain syndrome has an inflammatory profile consisting of the inflammatory mediators present in the pain syndrome. The inflammatory profile may vary from one person to another and may vary in the same person at different times. The key to the treatment of pain syndromes is to understand their inflammatory profiles. The concentrations of several substances, namely substance P, calcitonin gene-related peptide, bradykinin, and various cytokines, are measurably elevated in the milieu of the active trigger point, indicating a chemical inflammatory response (Figure 2).

Figure 2. Biochemical signaling mechanism and the role of biomarkers in pain activation. COX: cyclooxygenase; GNG7: G-protein subunit gamma 7; IL: interleukin; iNOS: inducible nitric oxide synthase; MFAP: microfibril-associated protein; TNF: tumor necrosis factor.



Inflammatory pain is felt through multiple mediators released at inflammation sites that send information to the CNS. At the inflammation site, arachidonic acid is released by phospholipase A2 into the cell membrane. Cyclooxygenase-2 (COX-2) catalyzes the conversion of arachidonic acid into prostaglandin G_2 (PGG_2), and the peroxidase further reduces PGG_2 to prostaglandin H_2 (PGH_2), which is eventually converted into prostanoids, prostacyclin, and thromboxanes. These products bind to various receptors that signal pain in the CNS. Extensive literature supports the relationship between COX-2 and pain, with the amount of COX-2 being proportional to the magnitude of pain. In addition, many of the most widely used pain medications (eg, aspirin and nonsteroidal anti-inflammatory drugs) act through the COX-2 pathway, implying that, among other things, the detection of COX-2 could represent a direct measure of pain [20-28].

The fundamental origin of inflammatory pain is the activation of pain receptors, which leads to pain transmission (Figure 2). At the biochemical level, several factors are essential including neurotransmitters, cytokines, and growth factors. Despite the underlying biochemical nature of pain, few studies have focused on medical assessments to determine the nature of pain at the molecular level. It is important to remember that in most situations, acute pain serves as a critical protective mechanism in preventing further tissue injury. By reducing the risk of continued trauma, the tissue can heal more rapidly, and the pain will subside. The main challenge in dealing with intensive acute pain is to prevent the overuse of strong opioids, which can lead to addiction to the euphoria created by the use of these medications.

Chronic pain presents a very different challenge. The perception of pain, particularly chronic pain, is a process that uses partial, multimodal, and noisy information to create the perception of a potential bodily threat even long after the tissue has healed [29]. The issue of addressing chronic pain is challenging, because, following tissue healing, there is no peripheral stimulation—that is, there is no pain source, but there is a sensation of pain. Pain arises directly within the CNS, usually through central sensitization. Central sensitization results in 3 pain-related outcomes: hypersensitivity, pain in response to nonnoxious stimuli, and pain response outside the area of injury. These responses are mediated in the dorsal roots of the spine by several chemical agents, including substance P and prostaglandins such as cyclooxygenases.

On the basis of these pain theories, we propose an approach that acquires accurate measurements of biomarkers of the underlying pain processes. One approach is to develop analytical biosensors that can accurately measure pain markers. A biosensor is an analytical device that consists of a recognition element (eg, enzymes, antibodies, nucleic acids, cells, or micro-organisms) combined with a transducer or detector element that responds to the interaction of an analyte, allowing for an easy method of measuring and quantifying data. Owing to their fast response, simplicity, cost-effectiveness, and portability, biosensors can be used for continuous monitoring point-of-care analysis and do not require highly trained staff.

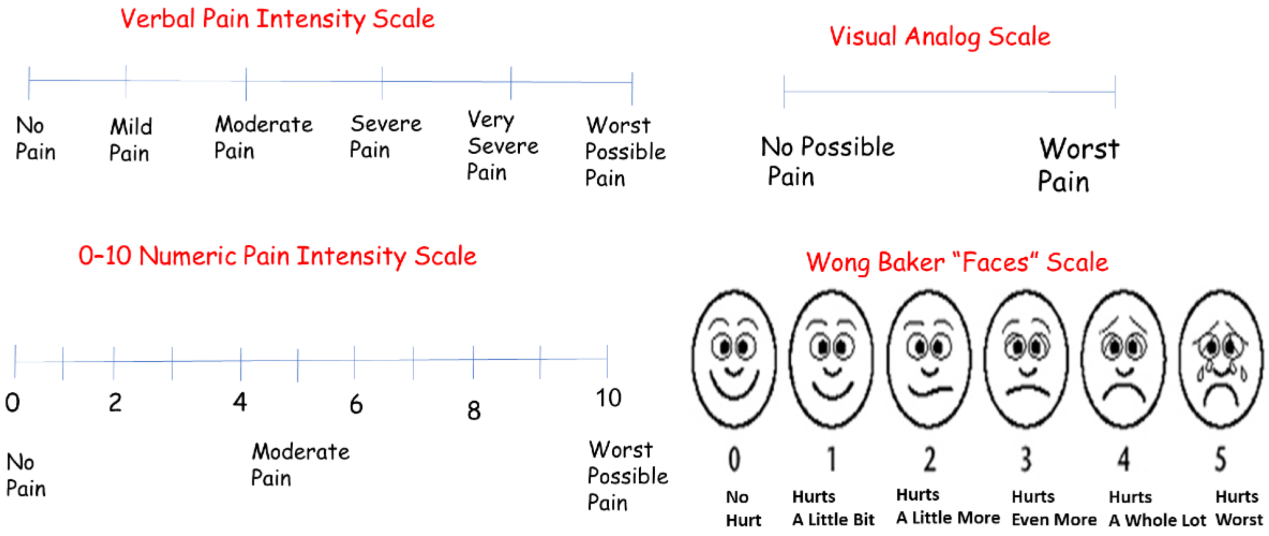
Conventional Methods of Pain Assessment

Currently, most pain measurements are based on patients' self-reports. For example, the visual analog scale (VAS) [30], McGill Pain Questionnaire [31], Wong-Baker Faces Scale [32], and Descriptor Differential Scale [33] have all been used as self-rating instruments for pain measurement in clinical and research settings. Figure 3 shows the different scales used to assess the levels of pain.

Despite the wide use of pain assessment tools, there is also an awareness of their inherent unreliability, as evidenced by reports on discrepancies [34-38]. Therefore, attempts have also been made to augment self-reports with other more objective measures, such as behavioral measures (eg, motor response, behavioral responses, facial expression, crying, sleep patterns, decreased activity or eating, body postures, and movements) [39-42]. Additional pain assessment methods include physiological measures such as changes in heart rate, blood pressure, oxygen saturation, palmar sweating, respiration, and sometimes neuroendocrine responses [43].

Conventional methods of assessing pain involve multisensory approaches or expensive devices such as brain imaging in a laboratory setting [44-47]. The feasibility and accuracy of this expensive advanced instrumentation in clinical settings remain challenging. A review article evaluated the current state of pain biomarkers developed using several commonly used methods, including structural magnetic resonance imaging, functional magnetic resonance imaging, and electroencephalography, with a model classification accuracy of 70% [46]. A study used functional MRI data during a visual stimulation task to distinguish between patients with fibromyalgia and healthy controls and recorded an accuracy of 82% [45]. Another study involving a combination of electrocardiograms to predict pain in healthy adults produced 75% to 81% accuracy [48]. A study has explicitly reported the use of microneedle-based biosensors for pain-free high-accuracy measurement of glycemia in interstitial fluid [49]. None had reported array self-reports using BN and pain-related biomarkers and biosensors.

Figure 3. Common pain assessment scales.



Rationale for Pain Sensor Development

The driving force behind the need to develop a pain sensor for the objective quantification of pain is that when different participants with the same disease or trauma report vastly different pain levels, it is tempting to assume that this reflects the differences in pain sensitivity. However, there are 2 reasons why this may not be true. First, although the diagnosis may be superficially the same, the severity of the disease or trauma may differ. Second, one might argue that the physical causes of pain may be initially similar across patients (eg, extraction of 2 wisdom teeth) but that these causes develop differently owing to differences in patients' pathophysiological conditions. Although better predictions of pain could be achieved through better characterization of pathology, there are reasons to doubt that differences in pathology are the only or even a major explanation for individual differences in pain.

Large differences in reported pain are ubiquitous and large when the cause of pain is homogenous and well defined (eg, surgery) as for illnesses with diffuse or unknown causes (eg, fibromyalgia). Inflammation and other physiological parameters are poorly correlated with pain intensity among patients with rheumatoid arthritis [50], and several studies have failed to find an association between the extent of breast surgery and acute postsurgical pain [51,52]. Most importantly, individual differences in reported pain are equally large for precisely controlled experimental pain stimuli [53].

In using either the VAS pain intensity ratings or the Wong-Baker Faces Scale [53,54], which is not only subjective but may only be qualitatively applied to patients with language or mental capacity difficulties, the variability in pain ratings of patients with the same disease or trauma is enormous. This occurs despite differences in individual pain sensitivity, and some clinical conditions experienced are more painful than others. Pain sensitivity can be estimated only through well-controlled experimental pain stimuli. Such estimates show substantial heritability but are equally critical environmental factors. The genetic and environmental factors that influence pain sensitivity differ across pain modalities. For example, genetic factors that influence cold-pressor pain have little impact on phasic heat pain and vice versa [53]. Individual differences in pain sensitivity can create complexities in diagnosis, because low sensitivity to pain may delay self-referral. The inclusion of patients with reduced pain sensitivity can attenuate treatment effects in clinical trials unless this is carefully controlled. Measures of pain sensitivity are predictive of acute postoperative pain, and there is preliminary evidence that heightened pain sensitivity increases the risk of future chronic pain conditions [54]. Experimental pain modalities have been suggested for use as predictors for future pain conditions, along with a careful assessment of each individual's pain sensitivity to prevent, evaluate, and treat pain. We propose a calibratable biomarker sensor and AI coupled with the patient's self-assessment of perceived pain. We surmise that such an approach may only be able to clearly distinguish between cases where the available evidence is consistent. However, this information may provide clinicians with valuable insight. Furthermore, as research progresses into how biomarkers are related to pain, more specific insights may emerge as to how specific evidence inconsistencies point to particular pain causes.

AI Methods for Summarizing Evidence

There is a long history of AI work that focuses on what is sometimes called evidential reasoning. These methods include BN [55], the Dempster-Shafer theory [56], and fuzzy logic and its derivatives [57,58]. For our present purposes, we will focus on Bayes nets and leave the possibility open that if our approach seems to stress the limitations of Bayes nets, other options may be available.

BN Proposal

Overview

The approach we propose is to devise a BN that will consider what evidence is available and report how the available evidence may be interpreted as a distribution of the likelihood that the participant is experiencing different levels of pain. This is similar to the proposal by Hill et al [59]. As a starting illustration, we consider only three forms of evidence: (1) a participant self-reports pain on a 0 to 5 scale (Figure 3), (2) a measured value of serum levels of COX-2 on a similar 0 to 5 scale, and (3) a measured value of inducible nitric oxide synthase (iNOS) on a 0 to 3 scale.

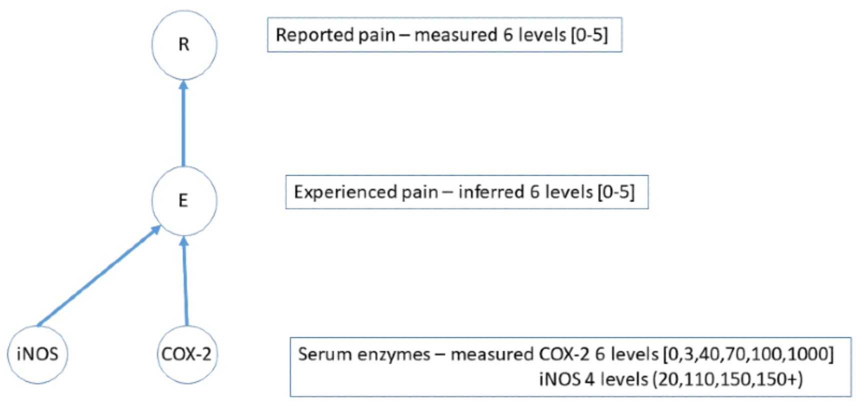
The process of developing a BN involves 2 basic steps:

1. Identify the key concepts needed in the domain, where each concept becomes a node in the network, and the causal relationships among them (known or assumed) are specified as directed links between the nodes. A node that has a "causal" influence on another node is called a parent node and the influenced node, the child. The nodes without links are assumed to be statistically independent. Figure 4 shows an example of a BN that can be applied to the pain domain. Each node is coded as a finite set of possible levels. For example, Figure 4 illustrates a node for "experienced pain" that may take any of the 6 levels: 0=no pain and 1, 2, 3, 4, and 5=most severe pain. This is the node we assume cannot be directly observed, and thus must be inferred from the other evidence.
2. Specify the set of parameters that will determine the probabilities to be computed. These consist of prior probabilities for each node and conditional probabilities of the child's probability, given knowledge of the parent's state. If a node's value is known, then the known level is assigned a probability of 1 and all other levels are assigned a probability of 0. If a node has no parent nodes and is unknown, then its prior probabilities are assigned to its levels. If a node is unknown but has one or more parent nodes, its posterior probabilities are computed using Bayes theorem (A is the child and B is the parent; in Figure 4, COX-2 is a parent and experienced pain is a child; equation 1):

where P(A) and P(B) are the prior probabilities, and P(A|B) is the conditional probability of A given the level of B. The usual approach assumes that all levels of a node's priors are the same (the principle of equal ignorance). However, in some applications, we may have the knowledge that the priors are not the same, and we can use this knowledge. If B is a set of parents rather than a singleton, a chain rule applies. Suppose it is the parent that is unknown (here, B is the child and A is the parent; in Figure 4, experienced pain is the parent and reported pain is the child). In this case, we compute the posterior probabilities for each level (*i*) of the parent node using the following (equation 2):

Figure 4. Illustrative Bayes network graphic for the pain application. COX: cyclooxygenase; E: experienced; iNOS: inducible nitric oxide synthase; R: reported.

Concept Sketch 2021-07-23

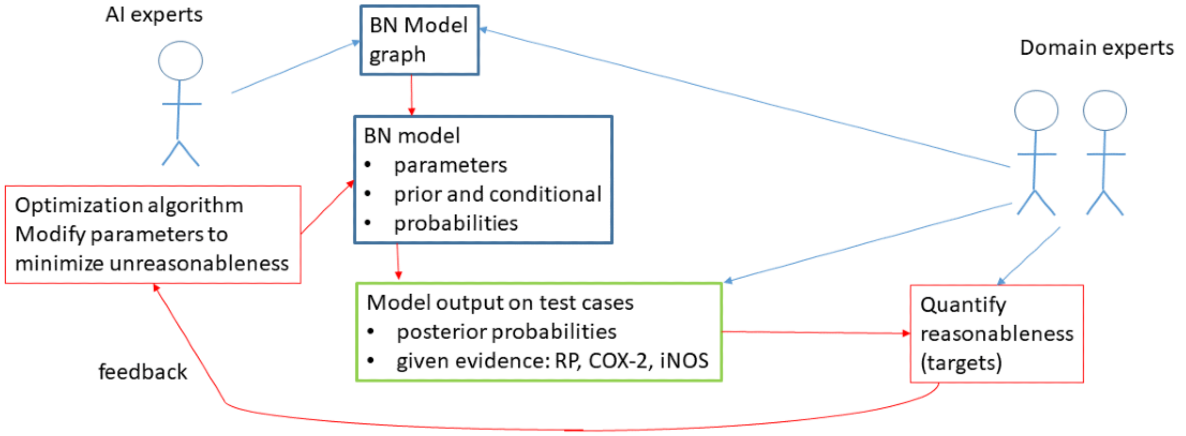

R
Reported pain – measured 6 levels [0-5]
E
Experienced pain – inferred 6 levels [0-5]
iNOS
COX-2
Serum enzymes – measured COX-2 6 levels [0,3,40,70,100,1000]
iNOS 4 levels (20,110,150,150+)

Approach to Setting BN Parameters When No Ground Truth Knowledge Exists

Each node in a specified graph requires a prior probability estimate. As mentioned earlier, one may always assume (by the principle of equal ignorance) that each level of a node has equal prior probability. If we have the knowledge or even reasonable consensus from experts that some probabilities other than all-equal are better, say, for a particular application or population of patients, we may set them accordingly. Such assumptions can always be tested using the methods described next.

The difficult challenge in devising a BN for an application is specifying the conditional probabilities for the parent-child links. One rarely has sufficient knowledge at this fine level of detail. However, what we may be able to accomplish with an adequate pool of domain experts is a specification of “reasonableness” tests for the final probabilities. That is, we ask the experts to say how the probabilities for the different levels of experienced pain (which we can never know with any certainty) “should” come out for some set of test cases. This set of test cases should span a “representative” range of possibilities of evidence combinations. With such information, it should be possible to set up an optimization procedure that can set the required conditional probability parameters to closely approximate the desired output behaviors of the system. This approach is illustrated in Figure 5.

Figure 5. An optimization approach for setting the conditional probability parameters for the Bayes network (BN). AI: artificial intelligence; COX: cyclooxygenase; iNOS: inducible nitric oxide synthase; RP: reported pain.



Some Illustrations of What Might be Accomplished

Our Patient Samples

Our earlier study explored the potential utility of serum COX-2 and iNOS as objective measures of pain in 102 American patients [60]. Sandwich enzyme-linked immunosorbent assay was used to determine COX-2 and iNOS levels in the blood serum. At the same time, statistical analysis was performed using Pearson product-moment correlation coefficients, regression, and receiver operating characteristics analyses. Our follow-up study examined the relationship between COX-2 and iNOS in the blood serum of >500 Turkish patients with different types of pain (Sadik, OA, unpublished data, November 2021) and assessed their potential as pain biomarkers. Serum COX-2 and iNOS levels were examined along with the level of pain caused by different types of pain, including lumbar or vertebral; lung; osteoporosis; inflammation; and fatigue, headache, or malaise related to problems. The data (Sadik, OA, unpublished data, November 2021) are now used to develop the current BN concept. For more details on our patient samples, see the Methods section.

Proposed Output From the BN

The BN will estimate the probabilities that the pain the participant is experiencing takes each of the possible levels: 0, 1, 2, 3, 4, and 5. Figure 6 shows the method for exhibiting these estimated probabilities as histograms. The left example shows the computed probabilities when the evidence is fully self-consistent, all pointing to no pain experienced. The right illustration shows computed probabilities when the evidence is not consistent; the participant reports experiencing pain at level 4 (out of 0-5), although the COX-2 evidence points to no pain (out of 0-5), and the iNOS evidence points to a pain level of 1 (out of 0-3). The probabilities over all levels must sum to one. The reader may notice in Figure 6 that we have modeled some "leakage" of probabilities to pain levels adjacent to the levels of the evidence. These behaviors are embodied in our choices of model parameters (conditional probabilities).

Because there is, as yet, no verifiable sensor for experienced pain, such a system would be capable of implementing some form of expert consensus as to how the available evidence should be combined. The best clinical decision support system can alert the caregiver to the extent that the available evidence is consistent. One possibility is to provide additional evidence that may be valuable in reconciling the situation.

Clearly, a point where important decisions are needed involves mapping actual measured biomarker values into the chosen discrete node levels for biomarkers, such as COX-2 and iNOS. Our current working model, the mapping, is presented in Tables 1 and 2. We hasten to point out that these decisions are provided for illustration purposes.

Evidently, the design of the system output can involve a small amount of creativity. Here, the primary source of guidance will be expert opinions from clinicians knowledgeable in this domain. There may be significant disagreements among pain experts, but it seems reasonable to assume that some considerable consensus may be reached regarding the nodes that should be included. Of course, it is always possible to explore different models for different applications. However, a significant challenge remains regarding the setting of many required internal probability parameters. For this, we propose the approach described in Section 3.1. For a somewhat differing approach, the reader may consult Hill et al [59].

Figure 6. Illustrations of envisioned Bayes network (BN)–clinical decision support system output. (A) consistent evidence (B) inconsistent evidence. COX: cyclooxygenase; iNOS: inducible nitric oxide synthase.

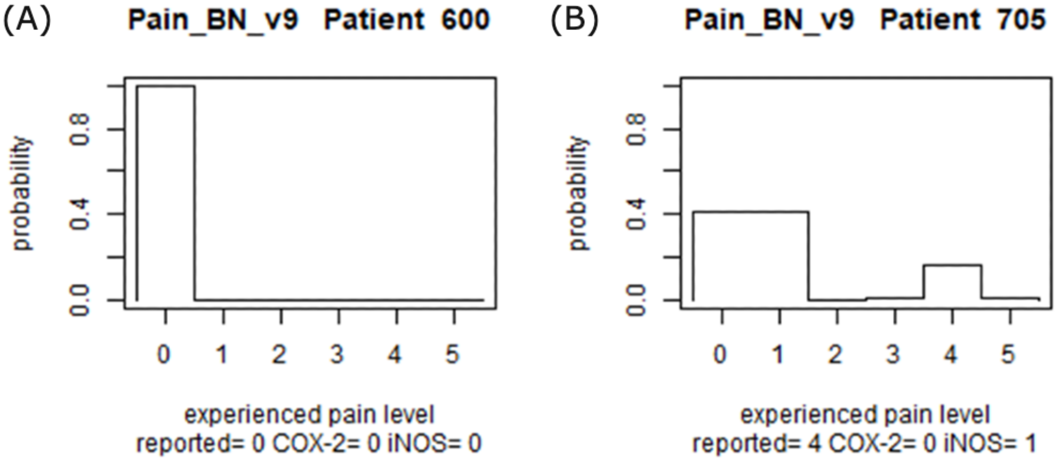


Table 1. Mapping measured values for cyclooxygenase-2 (COX-2) into experienced pain levels.

COX-2 measurement (ng/ml)	COX-2 code	Level of experienced pain most likely
<3	0	0
>3 to 40	1	1
>40 to 70	2	2
>70 to 100	3	3
>100 to 1000	4	4
>1000	5	5

Table 2. Mapping measured values for inducible nitric oxide synthase (iNOS) into experienced pain levels.

iNOS measurement (ng/ml)	iNOS code	Level of experienced pain most likely
<20	0	pain_0
>20 to 110	1	pain_1
>110 to 150	2	pain_2
>150	3	pain_3,4,5

Some Pain Evidence Examples

Figure 7 shows a distribution from our patient samples as to how consistent or not is the observed evidence that our BN model uses. We quantified the number of steps (node levels) that differed between the reported pain and biomarker levels. If a patient's difference was fewer than 3 steps for both biomarkers (an arbitrary threshold for illustration purposes), we called it "consistent" and colored it green. Of the 379 patients with all 3 pieces of evidence, 302 (79.7%) had consistent evidence. This suggests that our goal of using these biomarkers to corroborate reported pain may be reasonable. If one or both biomarkers was "inconsistent" (>2 steps different), we highlighted those patients in yellow. The 77 inconsistent patients fell into 5 groups:

1. A total of 8 patients with high COX2, low reported pain (RP), and low iNOS

2. A total of 31 patients with high RP, high COX2, and 0 iNOS

3. A total of 11 patients with high RP, low COX2, and low iNOS

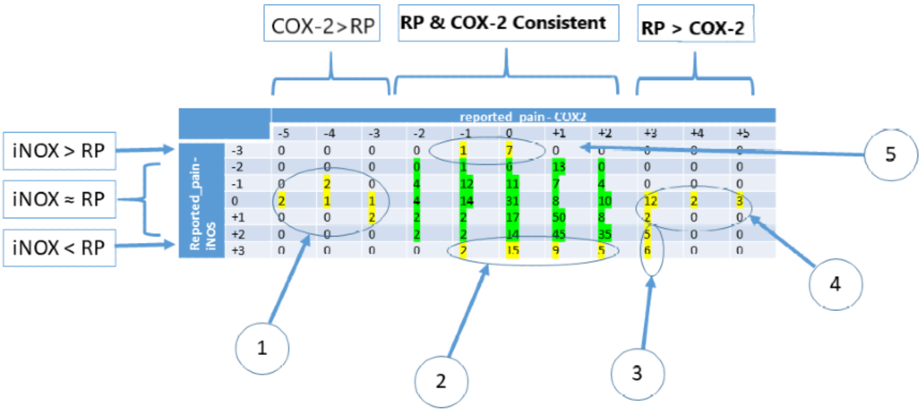
4. A total of 19 patients with high RP, high iNOS, and low COX2

5. A total of 8 patients with high iNOS, 0 RP, and low COX2

A reasonable next step would be to explore the data from these patients for other evidence that may help explain these observations.

The inconsistent evidence may be attributed to cultural or temporal variability of COX-2 and iNOS (as well as their serum variation and half-life), VAS, and other tools. Tables 1 and 2 assume that each categorized COX-2 and iNOS measurement most likely corresponds to a given pain experience. However, it is necessary to perform "sensitivity analysis" for greater precision. Inconsistencies may also be caused by potential miscalibration in pain management [35-37,61]. Ongoing work now includes supplementary information regarding the time-of-day data the biomarkers were taken and measured. The time-course measurements are being compared with other collected data. Information regarding gender and individual differences in COX-2 and iNOS data on RP is also being recorded. Different causes of inconsistencies could be attributed to memory bias and cognitive effects such as exaggerations or underestimations when reporting pain.

Figure 7. The distribution of evidence consistency or inconsistency in our samples. COX: cyclooxygenase; iNOS: inducible nitric oxide synthase; RP: reported pain.



Reported_pain - iNOS	reported_pain - COX2										
	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5
-3	0	0	0	0	1	7	0	0	0	0	0
-2	0	0	0	0	1	6	13	0	0	0	0
-1	0	2	0	4	12	11	7	4	0	0	0
0	2	1	1	4	14	31	8	10	12	2	3
+1	0	0	2	2	2	17	50	8	2	0	0
+2	0	0	0	2	2	14	45	35	5	0	0
+3	0	0	0	0	7	15	9	5	6	0	0

Summary

We have presented an approach to building a clinical decision support system to help clinicians assess the pain experienced by a patient. We base our approach on ongoing research into new biosensor technologies that we hope will soon make available quick, inexpensive, and minimally intrusive measures of biomarkers related to pain. Such evidence will then need AI technology to offer clinicians an easy-to-grasp summary of the available evidence and perhaps suggestions for useful next steps.

We present a simple Bayes net model as a prototype. Preliminary data on 379 patients suggest that this approach is appropriate, because the majority (302/379, 79.7%) of participants showed reasonable consistency between the biomarker data and the patients' self-reported pain. These data also showed 5 distinct types of inconsistencies, suggesting follow-up exploration of factors that might account for these inconsistencies. However, much work remains to be done. First, a community of clinical pain experts must be assembled to help define how such a prototype might be further developed (perhaps using alternative AI methods) to be of practical value. Portable biosensors need to be developed to allow for an easy method for measuring and quantifying data. Owing to their fast responses, simplicity, cost-effectiveness, and portability, biosensors can be used for continuous monitoring of point-of-care analysis and do not require highly trained staff to operate.

Methods

Biomarkers or Biosensors

In 2018, the National Institutes of Health launched the Helping to End Addiction Long-term Initiative to stem the national opioid public health crisis [60,62]. One component of the Helping to End Addiction Long-term Initiative is to support biomarker discovery and rigorous validation to accelerate high-quality clinical research on pain [62,63]. Biomarkers are objectively measured and evaluated as indicators of either normal or pathogenic biological processes or responses to therapeutic interventions. Multiple studies have observed significant differences in proinflammatory cytokines (eg, interleukin 6 [IL‐6], tumor necrosis factor α, IL‐8, and IL‐1β) in relation to pain intensity [63-65]. Serum protein levels and mRNA expression of tumor necrosis factor α have been shown to be significantly higher in participants experiencing a greater intensity of chronic pain. Biomarkers may be used alone or in combination to assess the health or disease state of an individual [66,67].

Our group conducted extensive research in this area, identifying COX-2 and iNOS or nitric oxide synthase 2 as good candidates for this purpose [58,60,62,63,65,68-71]. COX, also known as prostaglandin H_2 synthase, is a key bifunctional enzyme in the biosynthetic pathway that leads to the formation of prostanoids, including prostaglandins, prostacyclins, and thromboxanes. COX exists in different isoforms [72,73], COX-1, COX-2, and COX-3. COX-1 is an oxidoreductase enzyme constitutively expressed in many cell types. It is presumed to be responsible for the synthesis of housekeeping prostanoids that are critical for normal physiological functions such as regulating vascular homeostasis, gastric mucosa protection, and renal integrity [74]. COX-3 is a variant of COX-1, which has retained intron-1 during translation and is found in human tissues in a polyadenylated form [75]. It is a selective splicing product of COX-1 mRNA with 633 amino acids with less activity in the production of prostaglandin E_2, and it is mainly found in the hypothalamus, spinal cord, and pituitary choroid plexus. COX-2, on the other hand, is usually undetectable in healthy tissues but is rapidly induced and found to be upregulated in a variety of pathophysiological conditions such as neurological diseases [24], pain [76], inflammation, and cancer infection [13,77,78]. Some studies have indicated that the level of COX-2 at the point of inflammation translates to the degree of inflammation and may thus be used to determine the level of inflammatory pain [18]. Nitric oxide (NO) is a highly reactive free radical and, at the same time, an important signaling molecule involved in different functions [79]. Its inducible form, “iNOS,” is expressed in macrophages and other tissues in response to infection or inflammation, generating large amounts of NO in the blood [24,80]. Increased NO levels have been observed during inflammation and arthritis; therefore, iNOS can be considered a pain biomarker.

In addition to these biomarkers, preliminary results from our laboratory indicated that Contactin-1 (CNTN-1) could also be a promising pain biomarker. This is supported by previous studies that pointed to CNTN-1 as a pain suppressor [81,82] and found antibodies against CNTN-1 in patients with chronic inflammatory demyelinating polyradiculoneuropathy [26]. CNTN-1 levels have been shown to decrease in blood in high-pain states, with convergent evidence in other tissues in human studies for the involvement of pain. Anti–CNTN-1 autoantibodies block or decrease the levels of CNTN-1 in chronic inflammatory demyelinating polyneuropathy [22] and have been considered a bona fide pain marker [81-84]. Moreover, human G-protein subunit gamma 7 plays a strong role in signal transduction with decreased levels in high-pain states (ie, it is a pain suppressor gene with transdiagnostic evidence for involvement in psychiatric disorders) [85,86]. Its expression is decreased by omega-3 fatty acids [87,88]. Microfibril-associated protein 3 provides the most robust empirical evidence as a strong predictor of pain in both men and women. It decreases in expression in the blood during high-pain states [28,84,89,90].

Patient Recruitment

The General Secretary approved the institutional review board of the Manisa State Hospitals Union. The study was conducted at Manisa Merkez Efendi State Hospital, Manisa, Turkey. The participants included in the study were recruited from emergency, internal medicine, gynecology, general surgery, clinical microbiology, chest, urology, and physical therapy clinics. Only participants aged ≥18 years, who consented to participate were included in the study. All participant recruitment and data collection were performed by nurses at the clinics.

Patients were excluded from data analysis if they (1) aged <18 years, (2) did not provide sufficient description during anamnesis to determine their level of pain, and (3) had a blood sample not sufficiently large for analysis (Figure 8).

The survey questions were incorporated into the initial intake and anamnesis questions provided by the nurses. The survey questions are presented in Multimedia Appendix 1. The survey questions included information such as participant’s age, gender, smoking and alcohol habits (Figure 9), chronic disease, long-term medication, surgery history, the reason for and duration of the pain, and pain medication before coming to the hospital.

Figure 8. Flowchart of patient recruitment. COX: cyclooxygenase; iNOS: inducible nitric oxide synthase.

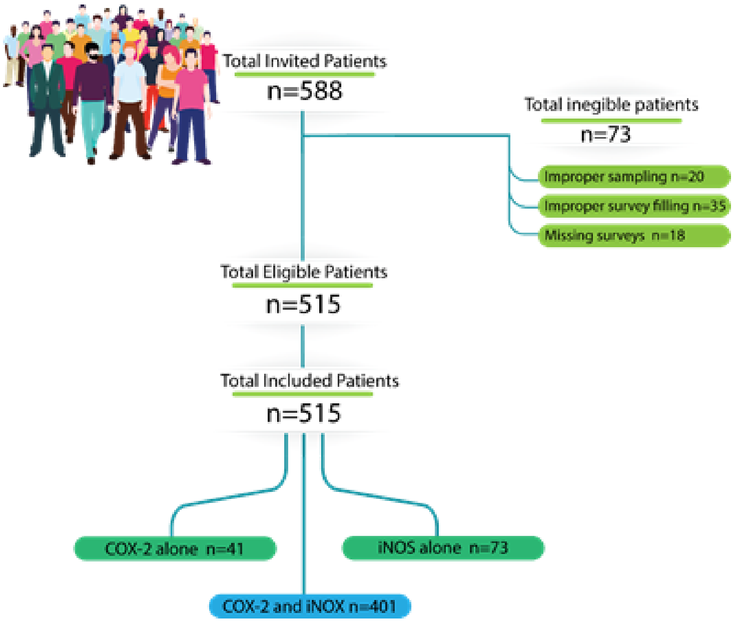
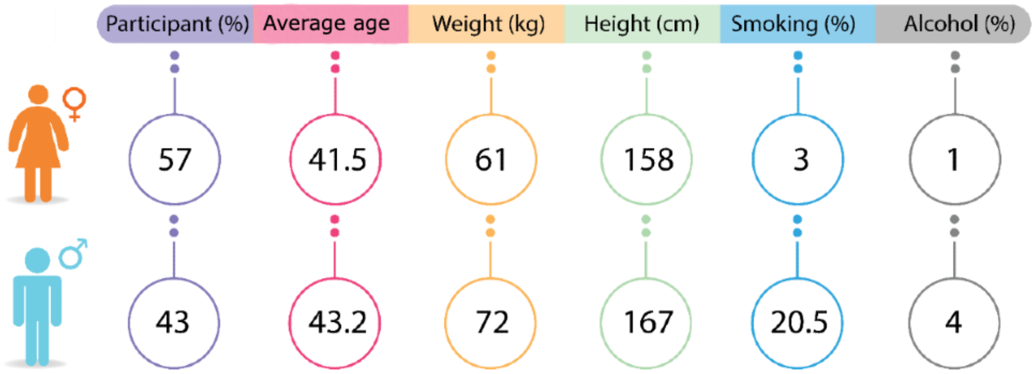


Figure 9. Demographic and habit distribution of patients.



Age and Height Are Averages

Following this initial questioning, patients were informed of the study. Informed consent was obtained if the patient agreed to participate in this study. Blood samples were collected via the leftover serum from blood samples taken for routine analysis. During the informed consent process, the participants consented to the use of their leftover serum; no patients were asked to donate blood samples specifically for the study.

The pain level for each participant was classified by the nurse performing anamnesis based on the patient's responses to the survey questions. Pain level was classified from 0 to 5 as 0=no pain, 1=feeling pain but not disturbing, 2=feeling pain and little disturbing, 3=severe pain and requiring painkiller intake, 4=very severe pain and distraction from working and requiring urgent painkiller administration, and 5=unbearable pain requiring urgent painkiller administration and rest as well as causing anxiety. Each pain level with characteristic conditions was explained to the patients, who were asked how they felt and if they had taken painkillers before they arrived at the hospital.

A sandwich enzyme-linked immunosorbent assay was used to monitor the levels of COX-2 and iNOS in the serum, as reported elsewhere [68].

Acknowledgments

This paper is based on the work supported by the Research Foundation of the State University of New York (SUNY) through the SUNY Health Network of Excellence Program under grant 71702. IY thanks the Turkish Ministry of National Education for their travel support. The authors also acknowledge Özlem Saf, who expended tremendous effort to collect the surveys; Professor Ali Kılınç of the Biochemistry Department at Ege University for opening his laboratory to the study; General Secretary of Manisa State Hospitals for their help in conducting the study; and all the physicians, nurses, and patients who contributed to the study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Survey form.
[PNG File , 71 KB - biomedeng_v7i2e35711_app1.png]

References

1. International Association for the Study of Pain. URL: https://www.iasp-pain.org/publications/iasp-news/iasp-announces-revised-definition-of-pain/ [accessed 2022-07-26]
2. One in Five American Adults Experience Chronic Pain, Limiting Daily Functioning and Productivity. SciTechDaily. 2021 Apr 20. URL: https://scitechdaily.com/one-in-five-american-adults-experience-chronic-pain-limiting-daily-functioning-and-productivity/ [accessed 202-07-26]
3. American Pain Society. Chronic pain costs U.S. up to $635 billion, study shows. ScienceDaily. 2012 Sep 11. URL: https://www.sciencedaily.com/releases/2012/09/120911091100.htm [accessed 2022-07-26]
4. Main CJ, Spanswick CC. Pain Management: An Interdisciplinary Approach. London, UK: Churchill Livingstone; 2000.
5. Cervero F, Laird JM. Visceral pain. Lancet 1999 Jun 19;353(9170):2145-2148. [doi: 10.1016/S0140-6736(99)01306-9] [Medline: 10382712]
6. Noah NM, Alam S, Sadik OA. Detection of inducible nitric oxide synthase using a suite of electrochemical, fluorescence, and surface plasmon resonance biosensors. Anal Biochem 2011 Jun 15;413(2):157-163. [doi: 10.1016/j.ab.2011.02.010] [Medline: 21316333]
7. Ji RR, Woolf CJ. Neuronal plasticity and signal transduction in nociceptive neurons: implications for the initiation and maintenance of pathological pain. Neurobiol Dis 2001 Feb;8(1):1-10. [doi: 10.1006/nbdi.2000.0360] [Medline: 11162235]
8. Martin P, Leibovich SJ. Inflammatory cells during wound repair: the good, the bad and the ugly. Trends Cell Biol 2005 Nov;15(11):599-607. [doi: 10.1016/j.tcb.2005.09.002] [Medline: 16202600]
9. Woolf CJ, Mannion RJ. Neuropathic pain: aetiology, symptoms, mechanisms, and management. Lancet 1999 Jun 05;353(9168):1959-1964. [doi: 10.1016/S0140-6736(99)01307-0] [Medline: 10371588]
10. Moayedi M, Davis KD. Theories of pain: from specificity to gate control. J Neurophysiol 2013 Jan;109(1):5-12 [FREE Full text] [doi: 10.1152/jn.00457.2012] [Medline: 23034364]
11. Melzack R, Wall PD. Pain mechanisms: a new theory. Science 1965 Nov 19;150(3699):971-979. [doi: 10.1126/science.150.3699.971] [Medline: 5320816]
12. Shimoji K, Yokota Y. Theories of pain. In: Shimoji K, Nader A, Hamann W, editors. Chronic Pain Management in General and Hospital Practice. Singapore, Singapore: Springer; 2021:11-19.
13. Bonica JJ. The Management of Pain. 2 edition. Palo Alto, CA, USA: Lea & Febiger; 1990.
14. Kucharski A, Todd EM. Pain: historical perspectives. In: Wootton RJ, Warfield CA, Bajwa ZH, editors. Principles & Practice of Pain Medicine. 2nd edition. New York, NY, USA: McGraw Hill; 2004.
15. Baldry P. Acupuncture, Trigger Points, and Musculoskeletal Pain: A Scientific Approach to Acupuncture for Use by Doctors and Physiotherapists in the Diagnosis and Management of Myofascial Trigger Point Pain. London, UK: Churchill Livingstone; 1993.
16. Garrison DW, Foreman RD. Decreased activity of spontaneous and noxiously evoked dorsal horn cells during transcutaneous electrical nerve stimulation (TENS). Pain 1994 Sep;58(3):309-315. [doi: 10.1016/0304-3959(94)90124-4] [Medline: 7838579]
17. Omoigui S. The Interleukin-6 inflammation pathway from cholesterol to aging--role of statins, bisphosphonates and plant polyphenols in aging and age-related diseases. Immun Ageing 2007 Mar 20;4:1 [FREE Full text] [doi: 10.1186/1742-4933-4-1] [Medline: 17374166]
18. Omoigui S. The biochemical origin of pain--proposing a new law of pain: the origin of all pain is inflammation and the inflammatory response. Part 1 of 3--a unifying law of pain. Med Hypotheses 2007;69(1):70-82 [FREE Full text] [doi: 10.1016/j.mehy.2006.11.028] [Medline: 17240081]
19. Omoigui S. The biochemical origin of pain: the origin of all pain is inflammation and the inflammatory response. Part 2 of 3 - inflammatory profile of pain syndromes. Med Hypotheses 2007;69(6):1169-1178 [FREE Full text] [doi: 10.1016/j.mehy.2007.06.033] [Medline: 17728071]
20. Ahn SH, Cho YW, Ahn MW, Jang SH, Sohn YK, Kim HS. mRNA expression of cytokines and chemokines in herniated lumbar intervertebral discs. Spine (Phila Pa 1976) 2002 May 01;27(9):911-917. [doi: 10.1097/00007632-200205010-00005] [Medline: 11979160]
21. Cerella C, Sobolewski C, Dicato M, Diederich M. Targeting COX-2 expression by natural compounds: a promising alternative strategy to synthetic COX-2 inhibitors for cancer chemoprevention and therapy. Biochem Pharmacol 2010 Dec 15;80(12):1801-1815. [doi: 10.1016/j.bcp.2010.06.050] [Medline: 20615394]

22. Fletcher BS, Lim RW, Varnum BC, Kujubu DA, Koski RA, Herschman HR. Structure and expression of TIS21, a primary response gene induced by growth factors and tumor promoters. J Biol Chem 1991 Aug 05;266(22):14511-14518 [FREE Full text] [Medline: 1713584]
23. Franson RC, Saal JS, Saal JA. Human disc phospholipase A2 is inflammatory. Spine (Phila Pa 1976) 1992 Jun;17(6 Suppl):S129-S132. [doi: 10.1097/00007632-199206001-00011] [Medline: 1631712]
24. Pasinetti GM, Aisen PS. Cyclooxygenase-2 expression is increased in frontal cortex of Alzheimer's disease brain. Neuroscience 1998 Nov;87(2):319-324. [doi: 10.1016/s0306-4522(98)00218-8] [Medline: 9740394]
25. Marnett LJ, Rowlinson SW, Goodwin DC, Kalgutkar AS, Lanzo CA. Arachidonic acid oxygenation by COX-1 and COX-2. Mechanisms of catalysis and inhibition. J Biol Chem 1999 Aug 13;274(33):22903-22906 [FREE Full text] [doi: 10.1074/jbc.274.33.22903] [Medline: 10438452]
26. Silva PJ, Fernandes PA, Ramos MJ. A theoretical study of radical-only and combined radical/carbocationic mechanisms of arachidonic acid cyclooxygenation by prostaglandin H synthase. Theor Chem Acc 2003 Dec 1;110(5):345-351. [doi: 10.1007/s00214-003-0476-9]
27. Sinicrope FA, Lemoine M, Xi L, Lynch PM, Cleary KR, Shen Y, et al. Reduced expression of cyclooxygenase 2 proteins in hereditary nonpolyposis colorectal cancers relative to sporadic cancers. Gastroenterology 1999 Aug;117(2):350-358. [doi: 10.1053/gast.1999.0029900350] [Medline: 10419916]
28. Niculescu AB, Levey DF, Phalen PL, Le-Niculescu H, Dainton HD, Jain N, et al. Understanding and predicting suicidality using a combined genomic and clinical risk assessment approach. Mol Psychiatry 2015 Nov;20(11):1266-1285 [FREE Full text] [doi: 10.1038/mp.2015.112] [Medline: 26283638]
29. Tabor A, Thacker MA, Moseley GL, Körding KP. Pain: a statistical account. PLoS Comput Biol 2017 Jan 12;13(1):e1005142 [FREE Full text] [doi: 10.1371/journal.pcbi.1005142] [Medline: 28081134]
30. Myles PS, Troedel S, Boquest M, Reeves M. The pain visual analog scale: is it linear or nonlinear? Anesth Analg 1999 Dec;89(6):1517-1520. [doi: 10.1097/00000539-199912000-00038] [Medline: 10589640]
31. Hawker GA, Mian S, Kendzerska T, French M. Measures of adult pain: Visual Analog Scale for Pain (VAS Pain), Numeric Rating Scale for Pain (NRS Pain), McGill Pain Questionnaire (MPQ), Short-Form McGill Pain Questionnaire (SF-MPQ), Chronic Pain Grade Scale (CPGS), Short Form-36 Bodily Pain Scale (SF-36 BPS), and Measure of Intermittent and Constant Osteoarthritis Pain (ICOAP). Arthritis Care Res (Hoboken) 2011 Nov;63 Suppl 11:S240-S252 [FREE Full text] [doi: 10.1002/acr.20543] [Medline: 22588748]
32. Lawson SL, Hogg MM, Moore CG, Anderson WE, Osipoff PS, Runyon MS, et al. Pediatric pain assessment in the emergency department: patient and caregiver agreement using the Wong-Baker FACES and the faces pain scale-revised. Pediatr Emerg Care 2021 Dec 01;37(12):e950-e954. [doi: 10.1097/PEC.0000000000001837] [Medline: 31335787]
33. Doctor JN, Slater MA, Atkinson HJ. The Descriptor Differential Scale of Pain Intensity: an evaluation of item and scale properties. Pain 1995 May;61(2):251-260. [doi: 10.1016/0304-3959(94)00180-M] [Medline: 7659435]
34. Cushman D, McCormick Z, Casey E, Plastaras CT. Discrepancies in describing pain: is there agreement between numeric rating scale scores and pain reduction percentage reported by patients with musculoskeletal pain after corticosteroid injection? Pain Med 2015 May;16(5):870-876. [doi: 10.1111/pme.12669] [Medline: 25715989]
35. Linton SJ, Götestam GK. A clinical comparison of two pain scales: correlation, remembering chronic pain, and a measure of compliance. Pain 1983 Sep;17(1):57-65. [doi: 10.1016/0304-3959(83)90127-6] [Medline: 6226918]
36. Marquié L, Raufaste E, Lauque D, Mariné C, Ecoiffier M, Sorum P. Pain rating by patients and physicians: evidence of systematic pain miscalibration. Pain 2003 Apr;102(3):289-296. [doi: 10.1016/S0304-3959(02)00402-5] [Medline: 12670671]
37. Savino F, Vagliano L, Ceratto S, Viviani F, Miniero R, Ricceri F. Pain assessment in children undergoing venipuncture: the Wong-Baker faces scale versus skin conductance fluctuations. PeerJ 2013 Feb 12;1:e37 [FREE Full text] [doi: 10.7717/peerj.37] [Medline: 23638373]
38. Varni JW, Thissen D, Stucky BD, Liu Y, Magnus B, He J, et al. Item-level informant discrepancies between children and their parents on the PROMIS(®) pediatric scales. Qual Life Res 2015 Aug;24(8):1921-1937 [FREE Full text] [doi: 10.1007/s11136-014-0914-2] [Medline: 25560776]
39. Fordyce WE. Behavioral Methods for Chronic Pain and Illness. St. Louis, MO, USA: Mosby; 1978.
40. Keefe FJ, Dunsmore J. Pain behavior: concepts and controversies. Am Pain Soc 1992 Jun 1;1(2):92-100.
41. Prkachin KM, Solomon PE. The structure, reliability and validity of pain expression: evidence from patients with shoulder pain. Pain 2008 Oct 15;139(2):267-274. [doi: 10.1016/j.pain.2008.04.010] [Medline: 18502049]
42. Prkachin KM. Assessing pain by facial expression: facial expression as nexus. Pain Res Manag 2009;14(1):53-58 [FREE Full text] [doi: 10.1155/2009/542964] [Medline: 19262917]
43. Korving H, Sterkenburg PS, Barakova EI, Feijs LM. Physiological measures of acute and chronic pain within different subject groups: a systematic review. Pain Res Manag 2020 Sep 3;2020:9249465 [FREE Full text] [doi: 10.1155/2020/9249465] [Medline: 32952747]
44. Boissoneault J, Sevel L, Letzen J, Robinson M, Staud R. Biomarkers for musculoskeletal pain conditions: use of brain imaging and machine learning. Curr Rheumatol Rep 2017 Jan;19(1):5 [FREE Full text] [doi: 10.1007/s11926-017-0629-9] [Medline: 28144827]

45. Harte SE, Ichesco E, Hampson JP, Peltier SJ, Schmidt-Wilcke T, Clauw DJ, et al. Pharmacologic attenuation of cross-modal sensory augmentation within the chronic pain insula. Pain 2016 Sep;157(9):1933-1945 [FREE Full text] [doi: 10.1097/j.pain.0000000000000593] [Medline: 27101425]
46. van der Miesen MM, Lindquist MA, Wager TD. Neuroimaging-based biomarkers for pain: state of the field and current directions. Pain Rep 2019 Aug 7;4(4):e751 [FREE Full text] [doi: 10.1097/PR9.0000000000000751] [Medline: 31579847]
47. Martucci KT, Ng P, Mackey S. Neuroimaging chronic pain: what have we learned and where are we going? Future Neurol 2014 Nov;9(6):615-626 [FREE Full text] [doi: 10.2217/FNL.14.57] [Medline: 28163658]
48. Kasaeyan Naeini E, Subramanian A, Calderon MD, Zheng K, Dutt N, Liljeberg P, et al. Pain recognition with electrocardiographic features in postoperative patients: method validation study. J Med Internet Res 2021 May 28;23(5):e25079 [FREE Full text] [doi: 10.2196/25079] [Medline: 34047710]
49. Strambini LM, Longo A, Scarano S, Prescimone T, Palchetti I, Minunni M, et al. Self-powered microneedle-based biosensors for pain-free high-accuracy measurement of glycaemia in interstitial fluid. Biosens Bioelectron 2015 Apr 15;66:162-168. [doi: 10.1016/j.bios.2014.11.010] [Medline: 25601169]
50. Montgomery GH, Bovbjerg DH. Presurgery distress and specific response expectancies predict postsurgery outcomes in surgery patients confronting breast cancer. Health Psychol 2004 Jul;23(4):381-387. [doi: 10.1037/0278-6133.23.4.381] [Medline: 15264974]
51. Chen AC, Dworkin SF, Haug J, Gehrig J. Human pain responsivity in a tonic pain model: psychological determinants. Pain 1989 May;37(2):143-160. [doi: 10.1016/0304-3959(89)90126-7] [Medline: 2664663]
52. Nielsen CS, Stubhaug A, Price DD, Vassend O, Czajkowski N, Harris JR. Individual differences in pain sensitivity: genetic and environmental contributions. Pain 2008 May;136(1-2):21-29. [doi: 10.1016/j.pain.2007.06.008] [Medline: 17692462]
53. Nielsen CS, Price DD, Vassend O, Stubhaug A, Harris JR. Characterizing individual differences in heat-pain sensitivity. Pain 2005 Dec 15;119(1-3):65-74. [doi: 10.1016/j.pain.2005.09.018] [Medline: 16298065]
54. Scott RW, Datye AK, Crooks RM. Bimetallic palladium-platinum dendrimer-encapsulated catalysts. J Am Chem Soc 2003 Apr 02;125(13):3708-3709. [doi: 10.1021/ja034176n] [Medline: 12656595]
55. Pearl J. On evidential reasoning in a hierarchy of hypotheses. Artif Intell 1986 Feb;28(1):9-15. [doi: 10.1016/0004-3702(86)90027-5]
56. Shafer G. A Mathematical Theory of Evidence. Princeton, NJ, USA: Princeton University Press; 1976.
57. Zadeh L. Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets Syst 1978 Jan;1(1):3-28 [FREE Full text] [doi: 10.1016/0165-0114(78)90029-5]
58. Zadeh LA. Fuzzy sets and information granularity. Adv Fuzzy Set Theory Appl 1979;11:3-18. [doi: 10.1142/9789814261302_0022]
59. Hill A, Joyner CH, Keith-Jopp C, Yet B, Tuncer Sakar C, Marsh W, et al. A Bayesian network decision support tool for low back pain using a RAND appropriateness procedure: proposal and internal pilot study. JMIR Res Protoc 2021 Jan 15;10(1):e21804 [FREE Full text] [doi: 10.2196/21804] [Medline: 33448937]
60. The Helping to End Addiction Long-term® Initiative. National Institutes of Health, Heal Initiative. URL: https://heal.nih.gov/ [accessed 2021-12-08]
61. Giusti GD, Reitano B, Gili A. Pain assessment in the Emergency Department. Correlation between pain rated by the patient and by the nurse. An observational study. Acta Biomed 2018 Feb 27;89(4-S):64-70 [FREE Full text] [doi: 10.23750/abm.v89i4-S.7055] [Medline: 29644991]
62. NINDS Program Description. National Institute of Neurological Disorders and Stroke. URL: https://www.ninds.nih.gov/Current-Research/Focus-Tools-Topics/Biomarkers [accessed 2021-12-08]
63. Roughley PJ, Alini M, Antoniou J. The role of proteoglycans in aging, degeneration and repair of the intervertebral disc. Biochem Soc Trans 2002 Nov;30(Pt 6):869-874. [doi: 10.1042/bst0300869] [Medline: 12440935]
64. Alvarez JA, Hardy Jr RH. Lumbar spine stenosis: a common cause of back and leg pain. Am Fam Physician 1998 Apr 15;57(8):1825-1840 [FREE Full text] [Medline: 9575322]
65. Raj PP. Intervertebral disc: anatomy-physiology-pathophysiology-treatment. Pain Pract 2008;8(1):18-44. [doi: 10.1111/j.1533-2500.2007.00171.x] [Medline: 18211591]
66. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther 2001 Mar;69(3):89-95. [doi: 10.1067/mcp.2001.113989] [Medline: 11240971]
67. Davis KD, Aghaeepour N, Ahn AH, Angst MS, Borsook D, Brenton A, et al. Discovery and validation of biomarkers to aid the development of safe and effective pain therapeutics: challenges and opportunities. Nat Rev Neurol 2020 Jul;16(7):381-400 [FREE Full text] [doi: 10.1038/s41582-020-0362-2] [Medline: 32541893]
68. Sadik OA, Yazgan I, Eroglu O, Liu P, Olsen ST, Moser AM, et al. Objective clinical pain analysis using serum cyclooxygenase-2 and inducible nitric oxide synthase in American patients. Clin Chim Acta 2018 Sep;484:278-283. [doi: 10.1016/j.cca.2018.06.005] [Medline: 29885320]
69. Omole MA, Noah N, Zhou L, Almaletti A, Sadik OA, Asemota HN, et al. Spectroelectrochemical characterization of pain biomarkers. Anal Biochem 2009 Dec 01;395(1):54-60. [doi: 10.1016/j.ab.2009.07.038] [Medline: 19646944]
70. Noah NM, Mwilu SK, Sadik OA, Fatah AA, Arcilesi RD. Immunosensors for quantifying cyclooxygenase 2 pain biomarkers. Clin Chim Acta 2011 Jul 15;412(15-16):1391-1398. [doi: 10.1016/j.cca.2011.04.017] [Medline: 21530501]

71. Noah NM, Marcells O, Almalleti A, Lim J, Sadik OA. Metal enhanced electrochemical cyclooxygenase-2 (COX-2) sensor for biological applications. Electroanalysis 2011 Sep 06;23(10):2392-2399. [doi: 10.1002/elan.201100241]
72. Somvanshi RK, Kumar A, Kant S, Gupta D, Singh SB, Das U, et al. Surface plasmon resonance studies and biochemical evaluation of a potent peptide inhibitor against cyclooxygenase-2 as an anti-inflammatory agent. Biochem Biophys Res Commun 2007 Sep 14;361(1):37-42. [doi: 10.1016/j.bbrc.2007.06.122] [Medline: 17640617]
73. Needleman P, Isakson PC. The discovery and function of COX-2. J Rheumatol Suppl 1997 Jul;49:6-8. [Medline: 9249644]
74. Subbaramaiah K, Dannenberg AJ. Cyclooxygenase 2: a molecular target for cancer prevention and treatment. Trends Pharmacol Sci 2003 Feb;24(2):96-102. [doi: 10.1016/S0165-6147(02)00043-3] [Medline: 12559775]
75. Vane JR, Botting RM. The mechanism of action of aspirin. Thromb Res 2003 Jun 15;110(5-6):255-258. [doi: 10.1016/s0049-3848(03)00379-7] [Medline: 14592543]
76. Millan MJ. The induction of pain: an integrative review. Prog Neurobiol 1999 Jan;57(1):1-164. [doi: 10.1016/s0301-0082(98)00048-3] [Medline: 9987804]
77. Bin W, He W, Feng Z, Xiangdong L, Yong C, Lele K, et al. Prognostic relevance of cyclooxygenase-2 (COX-2) expression in Chinese patients with prostate cancer. Acta Histochem 2011 Feb;113(2):131-136. [doi: 10.1016/j.acthis.2009.09.004] [Medline: 19836060]
78. Rizzo MT. Cyclooxygenase-2 in oncogenesis. Clin Chim Acta 2011 Apr 11;412(9-10):671-687. [doi: 10.1016/j.cca.2010.12.026] [Medline: 21187081]
79. Aktan F. iNOS-mediated nitric oxide production and its regulation. Life Sci 2004 Jun 25;75(6):639-653. [doi: 10.1016/j.lfs.2003.10.042] [Medline: 15172174]
80. Franco L, Doria D, Bertazzoni E, Benini A, Bassi C. Increased expression of inducible nitric oxide synthase and cyclooxygenase-2 in pancreatic cancer. Prostaglandins Other Lipid Mediat 2004 Jan;73(1-2):51-58. [doi: 10.1016/j.prostaglandins.2003.12.001] [Medline: 15165031]
81. Niculescu AB, Le-Niculescu H, Levey DF, Roseberry K, Soe KC, Rogers J, et al. Towards precision medicine for pain: diagnostic biomarkers and repurposed drugs. Mol Psychiatry 2019 Apr;24(4):501-522 [FREE Full text] [doi: 10.1038/s41380-018-0345-5] [Medline: 30755720]
82. Njoku UC, Amadi PU, Agomuo EN, Bhebhe M. The relationship between pain and vascular function biomarkers in dysmenorrheal university students. Chonnam Med J 2020 Sep;56(3):186-190 [FREE Full text] [doi: 10.4068/cmj.2020.56.3.186] [Medline: 33014757]
83. Cortese A, Lombardi R, Briani C, Callegari I, Benedetti L, Manganelli F, et al. Antibodies to neurofascin, contactin-1, and contactin-associated protein 1 in CIDP: clinical relevance of IgG isotype. Neurol Neuroimmunol Neuroinflamm 2019 Nov 21;7(1):e639 [FREE Full text] [doi: 10.1212/NXI.0000000000000639] [Medline: 31753915]
84. Levey DF, Niculescu EM, Le-Niculescu H, Dainton HL, Phalen PL, Ladd TB, et al. Towards understanding and predicting suicidality in women: biomarkers and clinical risk assessment. Mol Psychiatry 2016 Jun;21(6):768-785. [doi: 10.1038/mp.2016.31] [Medline: 27046645]
85. Liu CJ, Dib-Hajj SD, Black JA, Greenwood J, Lian Z, Waxman SG. Direct interaction with contactin targets voltage-gated sodium channel Na(v)1.9/NaN to the cell membrane. J Biol Chem 2001 Dec 07;276(49):46553-46561 [FREE Full text] [doi: 10.1074/jbc.M108699200] [Medline: 11581273]
86. Vawter MP, Ferran E, Galke B, Cooper K, Bunney WE, Byerley W. Microarray screening of lymphocyte gene expression differences in a multiplex schizophrenia pedigree. Schizophr Res 2004 Mar 01;67(1):41-52. [doi: 10.1016/s0920-9964(03)00151-8] [Medline: 14741323]
87. Bell RL, Kimpel MW, McClintick JN, Strother WN, Carr LG, Liang T, et al. Gene expression changes in the nucleus accumbens of alcohol-preferring rats following chronic ethanol consumption. Pharmacol Biochem Behav 2009 Nov;94(1):131-147 [FREE Full text] [doi: 10.1016/j.pbb.2009.07.019] [Medline: 19666046]
88. Gruber HE, Hoelscher GL, Ingram JA, Hanley Jr EN. Genome-wide analysis of pain-, nerve- and neurotrophin -related gene expression in the degenerating human annulus. Mol Pain 2012 Sep 10;8:63 [FREE Full text] [doi: 10.1186/1744-8069-8-63] [Medline: 22963171]
89. Andrus BM, Blizinsky K, Vedell PT, Dennis K, Shukla PK, Schaffer DJ, et al. Gene expression patterns in the hippocampus and amygdala of endogenous depression and chronic stress models. Mol Psychiatry 2012 Jan;17(1):49-61 [FREE Full text] [doi: 10.1038/mp.2010.119] [Medline: 21079605]
90. Liu J, Lewohl JM, Harris RA, Iyer VR, Dodd PR, Randall PK, et al. Patterns of gene expression in the frontal cortex discriminate alcoholic from nonalcoholic individuals. Neuropsychopharmacology 2006 Jul;31(7):1574-1582. [doi: 10.1038/sj.npp.1300947] [Medline: 16292326]

Abbreviations

AI: artificial intelligence
BN: Bayes network
CDSS: clinical decision support system
CNS: central nervous system

CNTN-1: Contactin-1
COX: cyclooxygenase
IL: interleukin
iNOS: inducible nitric oxide synthase
NO: nitric oxide
PGG2: prostaglandin G2
PGH2: prostaglandin H2
RP: reported pain
SUNY: State University of New York
TENS: transcutaneous electrical nerve stimulation
VAS: visual analog scale

Edited by A Mavragani; submitted 14.12.21; peer-reviewed by A Hill, R Rastmanesh; comments to author 18.01.22; revised version received 18.03.22; accepted 10.05.22; published 29.09.22.

Please cite as:
Sadik O, Schaffer JD, Land W, Xue H, Yazgan I, Kafesçiler K, Sungur M
A Bayesian Network Concept for Pain Assessment
JMIR Biomed Eng 2022;7(2):e35711
URL: https://biomedeng.jmir.org/2022/2/e35711
doi:10.2196/35711
PMID:27774989

Original Paper

Accuracy of Fully Automated 3D Imaging System for Child Anthropometry in a Low-Resource Setting: Effectiveness Evaluation in Malakal, South Sudan

Eva Leidman[1], MSPH; Muhammad Ali Jatoi[2], DPH, DHDN, MS; Iris Bollemeijer[3], MSc; Jennifer Majer[3], MA, MSc; Shannon Doocy[1], PhD

[1]Department of International Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States

[2]International Medical Corps, Juba

[3]International Medical Corps, Los Angeles, CA, United States

Corresponding Author:
Eva Leidman, MSPH
Department of International Health
Johns Hopkins Bloomberg School of Public Health
615 N. Wolfe Street
Baltimore, MD, 21205
United States
Phone: 1 4049085125
Email: eleidman@jhu.edu

Abstract

Background: Adoption of 3D imaging systems in humanitarian settings requires accuracy comparable with manual measurement notwithstanding additional constraints associated with austere settings.

Objective: This study aimed to evaluate the accuracy of child stature and mid–upper arm circumference (MUAC) measurements produced by the AutoAnthro 3D imaging system (third generation) developed by Body Surface Translations Inc.

Methods: A study of device accuracy was embedded within a 2-stage cluster survey at the Malakal Protection of Civilians site in South Sudan conducted between September 2021 and October 2021. All children aged 6 to 59 months within selected households were eligible. For each child, manual measurements were obtained by 2 anthropometrists following the protocol used in the 2006 World Health Organization Child Growth Standards study. Scans were then captured by a different enumerator using a Samsung Galaxy 8 phone loaded with a custom software, AutoAnthro, and an Intel RealSense 3D scanner. The scans were processed using a fully automated algorithm. A multivariate logistic regression model was fit to evaluate the adjusted odds of achieving a successful scan. The accuracy of the measurements was visually assessed using Bland-Altman plots and quantified using average bias, limits of agreement (LoAs), and the 95% precision interval for individual differences. Key informant interviews were conducted remotely with survey enumerators and Body Surface Translations Inc developers to understand challenges in beta testing, training, data acquisition and transmission.

Results: Manual measurements were obtained for 539 eligible children, and scan-derived measurements were successfully processed for 234 (43.4%) of them. Caregivers of at least 10.4% (56/539) of the children refused consent for scan capture; additional scans were unsuccessfully transmitted to the server. Neither the demographic characteristics of the children (age and sex), stature, nor MUAC were associated with availability of scan-derived measurements; team was significantly associated (P<.001). The average bias of scan-derived measurements in cm was −0.5 (95% CI −2.0 to 1.0) for stature and 0.7 (95% CI 0.4-1.0) for MUAC. For stature, the 95% LoA was −23.9 cm to 22.9 cm. For MUAC, the 95% LoA was −4.0 cm to 5.4 cm. All accuracy metrics varied considerably by team. The COVID-19 pandemic–related physical distancing and travel policies limited testing to validate the device algorithm and prevented developers from conducting in-person training and field oversight, negatively affecting the quality of scan capture, processing, and transmission.

Conclusions: Scan-derived measurements were not sufficiently accurate for the widespread adoption of the current technology. Although the software shows promise, further investments in the software algorithms are needed to address issues with scan transmission and extreme field contexts as well as to enable improved field supervision. Differences in accuracy by team provide evidence that investment in training may also improve performance.

(JMIR Biomed Eng 2022;7(2):e40066) doi:10.2196/40066

KEYWORDS

mobile health; mHealth; child nutrition; anthropometry; 3D imaging; imaging; accuracy; measurement; child stature; software; algorithm; automated; device; child health; pediatric health; height; length; arm circumference

Introduction

Background

Anthropometric measurement of children is a standard component of pediatric care to enable growth monitoring as well as population-level assessments and clinical research. Despite widespread reliance on anthropometry, there has been limited technological advancement in measurement equipment. The accuracy of weight measurement was improved with the transition from spring to digital scales in the 1980s [1,2]. However, until recently, measurements of recumbent length and standing height have not benefited from similar innovations; commonly used stadiometers, or height boards, are heavy, wooden devices that are robust to field conditions but inconvenient to transport, as is commonly done for field surveys and community screenings in low-resource settings.

In recent years, 2 different types of mobile device–based technologies have been proposed as alternatives to manual anthropometry using stadiometers: (1) apps using geometric morphometric models and (2) apps using 3D imaging systems. Using a portable camera attached to a standard tablet and preloaded software, these imaging systems are able to estimate child stature (length or height), head circumference, and mid–upper arm circumference (MUAC) from a 3D model developed from a series of image captures. Geometric morphometric models aim to directly classify a child as severely or moderately acutely malnourished; examples include the Severe Acute Malnutrition Photo Diagnosis App [3] developed by Action Against Hunger Spain and the Methods of Extremely Rapid Observation of Nutrition Status developed by Kimetrica [4]. In contrast, 3D imaging system technology—currently used by the Child Growth Monitor, developed by the nonprofit Welthungerhilfe, and AutoAnthro, developed by the company Body Surface Translations Inc (BST)—produces estimates of anthropometric measurements, which can then be used to characterize nutrition status.

Preliminary validation studies for software aiming to directly classify acute malnutrition have encountered methodological as well as logistic challenges. The Photo Diagnosis App validation phase study in Spain and Senegal found high accuracy of diagnosis but suggested significant morphometric differences among the populations sampled, implying a need to investigate this morphological variability [5,6]. The researchers involved in the study noted that, although morphological variability could likely be overcome with machine learning, the approach proved very expensive compared with current technologies and that capturing a viable scan required conditions that could not be repeated in the field (AV Brizuela, personal communication, February 25, 2022). Lower accuracy was obtained by the Methods of Extremely Rapid Observation of Nutrition Status software during an initial pilot in Kenya; work is ongoing to improve performance with further calibration using a larger, multicountry data set [4].

Although the Child Growth Monitor is still in development and beta testing, several studies have evaluated the performance of AutoAnthro. Initial efficacy studies demonstrated that, in a controlled setting in Georgia, United States, devices were able to achieve high precision—the reliability of repeated 3D scans was within 1 mm of manual measurement for stature, head circumference, and MUAC; however, systematic biases were reported [7]. Replication studies in Guatemala, Kenya, and China aimed to determine whether the systematic biases observed were generalizable across populations and, therefore, something that could be corrected analytically. However, the multicountry replication studies found lower accuracy and variability in the direction and magnitude of bias [8].

Further testing in a humanitarian setting was proposed given the additional challenges for nutrition surveillance in these locations. Settings hosting internally displaced persons and refugees are commonly remote, experience austere weather conditions, and have limited or no internet connectivity. These conditions present unique operating challenges for training and use of 3D imaging technology. Changes to the software and hardware were implemented to ensure that the device and operating software were robust to and acceptable in these conditions. In addition, the timing of the evaluation—during the acute phase of the COVID-19 pandemic—presented new challenges. Travel and movement restrictions necessitated a more autonomous operation. In addition, concern about transmission risk associated with the physical contact required for traditional anthropometry, particularly height and length measurement, created additional interest in the potential for 3D imaging technology.

Objectives

Widespread adoption of the AutoAnthro technology in humanitarian settings requires accuracy at least comparable with manual measurement notwithstanding additional constraints. Therefore, this study aimed to re-evaluate device accuracy following modifications to the software algorithm.

Methods

Overview

This study evaluated the accuracy of the third generation of the AutoAnthro 3D imaging system in comparison with manual measurements for child anthropometry. The third-generation software contained major updates to the previous version that were designed to achieve higher levels of durability and portability required in austere settings, improve user experience with scan capture and device performance, automate image processing, and implement changes to allow the software to operate on lower-cost hardware. Details on the hardware, positioning, data capture, and processing for the AutoAnthro technology used in this study and previous versions are compared in Table 1. Given the interest in the field-readiness of the technology, the study was embedded within a population-representative household nutrition survey conducted by International Medical Corps (IMC) to simulate the level of automation required to enable use by nonresearch actors in nutrition surveys [9]. The survey was undertaken in late 2021 (September 27 to October 2) at the Malakal Protection of Civilians site, which hosts approximately 34,000 internally displaced people and is located in the northeast of South Sudan [10].

Table 1. Hardware, data acquisition, review, and processing used by AutoAnthro technology to produce automated measurements of anthropometry for children.

	First generation	Second generation	Third generation
Hardware	iPad and a structure sensor 3D scanner	iPad and a structure sensor 3D scanner	Samsung Galaxy 8 phone running Android and an Intel RealSense 3D scanner
Positioning	Enumerators were unable to constrain the child's hands or feet to help position them	Enumerators were able to constrain the child's hands or feet to help position them	Enumerators were able to constrain the child's hands or feet to help position them
Real-time estimates	Not available	Not available	Available
Number of scans	Unlimited scans	Fixed number of scans automatically captured	Fixed number of scans automatically captured
Data acquisition	Automatically uploaded to a computer server	Automatically uploaded to a computer server	Automatically uploaded to a computer server
Data review	Scans manually screened for data quality by enumerators	No manual screening by enumerators	No manual screening by enumerators
Data processing	Semiautomatic	Fully automatic	Fully automatic
Evidence on performance	Initial efficacy study in the United States [7]	Replication studies in Guatemala, Kenya, and China [8]	Used in this study in South Sudan

Study Design and Data Collection

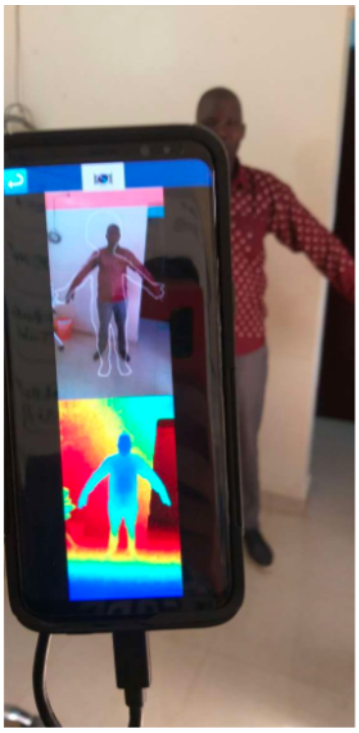
Households were sampled using a 2-stage cluster sampling design in which camp blocks were selected with probability proportional to size. Selected blocks were fully enumerated, and households were randomly selected using systematic random sampling. A sample of 485 children was targeted to achieve desired precision for estimating prevalence of global acute malnutrition, the primary survey aim. This sample was determined to be sufficient to detect a difference of 0.17 cm for height/length and 0.09 cm for MUAC given an α of .05, power of 0.8, and SDs observed in previous studies [5]. All children aged 6 to 59 months within selected households whose primary caregiver gave verbal informed consent were eligible to participate.

Staff from BST remotely trained the IMC survey manager; training included instructions on the positioning of children, use of the hardware and AutoAnthro software, and performing and saving scans. The survey manager replicated the training in person for the enumerators. Manual anthropometrics and 3D scan teams received a 4-day training. Teams jointly participated in a classroom training on study objectives and manual anthropometry. Practical exercises and a standardization test were organized separately for manual anthropometrists and scanners. All manual measurers passed the standardization test with an intra- and interenumerator technical error of measurement (TEM) for manual measurement of <1.4 for height/length and <3.0 for MUAC.

Measurements were performed by 6 teams of 4 individuals, including 2 (50%) measurers, 1 (25%) team leader trained on manual anthropometry, and 1 (25%) measurer trained to obtain the 3D scan-derived measurements. For a given child, manual measurements (weight, height/length, and MUAC) were separately obtained by 2 manual anthropometrists and entered into a survey programmed in Open Data Kit (Get ODK) on a tablet device. Anthropometrists first collected weight and MUAC; height or length was then collected following the protocol used for the 2006 World Health Organization (WHO) Child Growth Standards study [11]. After manual measurement, scans were taken by a different enumerator to ensure independence. Scans were captured using the AutoAnthro (version 3) software on a Samsung Galaxy 8 Android phone and an Intel RealSense 3D scanner (Figure 1). Each 3D imaging session comprised 10 scans, with 5 scans of both the front and back of the child. Two sets of measurements were produced: (1) a real-time, offline estimate of height/length and MUAC, which was produced by the app and displayed to data collection teams while in households for plausibility checks, and (2) an updated measurement. For the updated measurement, scan data were uploaded to a cloud server for fully automated processing, which is a slower, more computationally intensive, and generally more rigorous version of the phone-based algorithm with additional error checking for unanticipated positioning of the child. The software then compared the real-time result with the updated measurement and reported the more *consistent result*, which was identified by comparing the SD of the individual subscans (6-10 subscans produce 1 scan). In cases where the scan data were too poor to produce either real-time or updated measurements, AutoAnthro did not report results. For 11.9% (64/539) of the children, enumerators perceived the real-time estimate as implausible and took additional scans. If multiple scan sessions were performed for a given child, the median of all the individual scan sessions was used. Real-time scans were primarily used to evaluate whether they could be used for identification and referral of wasted children, whereas the analysis focused on the updated but still fully automated measurements.

Given a large number of lost scans following initial quality checks and data processing, in-depth interviews were conducted with enumerators (4 group interviews with a total of 12 enumerators) and BST staff (3 individual interviews) to document challenges. The interviews were conducted remotely in English from the United States using a semistructured interview guide and recorded to facilitate note taking.

Figure 1. Example scan capture using the AutoAnthro software (version 3; Body Surface Translations Inc) on a Samsung Galaxy 8 Android phone and an Intel RealSense 3D scanner.



Analysis Methods

Differences in demographic characteristics, nutritional status of the children, and team number between children with and without scan-derived measurements were evaluated to assess characteristics associated with successful scan captures. For unadjusted comparisons, the statistical significance of differences was evaluated using the Kruskal-Wallis test for continuous variables and the Fisher exact test for categorical variables. A multivariate logistic regression was fit to evaluate the adjusted odds of achieving a successful scan.

The quality of anthropometric measurements was assessed using standard indicators—digit preference scores, proportion of outlier values, and SDs [12]. The digit preference score was calculated for height and MUAC applying the MONICA procedure, which adjusts the chi-square statistic according to the size of the sample and the df of the test. Digit preference scores values of 0 indicate a uniform distribution, and the values increase with a greater imbalance [13]. Weight-for-height or weight-for-length *z* score (WHZ) and height-for-age or length-for-age *z* score (HAZ) were calculated using the WHO growth standard [14]. Outliers were calculated using two approaches: (1) fixed exclusions of WHZ values of <−5 or >5 and HAZ values of <−6 or >6 and (2) flexible exclusions of WHZ and HAZ values of <−3 or >3 from the observed median. Measurements outside the range for which *z* scores could be generated (length: 45-110 cm for children aged <2 years; height: 65-120 cm for children aged >2 years) were also excluded. The SDs of the MUAC, WHZ, and HAZ distributions were evaluated after exclusion of outliers.

The accuracy of the measurements was visually assessed using Bland-Altman plots [15] to evaluate whether accuracy remained constant across different child body sizes and look at random bias. For the y-axis of the Bland-Altman plots, the manual measurement was subtracted from the automated scan estimate and, for the x-axis, the mean of the manual and scan estimates was used. The average bias was assessed as a metric of systematic bias (Equation 1). The limits of agreement (LoAs), that is, the 95% precision interval for individual differences, were calculated as a metric of random bias. The Pitman test of difference in variance [16] was used to test the correlation between accuracy and size of the child. TEM, an accuracy index used to express the error margin, was calculated as described by Ulijaszek and Kerr [17] (Equation 2). Analyses were performed for all children who had both scan-derived and manual measurements as well as disaggregated based on team, sex, and age groups corresponding to measurement positioning (length: ages of 6-23 months; height: ages of 24-59 months).

where N is the number of children measured, $M_{i1}$is the scan-derived measurement for child i, and $M_{i2}$is the manual measurement for the same child.

In total, 2 distinct problems with data capture were identified during the analysis. First, the AutoAnthro software estimates height, length, and MUAC by measuring the distance between the reference markers on the 3D image of the child. Visual inspection of the 3D images used to generate scan values suggested that, in select instances, the software identified a caregiver in the background, resulting in a misplaced reference marker, which typically resulted in outlier estimates. Second, the scans and manual measurements were linked using a unique ID number. Age, sex, and weight were determined for each child and entered independently by the scan-derived and manual measurement. For select children, child IDs matched across the 2 data sets, but age, sex, or weight were discordant, suggesting a potential mismatch between scan-derived and manual measurements. To evaluate the implications of these two data capture errors, Bland-Altman plots and all accuracy metrics were also calculated after excluding records with outliers or mismatches in sex, age (>6 months), or weight (>5 kg).

To evaluate the implications measurement differences would have on the classification for each derived nutrition indicator (WHZ, HAZ, and MUAC), children were classified as severely, moderately acutely malnourished, or neither using both manual and scan-derived measurements. The concordance of the classification was tabulated and visually explored. For WHZ and HAZ, values of <−3 were considered severe, and values between ≥−3 and <−2 were considered moderate. For MUAC, values of <11.5 cm were considered severe, and values of ≥11.5 cm to <12.5 cm were considered moderate. All quantitative analyses were performed in RStudio (version 1.1.456 20; R Foundation for Statistical Computing). For the qualitative analysis, detailed notes were taken during in-depth interviews, supported by automated transcription available from the Microsoft Teams software, and reviewed to synthesize key themes. The results were triangulated with the quantitative data and used to interpret and explain the quantitative findings.

Ethics Approval

Johns Hopkins Institutional Review Board approved the study as “non-human subjects research” per DHHS regulations 45 CFR 46.102. The caregivers of the children enrolled in the study and key informants provided verbal informed consent. The study data retained for analysis were deidentified. Children identified as malnourished were referred for care, and no further compensation was provided.

Results

Study Sample

A total of 416 households were visited, of which 325 (78.1%) had age-eligible children and consented to participate. Manual anthropometric measurements were obtained for all children aged 6 to 59 months (N=539) in the enrolled households, and scan-derived measurements were successfully processed for 43.4% (234/539) of the children. Caregivers of 10.4% (56/539) of the children refused consent for scan capture; in addition, a large number of reportedly captured scans were unsuccessfully transmitted to the server and could not be recovered from the devices (Figure 2). A total of 485 scans were successfully transmitted to the server for processing, of which 373 (76.9%) were from unique children (when multiple scan sessions were conducted, they were combined to produce a single estimate for the child). After merging sessions and removing poor-quality scans, scan-derived estimates were available for 265 individuals, of which 234 (88.3%) could be matched to manual measurements. A detailed breakdown of the available data by cluster is provided in Table S1 in Multimedia Appendix 1.

Among the final sample with both manual and scan-derived measurements, approximately equal proportions were from male participants (119/234, 50.9%) and female participants (114/234, 48.7%), and two-thirds (154/234, 65.8%) were from participants aged 24 to 59 months. The prevalence of wasting as classified by WHZ (46/234, 19.7%) exceeded that of underweight (40/234, 17.1%) or stunting (32/234, 13.7%) when using manual measurements; no children with edema were identified. When comparing the demographic characteristics and nutritional status of children with and without scan-derived measurements, there were no significant differences apart from mean age; children with both measurements were older (31.7, SD 14.9 months vs 28.9, SD 14.5 months; *P*=.03; Table 2). However, the availability of scan-derived measurements was significantly associated with team (*P*<.001), where 81% (69/85) of the children measured by team 1 had successfully transmitted scans compared with only 4% (3/81) of the children measured by team 6. Differences in the availability of scan-derived measurements by team remained significant in the multivariate logistic regression, whereas child characteristics (age, sex, height/length, and MUAC) were not associated with scan availability (Table S2 in Multimedia Appendix 1).

Figure 2. Enrollment flowchart. *Scans captured in the field that were unsuccessfully transmitted to the server and could not be recovered from devices or child identification numbers that were misreported such that scan-derived values could not be matched to manual measurements. **In five of the 32 clusters, information on the outcomes of each household visit was recorded on paper forms lost in a large rainstorm during data collection. ***The child identification number associated with the scan was not a match to any children with manual measurements. ****Scan positioning or resolution was too poor to enable calculation of scan-derived measurements. *****Field teams conducted two or more scan sessions for 64 children. When multiple scan sessions were performed for a given child, scans from all available sessions were combined, and the median of all individual scan sessions was used for the analysis.

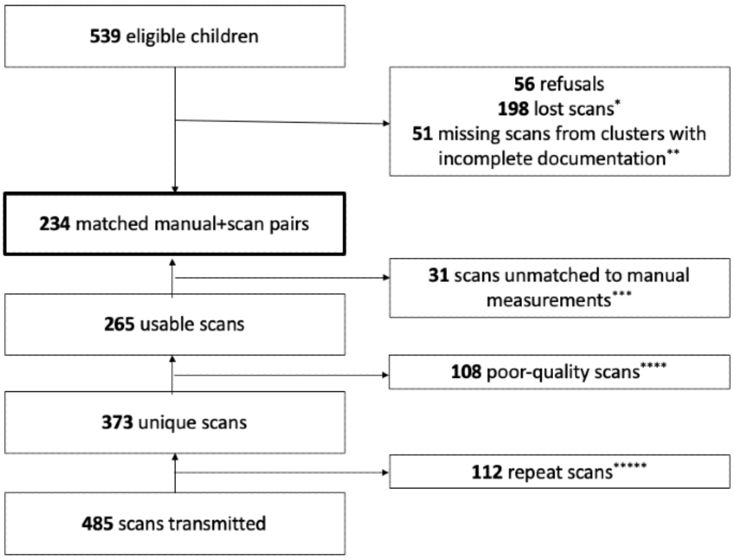

539 eligible children
56 refusals
198 lost scans*
51 missing scans from clusters with incomplete documentation**
234 matched manual+scan pairs
31 scans unmatched to manual measurements***
265 usable scans
108 poor-quality scans****
373 unique scans
112 repeat scans*****
485 scans transmitted

Table 2. Characteristics of the sample by availability of automated scan data (N=539).

	Children with manual measurements only (n=305)	Children with manual and scan-derived measurements (n=234)	*P* value[a]
Team, n (%)			<.001
Team 1 (n=85)	16 (18.8)	69 (81.2)	
Team 2 (n=112)	53 (47.3)	59 (52.7)	
Team 3 (n=106)	61 (57.5)	45 (42.5)	
Team 4 (n=91)	78 (85.7)	13 (14.3)	
Team 5 (n=64)	19 (29.7)	45 (70.3)	
Team 6 (n=81)	78 (96.3)	3 (3.7)	
Age (months), mean (SD)	28.9 (14.5)	31.7 (14.9)	.03
Age category (months), n (%)			.14
6 to 23	123 (40.3)	80 (34.2)	
24 to 59	182 (59.7)	154 (65.8)	
Sex, n (%)			.89
Female	148 (48.5)	115 (49.1)	
Male	157 (51.5)	119 (50.9)	
Underweight[b], mean (SD)	−1.2 (1.0)	−1.2 (1.0)	.36
Underweight category[c], n (%)			.99
Severe	13 (4.3)	10 (4.3)	
Moderate	42 (13.8)	30 (12.8)	
Stunting[b], mean (SD)	−0.9 (1.4)	−0.9 (1.2)	.39
Stunting category[c], n (%)			.92
Severe	17 (5.6)	11 (4.7)	
Moderate	32 (10.5)	21 (9)	
Wasting[b], mean (SD)	−1.1 (1.1)	−1.0 (1.2)	.86
Wasting category[c], n (%)			.58
Severe	10 (3.3)	9 (3.8)	
Moderate	42 (13.8)	37 (15.8)	
MUAC[d], mean (SD)	14.0 (1.2)	14.0 (1.3)	.52
MUAC category[c], n (%)			.52
Severe	4 (1.3)	1 (0.4)	
Moderate	23 (7.5)	21 (9)	

[a] Kruskal-Wallis test for continuous variables; Fisher exact test for categorical variables.

[b] Underweight was classified as weight-for-age *z* score, stunting was classified as height- or length-for-age *z* score, and wasting was classified as weight-for-height or weight-for-length *z* score using the World Health Organization growth reference based on manual measurements.

[c] For underweight, stunting, and wasting, moderate categories included *z* score values between −2 and ≥−3. The severe categories included values of <−3. For mid–upper arm circumference, the moderate category included values between 11.5 cm and 12.5 cm; values of <11.5 cm were classified as severe.

[d] MUAC: mid–upper arm circumference.

Measurement Quality

The quality of manual measurements was evaluated for the overall sample as well as for the subset matched to scan-derived measurements. Among children with both manual and scan-derived data (234/539, 43.4%), the digit preference score for height/length was nearly twice as high for manual measurements (13.5 vs 6.8) and nearly 3 times as high (19.6 vs 6.6) for MUAC measurements as compared with scan-derived measurements, suggesting more rounding of terminal digits by manual anthropometrists. For all other quality metrics, manual measurements outperformed scan-derived measurements. For children with both measurements, no outliers were identified with fixed exclusions, and only 11 were identified with flexible exclusions (n=3, 27% for WHZ and n=8, 73% for HAZ) from the manual measurements. By comparison, for scan-derived measurements, 29 outliers were identified applying fixed exclusions (n=13, 45% for WHZ and n=16, 55% for HAZ), and 61 were identified with flexible exclusions (n=22, 36% for WHZ and n=39, 64% for HAZ). SDs were notably wider for scan-derived measurements than for manual measurements for MUAC (2.33 vs 1.26), WHZ (1.56 vs 1.16), and HAZ (1.75 vs 1.23) for all children, and the same pattern was observed for length among younger children, for whom measurement can be a greater challenge. For all quality indicators, the results were similar when quality metrics for scan-derived measurements were compared with the sample of all children (N=539) with manual measurements (Table 3).

Table 3. Quality of manual and scan-derived measurements as evaluated using digit preference score, outliers, and SD (N=539).

	Manual measurement		Scan-derived measurements
	All children	All children with scans (n=234)	All children with scans (n=234)
Digit preference score			
Height or length	12.62	13.47	6.78
MUAC[a]	20.87	19.63	6.63
Outlier values (fixed)[b], N			
WHZ[c]	0	0	13
HAZ[d]	1	0	16
Outlier values (flexible)[e], N			
WHZ	6	3	22
HAZ	17	8	39
SD[f] (children aged 6-59 months)			
MUAC	1.21	1.26	2.33
WHZ	1.11	1.15	1.56
HAZ	1.25	1.23	1.75
SD[f] (children aged 6-23 months)			
MUAC	0.97	0.90	2.09
WHZ	1.19	1.28	1.59
HAZ	1.31	1.31	1.96

[a] MUAC: mid–upper arm circumference.

[b] Z score values <−5 or >5 for weight-for-height or weight-for-length and <−6 or >6 for height- or length-for-age were considered outliers, as were measurements outside the range for which *z* scores could be generated (length: 45-110 cm for children aged <2 years; height: 65-120 cm for children aged >2 years).

[c] WHZ: weight-for-height or weight-for-length *z* score.

[d] HAZ: height-for-age or length-for-age *z* score.

[e] Z score values <−3 or >3 from the median *z* score of the sample for WHZ and HAZ were considered outliers, as were measurements outside the range for which *z* scores could be generated (length: 45-110 cm for children aged <2 years; height: 65-120 cm for children aged >2 years).

[f] SD calculated after excluding outlier values (weight-for-height or weight-for-length and height-for-age or length-for-age).

Accuracy of Measurement

Analysis of scan-derived measurement accuracy used updated measurements generated after measurements were uploaded to cloud-based servers for automated processing. Real-time scan-derived measurements that were available in the field were reviewed to confirm adherence to the protocol to ensure that scan-derived measurements were not shared with manual anthropometrists (Figure S1 in Multimedia Appendix 1). Only 4 height/length and 3 MUAC manual and real-time scan-derived measurements were exact matches, providing evidence of the independence of the measurements. The correlations between manual and real-time scans were 0.54 for height/length and 0.17 for MUAC.

Accuracy was visually inspected using Bland-Altman plots (Figure 3). When using all scans, the average bias of the measurements in cm was −0.5 (95% CI −2.0 to 1.0) for height/length and 0.7 (95% CI 0.4-1.0) for MUAC (Table 4). For height and length, 48.7% (114/234) of scan-derived measurements were higher than manual measurements or positive, and the 95% LoA was within −23.9 cm and 22.9 cm. For MUAC, 67.9% (159/234) of scan-derived measurements were higher than manual measurements, and the 95% LoA was −4.0 cm to 5.4 cm. For both indicators, the Pitman test was statistically significant (*P*<.001), suggesting differential accuracy by child size. Mean differences in height/length were negative for all children but greater for children aged 24 to 59 months compared with children aged 6 to 23 months, whereas the reverse was true for MUAC. Of interest, for height/length, the LoAs were narrower for female participants than for male participants (Figure S2 in Multimedia Appendix 1), but the mean difference was greater (−1.1 vs 0.1); the sex difference in the accuracy of MUAC was less pronounced.

Accuracy metrics varied considerably by team, excluding team 6 given the small sample of scans (n=3). For height/length, the mean difference was the greatest for team 5 (−3.8) and smallest for teams 1 (−0.2) and 3 (0.2). The width of the 95% LoAs for team 5 (−42.8 to 35.2) was nearly 3 times that of team 1. For MUAC, the mean differences were positive for teams 1 to 4 but negative for team 5, and the 95% LoAs for team 5 (−7.2 to 4.8) exceeded those of all other teams (Table 4). Differences in correlation between the manual and scan-derived measurements are visualized by team in Figure S1 in Multimedia Appendix 1.

Given the relatively wide LoAs observed in the sample overall, a sensitivity analysis was used to explore how potential errors in data capture and matching of scan-derived and manual measurements contributed to the overall accuracy (Table 4 and Figure S3 in Multimedia Appendix 1). When outlier values (n=19) and discordant pairs (n=63) were excluded, the 95% LoA was reduced (−11.9 cm to 11.4 cm for height/length and −3.5 cm to 5.1 cm for MUAC). The mean difference was reduced to −0.2 (95% CI −1.2 to 0.7) for height/length and increased marginally for MUAC; the Pitman test remained significant for both indicators.

The TEM for height/length among children of all ages was 8.4 cm; TEM is analogous to the SD, indicating that the scan-derived measurements were within –8.4 cm to +8.4 cm of manual measurements for 2 out of 3 children and within –16.8 cm to +16.8 cm for 95% of the children. The TEM for height/length was higher for children aged 24 to 59 months, male participants, and team 5. The TEM was lowest (4.2) when flagged and discordant pairs were removed. The TEM for MUAC among all children was 1.8 cm, with more limited variation by age and sex. TEM was highest for team 5 for MUAC; excluding flagged and discordant values reduced the TEM for MUAC to 1.6 cm.

The implications of measurement differences on classification of nutrition status were characterized for each indicator (WHZ, HAZ, and MUAC). Children were classified as severe, moderate, or normal using scan-derived and manual measurements independently, and the classifications were compared (Figure 4). Classifications were concordant for 66.7% (156/234) of the children by WHZ, 61.9% (145/234) of the children by HAZ, and 77.4% (181/234) of the children by MUAC. However, among children with low WHZ (<−2) by manual measurement, only 32% (18/56) were identified as wasted by scan-derived measurements. Similarly, among stunted (HAZ <−2) and wasted children by MUAC (MUAC <125 cm), only 23% (9/40) and 14% (3/22) were identified as stunted and wasted by scan-derived measurements, respectively.

Key informants identified several issues unique to field data collection that may have affected device performance. The following section highlights issues related to (1) beta testing or validating the device algorithm, (2) training and field supervision, (3) data capture or field work, and (4) data transmission.

Figure 3. Bland-Altman plot of child stature (height and length) and mid–upper arm circumference (MUAC) comparing manual and scan-derived measurements.

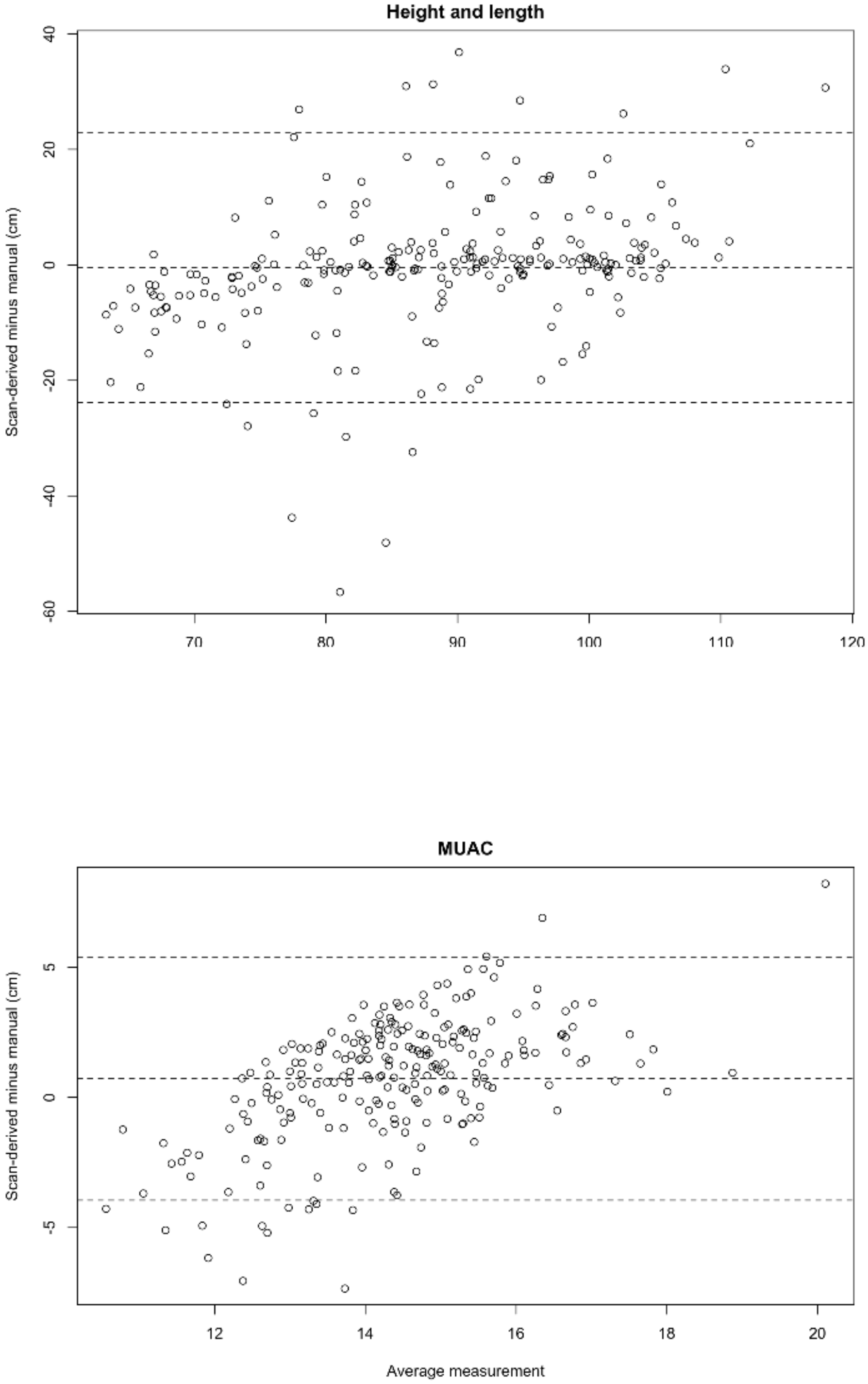


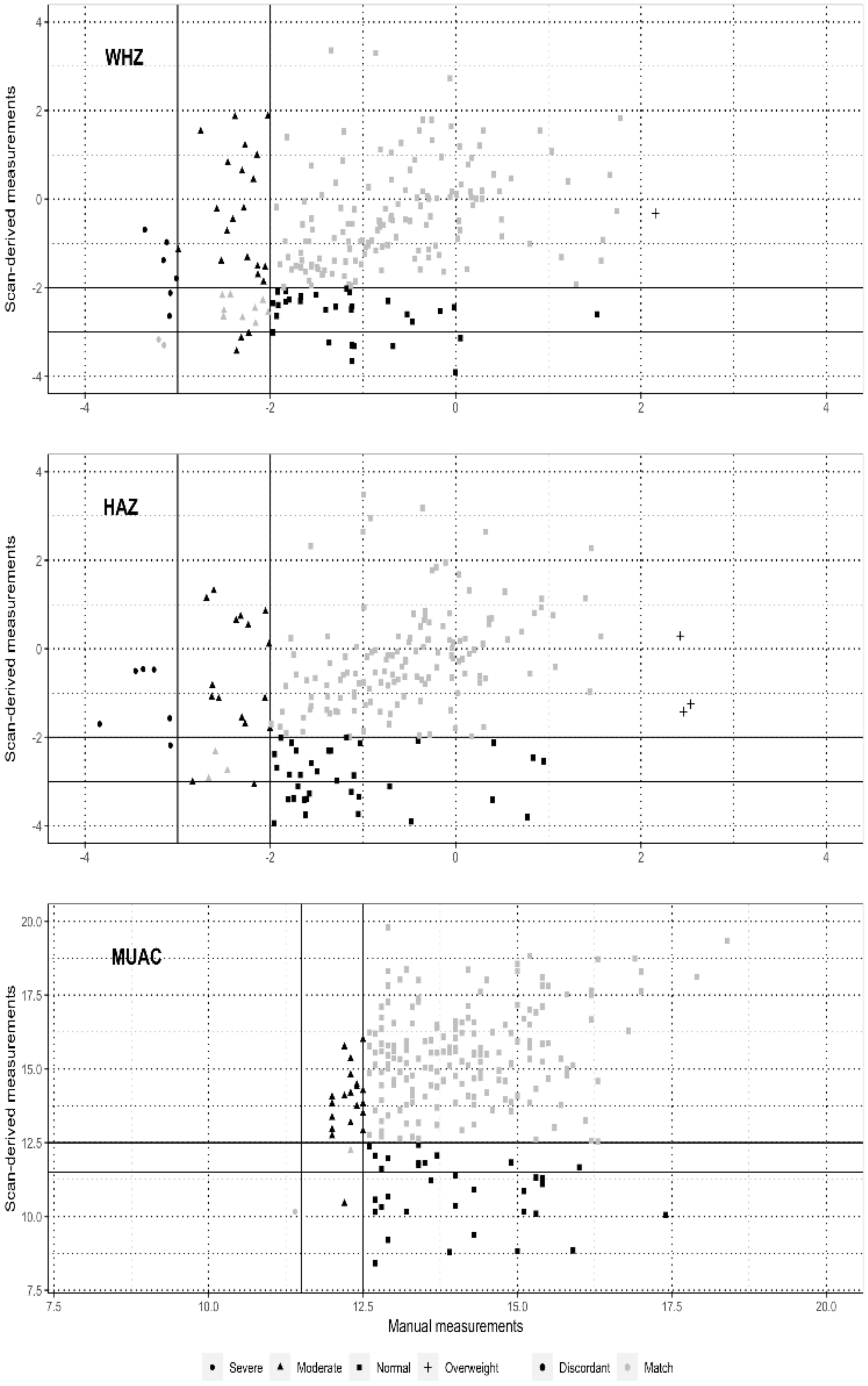
Table 4. Statistical evaluation of differences between manual and scan-derived measurements (N=234).

	Technical error of measurement	Mean difference in cm (95% CI)	95% limits of agreement (cm)	Pitman test		Children, n (%)
				r	*P* value	
Height or length						
All children	8.41	−0.50 (−2.03 to 1.04)	−23.86 to 22.86	0.34	<.001	234 (100)
Excluding flagged and discordant values[a]	4.18	−0.24 (−1.19 to 0.71)	−11.86 to 11.38	0.41	<.001	151 (64.5)
Age (months)						
0 to 23	8.02	−0.35 (−2.88 to 2.19)	−22.7 to 22.01	0.65	<.001	80 (34.2)
24 to 59	8.62	−0.58 (−2.52 to 1.37)	−24.51 to 23.36	0.46	<.001	154 (65.8)
Sex						
Female	6.85	−1.11 (−2.89 to 0.68)	−20.06 to 17.85	0.37	<.001	115 (49.1)
Male	9.69	0.09 (−2.41 to 2.59)	−26.88 to 27.07	0.33	<.001	119 (50.9)
Team[b]						
Team 1	5.10	−0.15 (−1.89 to 1.60)	−14.37 to 14.08	0.20	.10	69 (29.5)
Team 2	7.20	0.87 (−1.79 to 3.54)	−19.19 to 20.94	0.33	.01	59 (25.2)
Team 3	5.64	0.15 (−2.27 to 2.57)	−15.65 to 15.94	0.33	.03	45 (19.2)
Team 4	8.34	−2.42 (−9.68 to 4.84)	−25.98 to 21.14	0.44	.13	13 (5.6)
Team 5	14.18	−3.80 (−0.978 to 2.19)	−42.83 to 35.24	0.53	<.001	45 (19.2)
Mid–upper arm circumference						
All children	1.76	0.72 (0.41 to 1.03)	−3.95 to 5.39	0.56	<.001	234 (100)
Excluding flagged and discordant values[a]	1.64	0.78 (0.43 to 1.14)	−3.53 to 5.10	0.51	<.001	151 (64.5)
Age (months)						
0 to 23	1.64	0.97 (0.5 to 1.44)	−3.18 to 5.12	0.70	<.001	80 (34.2)
24 to 59	1.82	0.59 (0.19 to 0.99)	−4.33 to 5.50	0.58	<.001	154 (65.8)
Sex						
Female	1.73	0.69 (0.26 to 1.13)	−3.93 to 5.32	0.54	<.001	115 (49.1)
Male	1.78	0.75 (0.31 to 1.19)	−3.99 to 5.48	0.57	<.001	119 (50.9)
Team[b]						
Team 1	1.54	1.25 (0.82 to 1.68)	−2.28 to 4.78	0.55	<.001	69 (29.5)
Team 2	1.77	1.45 (0.91 to 1.98)	−2.60 to 5.49	0.52	<.001	59 (25.2)
Team 3	1.40	1.08 (0.58 to 1.59)	−2.19 to 4.35	0.41	.01	45 (19.2)
Team 4	1.37	0.07 (−1.14 to 1.29)	−3.86 to 4.01	0.28	.36	13 (5.6)
Team 5	2.30	−1.22 (−2.13 to –0.3)	−7.20 to 4.77	0.61	<.001	45 (19.2)

[a] Records were excluded if the absolute height measurement derived from the scans was out of the range, if the weight-for-height or weight-for-length *z* score or height- or length-for-age *z* score calculated from the scanned value was considered an outlier (fixed exclusion), if the sex recorded by the manual anthropometry and the scan teams was different, or recorded ages differed by >6 months or recorded weight differed by >10 kg.

[b] Team 6 was excluded given the small number (n=3) of children with both scan-derived and manual measurements.

Figure 4. Classification of nutritional status based on manual and scan-derived measurements. HAZ: height- or length-for-age z score; MUAC: mid–upper arm circumference; WHZ: weight-for-height or weight-for-length z score.



Validation of Device Algorithm (Beta Testing)

The third-generation software included major software changes aimed at full automation as well as a transition to the Android platform. However, the COVID-19 pandemic and the resulting social distancing policies limited the ability of developers to test and refine the new algorithms as they had done following previous substantive revisions of the software:

> *We came up with this newer device, which was on an Android phone with an Intel scanner to try to provide real time results. We knew that we needed to get some preliminary data here in the US to test out the system and also to validate the algorithms and to try and make some revisions to the software you always have to come back and kind of and tweak the estimation algorithms. Because of the pandemic, we really got*

> *limited to [the developer's] children...the kids that I really got a chance to test on were roughly from the ages of 10 to 16. So, none of the real younger kids. So, I'd say, that was really a big hold back for us. We really didn't get to test the device rigorously before we had to send it to South Sudan and let them try and run a pilot trial.*

The age of the children used for validating the algorithm may have been particularly relevant given the difference in how the software operated for younger children measured in a supine position (the software identified the child's heel) compared with older children measured while standing (the software identified the floor). Beta testing in South Sudan revealed that the software's ability to identify children in the supine position was more erratic. The algorithms were further adjusted before field work began. However, developers reported that further beta testing in the United States before deployment of the devices would have been valuable, particularly given the challenges in pushing software updates to South Sudan.

In addition to the demographics of the children included in the initial beta testing, pandemic-related travel restrictions meant that all prestudy testing to validate the algorithm was primarily performed in the United States in well-lit, indoor spaces. To address this, a pilot study in South Sudan was conducted in June 2021 and July 2021. Data collection in Malakal was delayed several months to allow developers time to update the software to address issues identified (eg, scan capture taking a prolonged period) before data collection. Donor deadlines prevented developers from taking more time to refine the algorithm before field-testing the updated software.

Training and Field Supervision

South Sudan mandated a 2-week quarantine for international travelers during the study period, which prevented BST developers from traveling to conduct training and field supervision as had been done in previous surveys. Training of trainers was conducted via web-based video conferencing by the BST team from the United States, in contrast to all previous evaluations of the AutoAnthro technology. This was perceived as a major barrier as it limited the ability to quickly identify small errors in data capture during training as well as support with technical troubleshooting.

The limitations of remote training were particularly relevant during the exercise in which the enumerators practiced taking scans on children. Observing these measurements remotely proved to be impractical:

> *At one point we were on the phone with the group at IMC and they were trying to capture data on a child and we were getting really just garbage data. It wasn't any good at all. Not usable. And we couldn't figure out what the problem was. But it turned out that two or three of the enumerators were using the device at the same time on a child. Now if any of us had been there, we would have corrected that problem in five seconds. Because we were using a structured light approach to generate these models, when one camera is looking for its pattern, if another camera is also running at the same time, it's generating a pattern that interferes with the first one. It's a 5 second problem in person that we didn't figure out for a day or two.*

During the training of enumerators, a single individual on each team was identified to be trained on the AutoAnthro technology. Enumerators felt that training all enumerators would have enabled them to better support each other; in particular, team leads (who were not trained on the software or positioning) felt unempowered to supervise the quality of the scans by their teams. In previous studies, all team members were trained on the technology:

> *It's technical work. We need more training for all people...People are trained together, and some are very quick at capturing what [information] we were given in the training. In class, we are not equal...It's good for people to be trained in one place together and then select [individuals to do scans] who would be the best to do the job.*
>
> *[We] needed additional days for training on the device. [BST] needed to train all of us (not just the scanner) so we could help each other especially on positioning. More than 2 days for piloting and device training are needed. Maybe 4 or 5 days on the scanners so we have enough time to practice.*

Key informants, including both enumerators and BST developers, felt that the training could have been improved by increasing the duration and improving the training materials and protocol. Although the duration of the training for this study was similar to that for previous evaluations of the AutoAnthro technology, key informants felt that, in retrospect, the training was too short. In addition, training was organized using a manual that was written in English with few photographs. As most enumerators did not speak English or spoke English as a second or third language, a more visual field manual was recommended. Finally, a standardization test (where 10 children were measured twice by each enumerator) was performed for both manual and scan-derived measurements, but only manual results were evaluated for accuracy and precision, which was perceived as a limitation.

Data Capture

Data collection occurred during the summer period in South Sudan, where the temperature was consistently >100 °F (38 °C) and, on many days of data collection, the teams experienced a downpour of rain. When it was sunny, scans were commonly performed outside in direct sunlight. When it was raining, scans were typically performed indoors with doors and windows closed to prevent water from entering such that space and light were limited. Enumerators highlighted the small spaces and low light conditions as key barriers to obtaining successful scans:

> *[What we were taught was that] the result will not come accurate if the child is not positioned well, [but there was] not enough space to put this child in a good position. That is where the difference is coming from.*

Limited testing was performed under these conditions; however, developers believed that neither should have affected scan performance:

> *You really only have to be 4 feet away from the child...When you take a picture, how far back do you normally stand? You know, I would say at least six feet. Probably more when you're taking a picture of your friends. And so [the appropriate distance for the scans] might just not be where it is natural to stand.*
>
> *[Poorly lit households] should not affect scan quality. In fact, it should improve scan quality, in as much as you get rid of direct sunlight because then you can rely purely on the structured light. That should be good. [The AutoAnthro technology] has two ways of measuring distance: through this structured light by putting like a pattern on the kid and knowing where those [infrared] light dots are or [by geometry using differences in angles from] two cameras simultaneously on the object. In really bright light, you can only use the dual cameras, and you can't rely so much on the structured light.*

An additional challenge noted by enumerators was that the phone frequently overheated, shut down, and gave invalid readings, an issue that was not observed in previous studies or during stress testing. However, stress testing in advance of the survey was limited to 2 hours, whereas field work lasted >8 hours per day. Key informants noted that the maximum operating temperature of the 3D scanner is 95 °F (35 °C) and that large swings in temperature can affect the trigonometry used in the scanner to assess distance (eg, affect calibration of distances between the camera and infrared light emission). However, they were uncertain regarding whether operating the devices 5 °F above the maximum operating temperature would be meaningful:

> *On the third day [of data collection] the device began to be hot. I reported those challenges to [the study supervisor]. The device is showing that it locked itself. When it failed, we know there was a problem. When the device failed it may bring you "00" or may give [a measurement of] 30 [cm for] MUAC...When it failed, you could click save and select end session. When the device gave you the "00," I selected end session, [restarted] and then scanned the child again. To me it happened many times. In the third and fourth days [of data collection], maybe 3 times* [each day]

Scans used for analysis were automatically processed but manually reviewed for the purpose of understanding the source of errors. The 2 most common issues identified were scans captured with the feet of the child obscured (affecting height and length measurements) and with the enumerator too close to the child such that the algorithm mistook the enumerator's arm for the child's (affecting MUAC measurements). Unusual light conditions were also noted to have affected scans but were observed less frequently:

> *The feet were almost always obscured by the enumerators hand or arms. So, there were some guesswork there (I didn't touch the feet manually, but algorithmically there's going to be a lot of uncertainty as to where the feet are). In that case, you know, I think we could have written the instructions differently, meaning like just hold the child by the calves or ankles not the balls of the feet. For the older children who were standing, I saw so many cases where you couldn't see the feet at all. They were just too close, or they angled the phone incorrectly. In this example [scan from training shown] there's plenty of space between the outline of the feet in the bottom. And yet there were so many instances where you couldn't even see the child's ankles. Sometimes the hands could get cut off as well but that's not a problem because [the algorithm] just looks for the elbows. It was rare that the...head was not captured. That happened a couple of times. It was pretty rare. The procedural problem that we saw far more often is that the enumerators arms were really close to the child's arms and that would throw a joint (e.g., move it from the child elbow to the enumerators elbow).*
>
> *On the iOS device, aiming screen, you were clearly moving a cube through 3-dimensional space—that is what it looked like on the screen—and you're trying to put the kid in that cube. You had a very clear representation of what is in the cube and what is out of the cube. Your goal is to put the kid entirely in that cube. On the Android system it is much more like aiming a camera. It feels more 2-dimensional. We ended up adding a body outline which was useful, but kids were still, I think, cut off. Where with the iOS system your cut off was not the edge of the screen. You're cut off was like this aiming box within the screen. And because you're using an iPad you have a bigger kind of field of vision [compared to the Galaxy phones] is what it feels like when you're aiming it, using both hands.*

Finally, enumerators noted that refusals by caregivers were very uncommon. Despite cultural sensitivities regarding capturing photographs of young children, particularly naked young children, they were typically able to reassure caregivers, showing them that 3D models (not pictures) were captured and, ultimately, most caregivers consented. However, children would sometimes cry and have tantrums and, based on this, the caregiver would withdraw consent. According to the enumerators, this was commonly observed when trying to capture scans for children aged <2 years whom they needed to lie flat on their back with arms extended, ideally separated from their caregiver.

Data Transmission

The transmission of the data was performed by the study coordinator at the end of data collection; daily upload by enumerators was not feasible given connectivity. A total of 198 scans were lost during the transmission process. Developers have been unable to replicate the error observed during transmission such that the source of the error occurred has not been determined. On the basis of the metadata retained, developers believe it is more likely that the scans were not captured (eg, the children were positioned, but scan acquisition was not successfully initiated) than that scan transmission failed:

> *The data transmission is still, to my mind, somewhat of a mystery. I don't understand why we could essentially run Skype, so running video back and forth between them and our servers but we could never consistently get the data to automatically upload. We went through all kind of hoops and work arounds trying to get to make sure that we actually had all of the data. And to this day, I don't know why our software didn't work cleanly on data uploads. In other places it has worked. [In South Sudan] it works very inconsistently.*
>
> *I do not believe that the data exists for a lot of those sessions. Because I could not replicate what we observed in South Sudan, where we have...a hundred children on the server, but there's only data for 20. And as far as I can tell, the only way that can happen is if you enter the child's information to create new session, but don't actually acquire the data for it.*

Discussion

Principal Findings

This study evaluated the accuracy of scan-derived anthropometric measurements among children aged 6 to 59 months calculated using the third-generation AutoAnthro technology. This version of the AutoAnthro system aimed to optimize 3D imaging technology for adoption at scale in nonresearch settings, including austere contexts such as rural South Sudan. In contrast to previous pilots in controlled settings but consistent with other effectiveness evaluations, the quality of scan-derived measurements was substantially poorer than that of manual anthropometry [7,8]. Measurement derived from scans in our study were biologically implausible for more than 1 in 20 children applying the cutoffs recommended by the WHO [11]. Although the mean differences between measurements for both height/length and MUAC were within 1 cm, only 1 in 5 measurements of stature and 1 in 3 measurements of MUAC were within 1 cm of the manual measurement. The half width of the 95% LoAs observed was >5 times wider for MUAC and nearly 20 times wider for height/length than that observed in the previous efficacy study [7]. In addition, the magnitude of the error was associated with the size of the child such that larger errors were observed among taller children and those with greater arm circumference. The magnitude of the differences observed translated into poor classification of malnutrition; most children with wasting and stunting would not have been identified for referral using the current version of the AutoAnthro technology.

The context of the COVID-19 pandemic as well as the low-resource setting of South Sudan served to highlight logistic challenges not previously identified with the use of the 3D imaging technology and likely contributed to the low accuracy observed in our study. Successful scans were processed for only 4 in 10 children included in the study as a result of higher refusal rates, poorer scan quality, and a large number of scans unsuccessfully transmitted. Although high refusal rates have been reported in previous studies of 3D scan technologies, the magnitude of the problem in South Sudan is distinct [5,8]. COVID-19 pandemic–related social distancing orders limited the number of children, particularly younger children, measured in a supine position available to use for validating the device algorithm before the initiation of field pilots and data collection. In addition, global COVID-19 travel restrictions prevented BST developers from conducting in-person trainings or providing real-time feedback to IMC IT staff or enumerators. Both of these factors were seen as critical barriers to the success of the AutoAnthro evaluation from the perspective of the software developers and the users.

Although quantitative analysis documented accuracy too poor to support the widespread adoption of the AutoAnthro software at present, key informant interviews provided insights into investments that may improve scan capture and processing. With respect to the software platform, further enhancements are needed to ensure that scans can be transmitted successfully on low-bandwidth networks and that scans captured in extreme light conditions (direct sunlight or very low light) can be processed without issue. To support a transition to full automation, the ability of enumerators and field supervisors to review scans and metadata was more restricted than in previous versions of the technology evaluated in the studies by Conkle et al [7] and Bougma et al [8]. Although automation will be essential for adoption at scale, revisiting what information remains available to enumerators and field supervisors locally on the device may be key to ensuring that field teams can aid in guaranteeing that scans are well captured and successfully saved. In addition, empowering teams to indicate which scans should be retained for analysis may further serve to address the barriers in scan quality identified. The teams used real-time, scan-derived height/length estimates to inform whether they should collect additional scans. The analysis approach, determined a priori, used the median of all scan values without input from enumerators on field challenges. Adaptations to the algorithm that allow teams to indicate scans that should be discarded may serve to improve accuracy.

In addition, further improvements in training materials are needed to ensure a more optimal implementation without direct support from the BST team; this would ultimately be needed to allow for use at scale. Updates to training protocols and materials would benefit from translation to local languages and more illustrations to support non–English-speaking and low-literacy enumerators. Ensuring adequate time for practice is also essential. Consistent with previous research, we identified a need for further guidance on scan capture and positioning in field conditions experienced (eg, low light, small spaces, and direct sunlight) [18]. Where scans were captured with the child’s head, feet, and arms clearly in the frame and not obscured by the enumerator or caregiver, the quality was acceptable. Both the availability and accuracy of scans were strongly associated with the enumeration team, notably more so than any characteristic of the child measured. Although the study was not designed to isolate the contribution of software, hardware, and the user to device accuracy, large differences in accuracy by team help illustrate how data acquisition (eg, positioning of the child relative to the scanner and caregiver, control of lighting, adaptation to space constraints, and other environmental factors) can affect scan-derived measurements. To the extent that user variation can be controlled with additional improvements in training, this may present an opportunity for future performance improvements.

This study is subject to at least six limitations. First, scans for over one-third of the sampled children were not successfully transmitted to the cloud server and could not be recovered from the devices. Although successful transmission of scans was not associated with child demographic characteristics or nutritional status, the loss of data resulted in a smaller sample size and limited power for planned analysis. Second, to ensure that the manual and scan-derived measurements were correctly matched, the child's age, sex, and weight were entered into both data sets as well as a child identification number. However, for many children with matched identification numbers, these other values were not a perfect match, prompting concerns about whether both measurements were truly from the same child. Third, manual measurements were used as the standard for evaluating the scan-derived measurements; however, there is some indication of terminal digit preference score, and the SD of WHZ and HAZ values exceeded 1.1, an indication of potential measurement error [11,19]. Fourth, given the humanitarian context, repeat manual or scan-derived measurements were not collected. As a result, we were unable to evaluate the reliability of these measurements. Fifth, information on the number of eligible children measured per household was recorded on paper forms. The forms for 5 clusters were damaged in a heavy rainstorm such that the total number of refusals in these clusters is unknown. Finally, the study sample is unique in that the population sampled was from a single internally displaced person site, and data collection occurred during the COVID-19 pandemic, both factors that may affect the generalizability of the findings to other populations and periods.

Conclusions

This study was initiated given considerable interest in 3D imaging technology, the potential use of the lightweight hardware, strong user acceptability, and evidence supporting the potential time savings relative to manual anthropometry [20]. Previous studies in controlled settings provided evidence that repeated scans could reliably estimate height/length and MUAC, suggesting the potential of 3D imaging technology as an alternative to manual measurement [7]. This study aimed to evaluate whether these results could be replicated at scale with full automation of scan processing and minimal oversight of training and data collection. Enumerators communicated an overall interest in the device performing well given that the scan capture generally took less time than manual measurement and eased field work. However, our findings suggest that the scan-derived measurements produced by AutoAnthro were not of sufficient accuracy for widespread adoption. Developers generally concluded that they needed more time to test and improve training and software; pandemic and financial barriers prevented them from ensuring that the software worked as intended before final testing. Further investments in the software algorithms are needed to address issues with scan capture and transmission—to ensure that scans can be captured in difficult field contexts (eg, extreme light conditions and temperature conditions) and efficiently transmitted on low-bandwidth networks. In addition, software revisions aimed at empowering field enumerators and supervisors were proposed, including local retention of data to facilitate field review of scan capture completeness and quality. Finally, differences in accuracy by team provide evidence that investments in training may also be able to improve performance.

Acknowledgments

The authors are immensely grateful for the efforts of the International Medical Corps team, which provided field supervision and support for data collection activities, including Ibrahim Maalim Bashir, Dugsiye Ahmed Sirad, and Logworong Alex Duku. Although blinded to manual measurements, the involvement of Gene Alexander, Harris Bergman, and Peter Miller from Body Surface Translations Inc was essential to articulating hardware and software changes being evaluated, processing scan-derived measurements, and providing insights into early piloting and training that help contextualize findings. AutoAnthro: 3D Scanning for Improved Malnutrition Assessment in Conflict Areas was supported by Creating Hope in Conflict: a Humanitarian Grand Challenge, a partnership of the US Agency for International Development, the UK Government, the Ministry of Foreign Affairs of the Netherlands, and Global Affairs Canada with support from Grand Challenges Canada. The recipient of the award was Body Surface Translations Inc, the developer of the AutoAnthro software. The US Agency for International Development Bureau for Humanitarian Assistance provided separate financial assistance for the 2-stage cluster survey within which this study was nested. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflicts of Interest

The authors declare no conflicts of interest. The paper was prepared by the coauthors without engagement from Body Surface Translations Inc, and the authors affirm that reporting on software performance is objective and independent.

Multimedia Appendix 1
Additional tables and figures with data related to scan availability and accuracy.
[DOCX File , 2391 KB - biomedeng_v7i2e40066_app1.docx]

References

1. Yorkin M, Spaccarotella K, Martin-Biggers J, Quick V, Byrd-Bredbenner C. Accuracy and consistency of weights provided by home bathroom scales. BMC Public Health 2013 Dec 17;13:1194. [doi: 10.1186/1471-2458-13-1194] [Medline: 24341761]
2. Patents by Inventor Edward G. Pryor. Justia. URL: https://patents.justia.com/inventor/edward-g-pryor [accessed 2022-09-28]
3. Severe acute malnutrition (SAM) photo diagnosis app® project. Action Against Hunger. 2015. URL: https://knowledgeagainsthunger.org/research/prevention/severe-acute-malnutrition-sam-photo-diagnosis-app-project/ [accessed 2022-09-28]
4. Methods for Extremely Rapid Observation of Nutritional Status (MERON). Kimetrica. URL: https://tinyurl.com/5a3tmeuj [accessed 2022-09-28]
5. Medialdea L, Bazaco C, D'Angelo Del Campo MD, Sierra-Martínez C, González-José R, Vargas A, et al. Describing the children's body shape by means of Geometric Morphometric techniques. Am J Phys Anthropol 2019 Apr;168(4):651-664. [doi: 10.1002/ajpa.23779] [Medline: 30629739]
6. Medialdea L, Bogin B, Thiam M, Vargas A, Marrodán MD, Dossou NI. Severe acute malnutrition morphological patterns in children under five. Sci Rep 2021 Feb 19;11(1):4237 [FREE Full text] [doi: 10.1038/s41598-021-82727-x] [Medline: 33608567]
7. Conkle J, Suchdev PS, Alexander E, Flores-Ayala R, Ramakrishnan U, Martorell R. Accuracy and reliability of a low-cost, handheld 3D imaging system for child anthropometry. PLoS One 2018 Oct 24;13(10):e0205320 [FREE Full text] [doi: 10.1371/journal.pone.0205320] [Medline: 30356325]
8. Bougma K, Mei Z, Palmieri M, Onyango D, Liu J, Mesarina K, et al. Accuracy of a handheld 3D imaging system for child anthropometric measurements in population-based household surveys and surveillance platforms: an effectiveness validation study in Guatemala, Kenya, and China. Am J Clin Nutr 2022 Jul 06;116(1):97-110. [doi: 10.1093/ajcn/nqac064] [Medline: 35285874]
9. Sampling methods and sample size calculation for the SMART methodology. Standardized Monitoring and Assessment of Relief and Transitions (SMART). 2012. URL: http://smartmethodology.org/survey-planning-tools/smart-methodology/ [accessed 2022-09-28]
10. South Sudan — Malakal PoC Brief (September 2021). International Organization for Migration. 2021 Sep 15. URL: https://dtm.iom.int/reports/south-sudan-%E2%80%94-malakal-poc-brief-september-2021 [accessed 2022-09-28]
11. de Onis M, Onyango AW, Van den Broeck J, Chumlea WC, Martorell R. Measurement and standardization protocols for anthropometry used in the construction of a new international growth reference. Food Nutr Bull 2004 Mar;25(1 Suppl):S27-S36. [doi: 10.1177/15648265040251S104] [Medline: 15069917]
12. World Health Organization, United Nations International Children's Emergency Fund. Recommendations for data collection, analysis and reporting on anthropometric indicators in children under 5 years old. World Health Organization. 2019. URL: https://apps.who.int/nutrition/publications/anthropometry-data-quality-report/en/index.html [accessed 2022-09-28]
13. Hense HW, Koivisto AM, Kuulasmaa K, Zaborskis A, Kupsc W, Tuomilehto J. Assessment of blood pressure measurement quality in the baseline surveys of the WHO MONICA project. J Hum Hypertens 1995 Dec;9(12):935-946. [Medline: 8746637]
14. WHO child growth standards: length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: methods and development. World Health Organization. Geneva, Switzerland: World Health Organization; 2006 Nov 11. URL: https://www.who.int/publications/i/item/924154693X [accessed 2022-09-28]
15. Bland JM, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986 Feb 08;1(8476):307-310. [Medline: 2868172]
16. Bartlett JW, Frost C. Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. Ultrasound Obstet Gynecol 2008 Apr;31(4):466-475 [FREE Full text] [doi: 10.1002/uog.5256] [Medline: 18306169]
17. Ulijaszek SJ, Kerr DA. Anthropometric measurement error and the assessment of nutritional status. Br J Nutr 1999 Sep;82(3):165-177. [doi: 10.1017/s0007114599001348] [Medline: 10655963]
18. Jefferds ME, Mei Z, Palmieri M, Mesarina K, Onyango D, Mwando R, et al. Acceptability and experiences with the use of 3D scans to measure anthropometry of young children in surveys and surveillance systems from the perspective of field teams and caregivers. Curr Dev Nutr 2022 Jun;6(6):nzac085. [doi: 10.1093/cdn/nzac085] [Medline: 35755937]
19. Standardized Monitoring and Assessment for Relief and Transitions (SMART) Manual 2.0. Standardized Monitoring and Assessment for Relief and Transitions, Action Against Hunger. Toronto, Canada: Canada and Technical Advisory Group; 2017. URL: https://smartmethodology.org/wp-content/uploads/2018/02/SMART-Manual-2.0_Final_January-9th-2017-for-merge-3.pdf [accessed 2022-09-28]
20. Conkle J, Keirsey K, Hughes A, Breiman J, Ramakrishnan U, Suchdev PS, et al. A collaborative, mixed-methods evaluation of a low-cost, handheld 3D imaging system for child anthropometry. Matern Child Nutr 2019 Apr;15(2):e12686 [FREE Full text] [doi: 10.1111/mcn.12686] [Medline: 30194911]

Abbreviations

BST: Body Surface Translations Inc

HAZ: height-for-age or length-for-age z score
IMC: International Medical Corps
LoA: limit of agreement
MUAC: mid–upper arm circumference
TEM: technical error of measurement
WHO: World Health Organization
WHZ: weight-for-height or weight-for-length z score

Edited by A Mavragani; submitted 05.06.22; peer-reviewed by M Jefferds, Z Reis; comments to author 10.08.22; revised version received 26.08.22; accepted 12.09.22; published 21.10.22.

Please cite as:
Leidman E, Jatoi MA, Bollemeijer I, Majer J, Doocy S
Accuracy of Fully Automated 3D Imaging System for Child Anthropometry in a Low-Resource Setting: Effectiveness Evaluation in Malakal, South Sudan
JMIR Biomed Eng 2022;7(2):e40066
URL: https://biomedeng.jmir.org/2022/2/e40066
doi:10.2196/40066
PMID:

Original Paper

Transforming Rapid Diagnostic Tests for Precision Public Health: Open Guidelines for Manufacturers and Users

Peter Lubell-Doughtie[1], MSc; Shiven Bhatt[1], MSc; Roger Wong[1], BSc; Anuraj H Shankar[2,3], DSc

[1]Ona Systems Inc, Burlington, VT, United States

[2]Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom

[3]Eijkman-Oxford Clinical Research Unit, Jakarta, Indonesia

Corresponding Author:
Anuraj H Shankar, DSc
Centre for Tropical Medicine and Global Health
Nuffield Department of Medicine
University of Oxford
New Richards Building, Old Road Campus, Roosevelt Drive
Oxford, OX3 7LG
United Kingdom
Phone: 44 1865 612 900
Email: anuraj.shankar@ndm.ox.ac.uk

Abstract

Background: Precision public health (PPH) can maximize impact by targeting surveillance and interventions by temporal, spatial, and epidemiological characteristics. Although rapid diagnostic tests (RDTs) have enabled ubiquitous point-of-care testing in low-resource settings, their impact has been less than anticipated, owing in part to lack of features to streamline data capture and analysis.

Objective: We aimed to transform the RDT into a tool for PPH by defining information and data axioms and an information utilization index (IUI); identifying design features to maximize the IUI; and producing open guidelines (OGs) for modular RDT features that enable links with digital health tools to create an RDT-OG system.

Methods: We reviewed published papers and conducted a survey with experts or users of RDTs in the sectors of technology, manufacturing, and deployment to define features and axioms for information utilization. We developed an IUI, ranging from 0% to 100%, and calculated this index for 33 World Health Organization–prequalified RDTs. RDT-OG specifications were developed to maximize the IUI; the feasibility and specifications were assessed through developing malaria and COVID-19 RDTs based on OGs for use in Kenya and Indonesia.

Results: The survey respondents (n=33) included 16 researchers, 7 technologists, 3 manufacturers, 2 doctors or nurses, and 5 other users. They were most concerned about the proper use of RDTs (30/33, 91%), their interpretation (28/33, 85%), and reliability (26/33, 79%), and were confident that smartphone-based RDT readers could address some reliability concerns (28/33, 85%), and that readers were more important for complex or multiplex RDTs (33/33, 100%). The IUI of prequalified RDTs ranged from 13% to 75% (median 33%). In contrast, the IUI for an RDT-OG prototype was 91%. The RDT open guideline system that was developed was shown to be feasible by (1) creating a reference RDT-OG prototype; (2) implementing its features and capabilities on a smartphone RDT reader, cloud information system, and Fast Healthcare Interoperability Resources; and (3) analyzing the potential public health impact of RDT-OG integration with laboratory, surveillance, and vital statistics systems.

Conclusions: Policy makers and manufacturers can define, adopt, and synergize with RDT-OGs and digital health initiatives. The RDT-OG approach could enable real-time diagnostic and epidemiological monitoring with adaptive interventions to facilitate control or elimination of current and emerging diseases through PPH.

(JMIR Biomed Eng 2022;7(2):e26800) doi:10.2196/26800

KEYWORDS

rapid diagnostic test; precision public health; digital health; diagnostic; testing; guideline; manufacture; surveillance; FHIR; Fast Healthcare Interoperability Resources

Introduction

Background

Rapid diagnostic tests (RDTs), specifically immunochromatographic lateral flow assays, can provide accurate real-time point-of-care diagnoses in low-resource settings and have been an important tool in the global health arsenal. Advances in microfluidics have enabled the medical device community to design smaller RDTs—as small as half the area of a business card and the thickness of a watch—that can be used to diagnose more conditions and are low cost—less than US $1 per device [1]. Manufacturers and international health organizations have collaborated to deliver hundreds of millions of RDTs to countries and communities [2]. However, the full potential of RDTs as a public health tool has not yet been realized. This is due to field deployment challenges amplified by a fragmented market, nonstandard designs, and lack of features that facilitate systematic capture and use of RDT results and patient data. Fortunately, current technology—computer vision, widely deployed smartphones, and mobile networks—applied to RDTs provides a scalable path to improved health and well-being through the emerging paradigm of precision public health (PPH). PPH is a field that aims to maximize impact with the active use of data for surveillance and targeted interventions by temporal, spatial, and epidemiological characteristics of populations.

Therefore, we aimed to define axioms to underpin steps to incorporate RDTs into a PPH approach, identify an initial set of features (RDT's hardware and software features) that would be needed to implement these axioms, and select a final evidence-based set of features that could be used by health policy and program implementers in integrating RDTs as tools for frontline health care workers at the community and clinic levels [3,4].

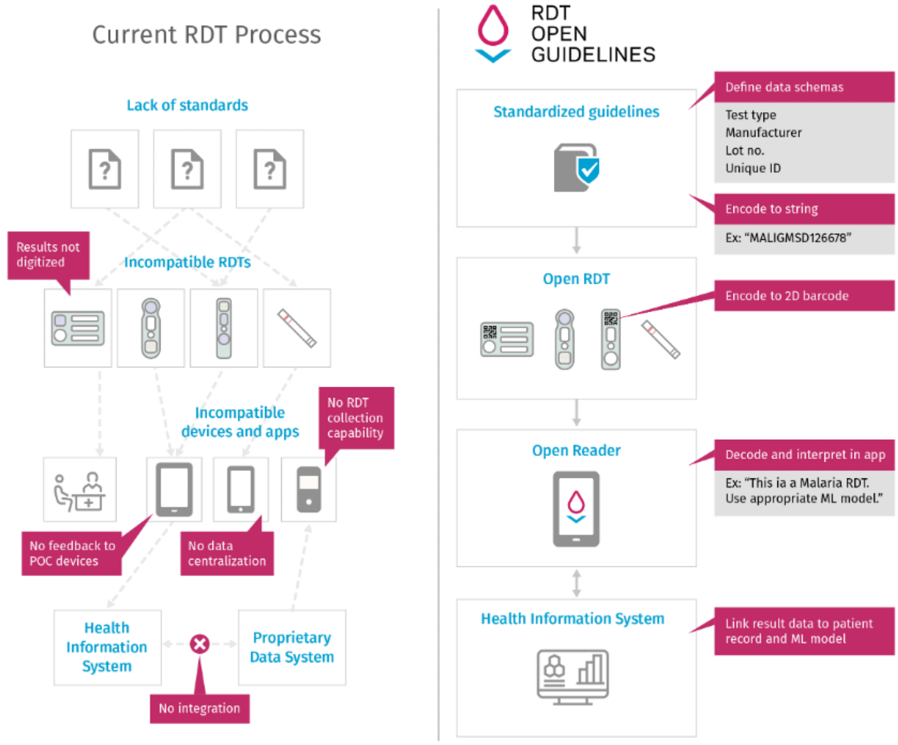
Challenges Facing the Current RDT Ecosystem

There are 3 core challenges faced within the current RDT ecosystem, which impede application and widespread implementation for PPH.

Lack of Data Standards

The lack of uniformity in RDT hardware and software substantially limits the integration of public health data from RDTs into current health information systems, thereby impeding their ability to respond to emerging crises [5]. Under these conditions, bridging these limitations requires analytics-intensive tasks to convert, code, recode, and integrate data. Current RDT versions require specific knowledge and tools that are typically not compatible (Figure 1), leading to a combinatorial explosion of integration and data interoperability requirements. Without uniformity, health professionals and individual consumers using RDTs cannot benefit from integration with point-of-care smartphone apps to enable personal tailored care guided by modern machine learning techniques that can calculate the prior probabilities of having a condition from data on demography, the environment, and etiology.

Figure 1. Current rapid diagnostic test processes (left) use undefined or proprietary standards, which lead to multiple incompatible protocols (ie, incompatible apps and devices). The rapid diagnostic test open guideline process (right) uses standard guidelines to produce modular reference rapid diagnostic tests that are compatible and can be read by a well-defined protocol for devices and apps. ID: identifier code or number for each device; ML: machine learning; POC: point-of-care; RDT: rapid diagnostic test.



Heterogeneity of RDT Reader Hardware

One strategy to improve the uniformity, amount, and quality of information collected from RDTs is to use custom hardware readers and image capture and analysis devices (eg, the DekiReader [6], specialized microscopes, and device holders). However, these devices can be problematic owing to the expense, continuous supply, and maintenance required. As such, custom hardware is incompatible with large-scale deployments and the broad consumer use necessary for RDTs to cover high- and emerging-risk areas; therefore, such devices hinder the continuous stream of accurate diagnostic data needed to expose outbreaks of known diseases and predict the emergence of new diseases.

Diversity of RDT Form Factors and Instructions

Any diagnostics integrated with smartphones would still involve manual use of an RDT and interaction with patients. However, several studies have documented the challenges faced by health care workers in translating their competency to use one RDT to comparable competency with another (ie, those for similar diseases, from other manufacturers, or with revised procedures). The lack of consistency contributes to high error rates in RDT usage and interpretation and limits their impact [7].

Solving These Challenges With Open Guidelines

Based on these challenges, we define 3 axioms that underpin solutions (Table 1). To solve the challenges and maximize information usage for PPH, RDTs should adhere to a set of open guidelines (OGs), and use smartphone readers and data protocols that standardize the information—both horizontally between different manufacturers or providers and vertically between different steps in the RDT life cycle (Figure 1). These axioms can transform the current state of the RDT ecosystem into one that supports PPH. RDT-OGs address the data uniformity challenge of RDTs by standardizing the capture and use of data.

Table 1. Rapid diagnostic test open guideline axioms.

Axiom	Description
Axiom 1: Maximize rapid diagnostic test (RDT) data usage by capturing and structuring information for integration	To address the custom hardware challenge, we designed RDT open guidelines (OGs) in line with the existing realities of the rapid diagnostic test manufacturing world. Manufacturers focused on creating tests simple enough to be used in clinical, community, and household settings by minimally trained community health workers, and eventually clients themselves. This need for widespread use leads us to define RDT-OGs to satisfy and facilitate these needs.
Axiom 2: Any solution must rely on only readily available local resources	This allows implementers of RDT-OGs to create solutions accessible by locally available technology and capability, including those that use pre-existing devices in target communities, such as low-cost smartphones [8]. RDT-OGs address the problems caused by lack of physical device uniformity by designing for human and device interoperability.
Axiom 3: Diagnostic interfaces should remain uniform or compatible	This applies to RDT hardware, the software reading and interpreting RDTs, and the data schemas integrating with external systems. When using software-based RDT readers, following Axiom 3 leads us to link individual RDTs to uniform interactive guidance of users as they conduct a diagnostic test.

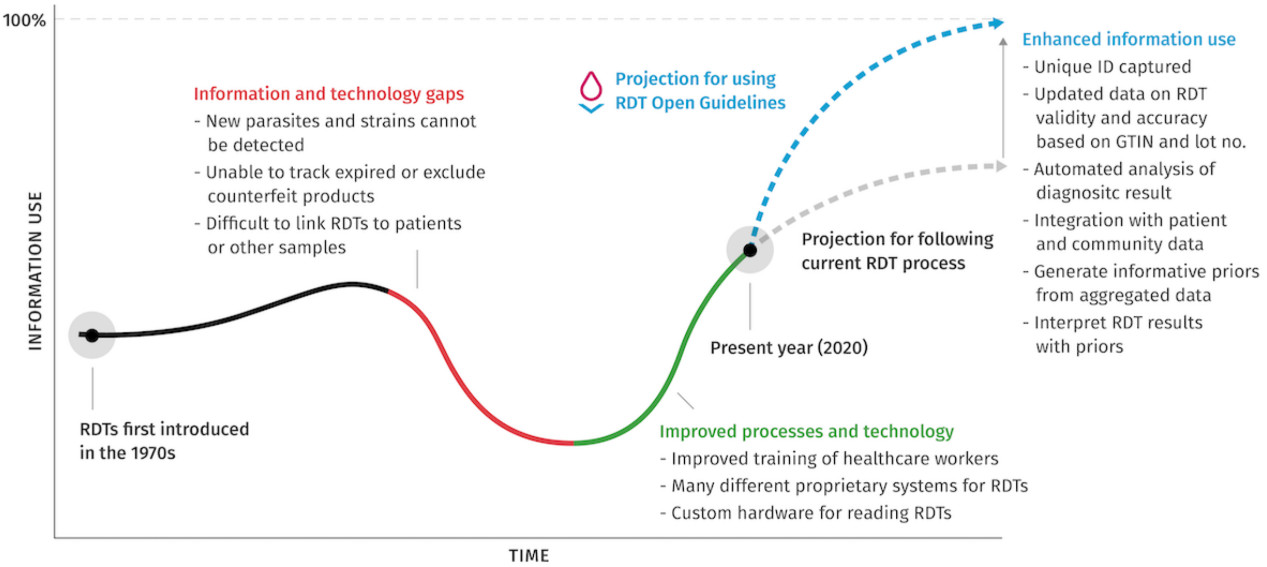
The Need to Catalyze the Era of RDT-OGs

We have chosen to address the current challenges in RDTs with open guidelines in order to focus directly on the systems and integration problems they face. Figure 2 shows initial progress as RDTs were developed in the laboratory; as researchers predicted their impact, optimism surrounding their potential grew. As RDT rollouts began, the ability to capture diagnostic data increased, but these increases did not keep pace with global growth in the technological capacity to store, communicate, and analyze information [9].

This changing technology landscape, combined with a lack of individual RDT identifiers, inconsistent test use protocols, and the appearance of fraudulent and counterfeit RDTs, led to a relative decrease in information use; however, as the RDT community began to effectively encode and aggregate information and improve the training of health care workers, information use increased. Currently, RDTs have reached a tipping point—there are multiple proprietary hardware and software solutions, and medical systems are facing “information overload” [10]. The future trend in information use could take 1 of 2 diverging paths—modest growth with eventual stagnation or a promising future with open guidelines aligned with the aforementioned PPH axioms to accelerate impact by enhancing information usage (Figure 2).

Below, we describe our methods, present survey results from experts in RDT technology, formally establish an information utilization index (IUI), and define how various RDT features gather information. We use this to assess World Health Organization (WHO)–prequalified RDTs in comparison to a prototype RDT based on open guidelines and then discuss these results and related work in field-based hardware and software diagnostics and standards. As access to telecommunications networks improves worldwide, and advanced information systems become more widely used by ministries of health and global health organizations, the community can dramatically accelerate the transformational impact promised by RDTs.

Figure 2. The trajectory of information use in response to rapid diagnostic test technology innovation, which shows the introduction of rapid diagnostic tests and their initial impact (black line) followed by subsequent challenges (red line) and improvements (green line). Potential future paths are also shown: a lower growth in information utilization under the current incremental improvements (gray dashed line) or an accelerating trajectory enabled with rapid diagnostic test open guidelines (blue dashed line). GTIN: global trade item number; RDT: rapid diagnostic test.


100%
INFORMATION USE
TIME
Information and technology gaps
- New parasites and strains cannot be detected
- Unable to track expired or exclude counterfeit products
- Difficult to link RDTs to patients or other samples
Projection for using RDT Open Guidelines
Enhanced information use
- Unique ID captured
- Updated data on RDT validity and accuracy based on GTIN and lot no.
- Automated analysis of diagnositc result
- Integration with patient and community data
- Generate informative priors from aggregated data
- Interpret RDT results with priors
Projection for following current RDT process
Present year (2020)
RDTs first introduced in the 1970s
Improved processes and technology
- Improved training of healthcare workers
- Many different proprietary systems for RDTs
- Custom hardware for reading RDTs

Methods

Review of Published Literature

To identify key issues related to data capture and use from RDTs, we conducted a review of published papers using the Semantic Scholar artificial intelligence–enabled research search engine [11]. We focused on, but did not constrain ourselves to, PubMed-indexed medical journal papers. The search was conducted on December 31, 2018, and updated on December 31, 2019, using the keywords “rdt,” “smartphone,” and “mobile phone.” The search revealed 480 papers that were further screened for the study of lateral flow immunochromatographic rapid tests, yielding 58 papers. In addition to reviewing these papers, we reviewed their citations to identify additional papers in telecytology, immunochromatography, and diagnostic hardware. From these papers, we extracted themes and concepts related to barriers to information capture and usage from RDTs, and applied a grounded theory conceptual framework to compile and code themes and concepts into core ideas and then high-level abstractions and classifications. These were discussed and reviewed by 3 different members of the team. We considered this process complete when saturation was reached (ie, no additional novel ideas or abstractions emerged upon review of additional papers).

RDT Stakeholder Survey

Procedure

The literature review and PPH axioms were used to identify the fundamental features and feasibility of open guidelines for RDTs that maximize information usage for PPH. A survey was designed, which comprised 30 questions (Multimedia Appendix 1). Respondent-driven sampling was used and initiated by contacting authors of the papers reviewed and professional referrals, which included researchers, medical technologists, manufacturers, medical professionals, and frontline health workers. These stakeholders were asked to participate in a web-based survey that comprised specific statements that corresponded to general, user-specific, manufacturer-based issues or issues regarding informatics. A 5-point Likert scale was used for response: 0=strongly disagree, 1=disagree, 2=neutral, 3=agree, 4=strongly agree, and unable to reply. Replies were accrued from January 2019 to March 2019, and submitted entries were downloaded and tabulated. Results were summarized by tabulations and analysis of proportions using Excel (version 16; Microsoft Inc).

Ethics Approval

We note that the survey was exempt from human subjects research as per guidelines from the US Department of Health and Human Services as it assessed a public benefit or service and was not about humans, and did not collect sensitive information.

Defining the IUI

Survey results and the literature review were used to identify essential information features for RDTs and their integration into health care platforms to support PPH. The presence or absence of a feature for a specific RDT could be used to calculate an IUI defined as *number of features present* or the *number of features defined*. We then selected WHO prequalified cassette-based RDTs for malaria and HIV that had been assessed for performance [12] and calculated the IUI.

Results

Literature Review

The review of published literature and thematic extraction of concepts related to information capture and usage from RDTs led us to identify the following core areas that affect information usage (Table 2).

Table 2. Core areas that affect information usage.

Core areas	Description
Challenges in using commonly deployed rapid diagnostic tests (RDTs).	This is related to issues of RDT choreography, and proper reading and interpretation even when control and results lines were clear.
Existing barriers for mobile imaging of RDTs.	This referred to shadows from the cassette on the surface of the immunochromatographic strip, or glare from the cassette and surface of the strip, all of which hindered image capture quality.
Criteria for designing RDT standards.	This included specific characteristics of RDTs that could be feasibly standardized.
Barriers to RDT manufacturing standards.	This referred to cost and other factors that could hinder manufacturing to an enhanced standard.
Feasibility and features for smartphone-read RDTs.	This included the practicality of using identified features in the clinical or field setting.
Perceptions of non–human-readable RDTs (eg, electrochemical readouts).	This included whether or not read-out systems for RDTs would be acceptable for clinical or field personnel, if the actual reaction was not observable.

RDT Stakeholder Survey

We contacted 81 stakeholders, and 33 completed the questionnaire (16 researchers, 7 technologists, 3 manufacturers, 2 doctors or nurses, and 5 others). Respondents were most concerned about the proper use of RDTs (agreed: 30/33, 91%), their interpretation (agreed: 28/33, 85%), and reliability (agreed: 26/33, 79%). Respondents were confident that smartphone-based RDT readers could address some reliability concerns (agreed: 28/33, 85%) and that readers were more important for complex or multiplex RDTs (agreed: 33/33, 100%).

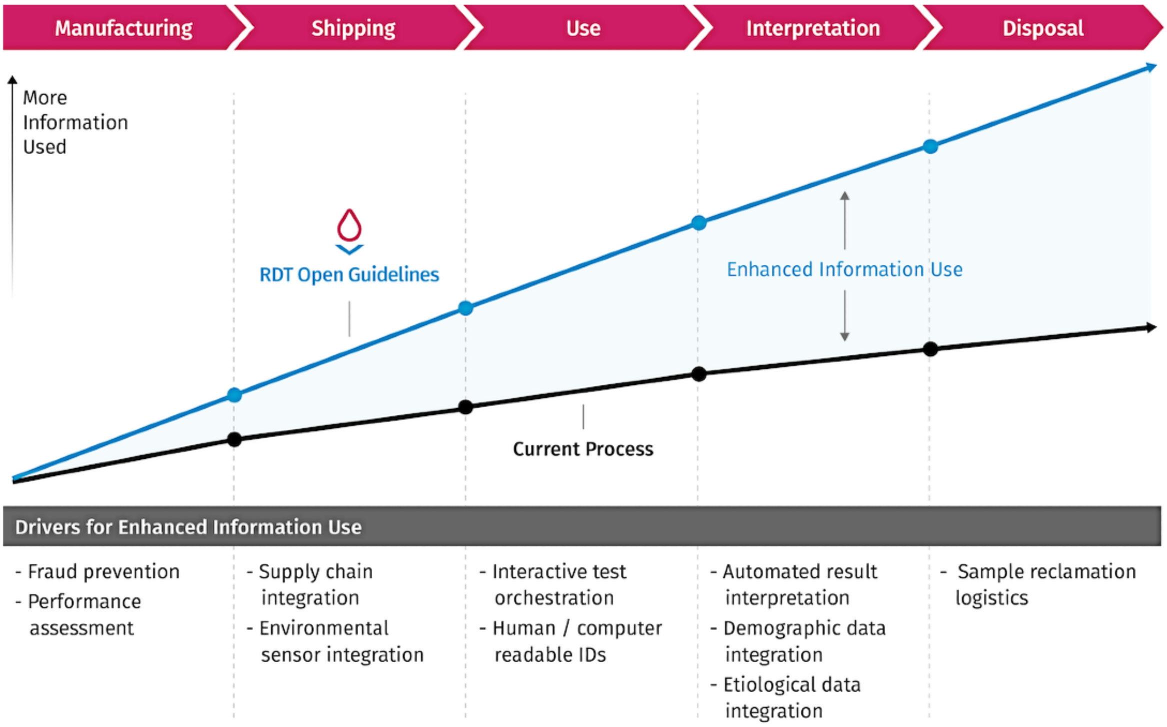
IUI

Based on these results, and because RDTs are embedded in a set of protocols and practices defined by health care workers, institutions, clients, and communities, proper usage of rapid diagnostic tests depends not only on a physical device but also on its integration into the larger ecosystem. In this context, to maximize information usage, an RDT platform must function effectively in all phases of its life cycle, with added value at each phase.

Specific stakeholder results are divided into discrete phases of the rapid diagnostic test life cycle (Manufacture, Shipping, Use, Interpretation, and Disposal), and RDT capabilities are divided into themes (Metadata, Molding of the Cassette, Printed Data, and Smartphone Reader) (Figure 3). The open guidelines contribute to each theme, and they are essential for the Smartphone Reader theme. A conceptual framework that contrasts the accumulated value of RDT open guidelines and the current RDT process shows an increase at each life cycle phase. Specific capabilities drive these increases (Figure 3).

We identified 11 essential information features for RDTs and their integration into health care platforms, which define components of the IUI (Textbox 1).

Figure 3. Conceptual framework of gaps in information use through the rapid diagnostic test life cycle. Information utilization (vertical axis) is quantified relative to 5 distinct phases of the rapid diagnostic test life cycle (horizontal axis): Manufacturing, Shipping, Use, Interpretation, and Disposal. Over the life cycle, the information utilization of the current process increases (black line), but with the use of rapid diagnostic test open guidelines, information utilization would increase (blue line) as a result of several features (list at bottom).


Manufacturing
Shipping
Use
Interpretation
Disposal
More Information Used
RDT Open Guidelines
Enhanced Information Use
Current Process
Drivers for Enhanced Information Use
- Fraud prevention
- Performance assessment
- Supply chain integration
- Environmental sensor integration
- Interactive test orchestration
- Human / computer readable IDs
- Automated result interpretation
- Demographic data integration
- Etiological data integration
- Sample reclamation logistics

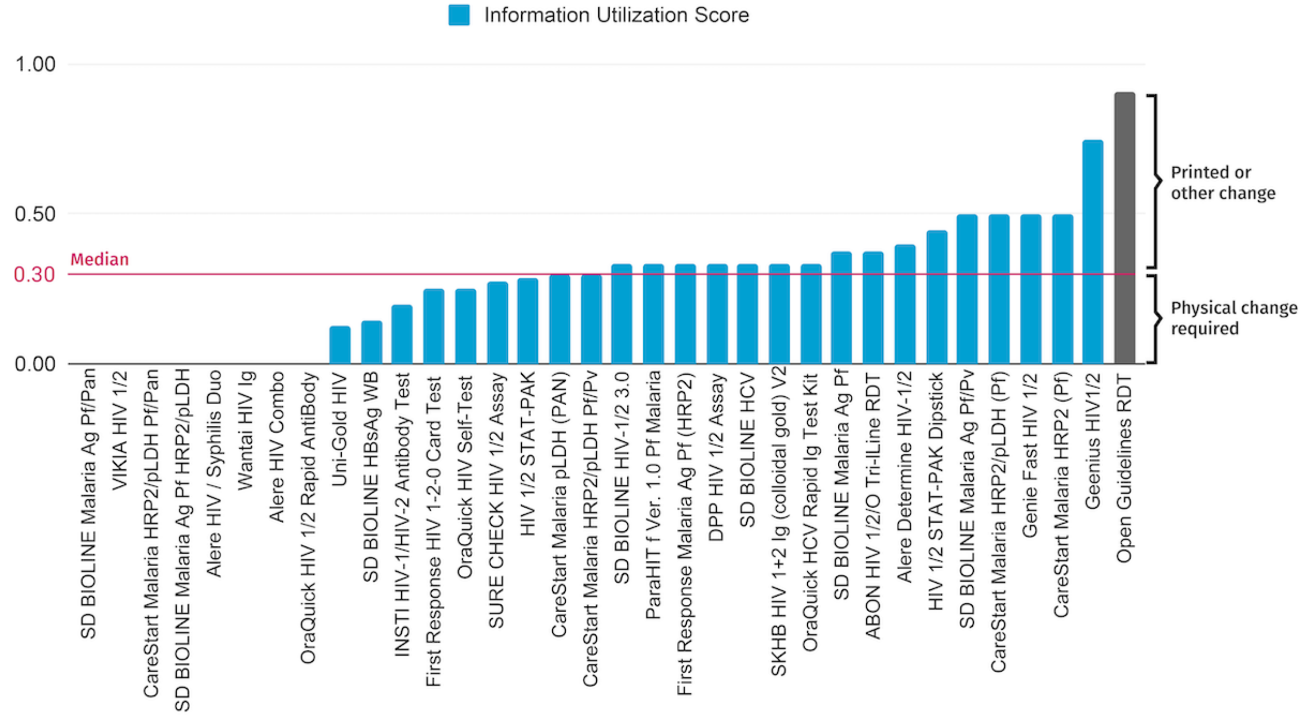
Textbox 1. Components of the information utilization index.

1. Smartphone or other device reader exists
2. Instructions included
3. Cassette is not reflective
4. Test strip is not reflective
5. Shadow does not exist on test window
6. Expiration date printed on device
7. Identifier printed on device
8. Color calibration panel on device
9. 2D barcode on device
10. Test name clearly printed on device
11. Regulator (eg, World Health Organization) approved for lab and field use

Assessment of Current Rapid Diagnostic Tests

The IUI—which provides an overview of how much information current diagnostics can capture and where there is room for improvement—for prequalified RDTs and an OG RDT had values ranging from 0 to 0.75 (mean 0.27; median 0.30, IQR 0.25) (Figure 4). The large bracket shows that 70% of this information usage score can be attributed to printed or other nonphysical changes, while the remaining 30% require physical changes to the RDT.

Figure 4. Information utilization index for WHO prequalified rapid diagnostic tests (RDTs). Scores were calculated for 33 WHO prequalified devices that had accessible information (blue, names listed below), as well as an RDT based on the RDT Open Guidelines (grey). The median information utilization score was 0.30 (magenta line), in contrast to 0.91, the score for an Open Guidelines RDT. 70% (top bracket) of the Open Guidelines RDT score can be attributed to non-physical changes, while the remaining 30% (bottom bracket) requires physical changes to the RDT.



Discussion

RDT Open Guidelines

Our RDT-OGs recommend the horizontal integration of RDT hardware through consistent physical modules, thereby enabling vertical integration of RDT software through consistent protocols linking supply chain, test choreography, and interpretation. In Figure 5, a reference example of RDT-OGs with colored overlays identifying the core modifications is shown. These include a 2D barcode to embed information needed for an app to identify, read, and interpret the RDT; fiducials as reference points to assist the camera and phone to quickly and accurately identify the RDT areas of interest and reference; and a color calibration panel to enable reliable colorimetric inference. In addition, 3 WHO-prequalified RDTs with overlays are shown to highlight current inconsistencies between tests (Figure 5).

In comparing the optimized IUI and design of the reference RDT-OGs to others, we observed both the heterogeneity and common structures across all RDTs. We have designed RDT-OGs to be useful whether adopting all recommendations, a subset of modules, or using existing cassettes linked to an RDT-OG–compatible software platform. Defining common data models and schemas provides an information architecture that would encapsulate data from any module combination that exists on the RDT. The RDT-OG data schema can be effectively encoded by the 2D barcode and easily drive the process forward via a reader app. The design and production aspects have proved feasible given the successful production of the prototype, and the field assessments, such as assessment of integration with epidemiological monitoring systems, are ongoing.

Creating a systematic way (Figure 6) to collect and aggregate structured RDT data allows the community to continuously monitor device performance, disease prevalence, and the relationship between demographic priors and diagnostic outcomes. This workflow, like the RDT-OG, is not tied to a particular diagnostic and is designed to accommodate both existing and emerging diagnostics. RDTs can increase their IUI by using a universal RDT-OG–compatible reader and data storage. New RDTs that become available in the market can further increase their IUI by using the hardware recommendations of the RDT-OGs (Figure 6). The workflow integrates automatic result interpretation modules using machine learning from image libraries or template-based approaches and can dynamically accommodate new RDTs through database-backed parameters defining rapid diagnostic test components and hyperparameters identifying the specific RDT (Figure 7). The rapidly growing number of COVID-19 serology and antigen-based RDTs show the critical role of dynamically supporting newly released RDTs [13,14].

To the best of our knowledge, this is the first paper to propose guidelines to harmonize the hardware, software, and data standard used to read and interpret RDTs.

Figure 5. Unifying rapid diagnostic test functionalities based on formal guidelines. A rapid diagnostic test based on the rapid diagnostic test open guidelines (left) should have certain functional components, as indicated by the color-coded overlay. In contrast, 3 RDTs currently on the World Health Organization–prequalified list have only some of these components (right, with color-coded overlays).

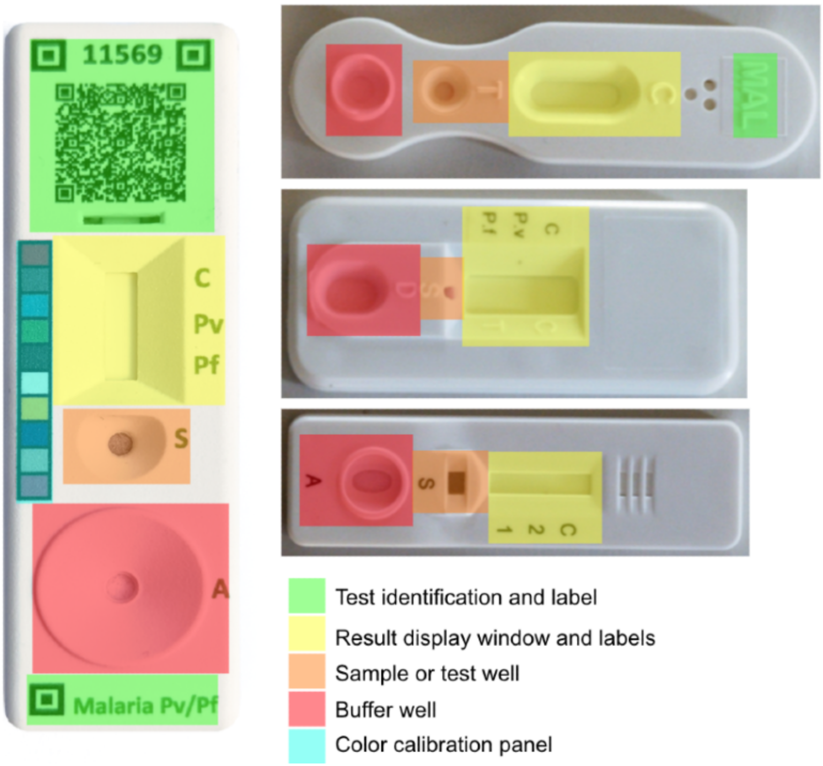

11569
C
Pv
Pf
S
A
Malaria Pv/Pf
T
C
MAL
C
Pv
Pf
S
D
T
A
S
C
2
1
🟩 Test identification and label
🟨 Result display window and labels
🟧 Sample or test well
🟥 Buffer well
🟦 Color calibration panel

Figure 6. In a system that incorporates rapid diagnostic test open guidelines, data are captured and digitized data from rapid diagnostic tests using a smartphone app that is compatible with the rapid diagnostic test open guidelines. These data are transmitted to a health information system platform and integrated with health system data, laboratory data, and other relevant data, then used to build machine learning models that both feed upstream, to smartphone apps to model symptoms and to be used to better interpret results, as well as downstream, for monitoring. Planners, managers, and researchers can use the real-time data to decide on modifications to existing programs and plan new programs. The color of the lines identifies the primary participant in that portion of the workflow, and the badges depict where to apply the features of the rapid diagnostic test open guideline. ML: machine learning; RDT: rapid diagnostic tests.

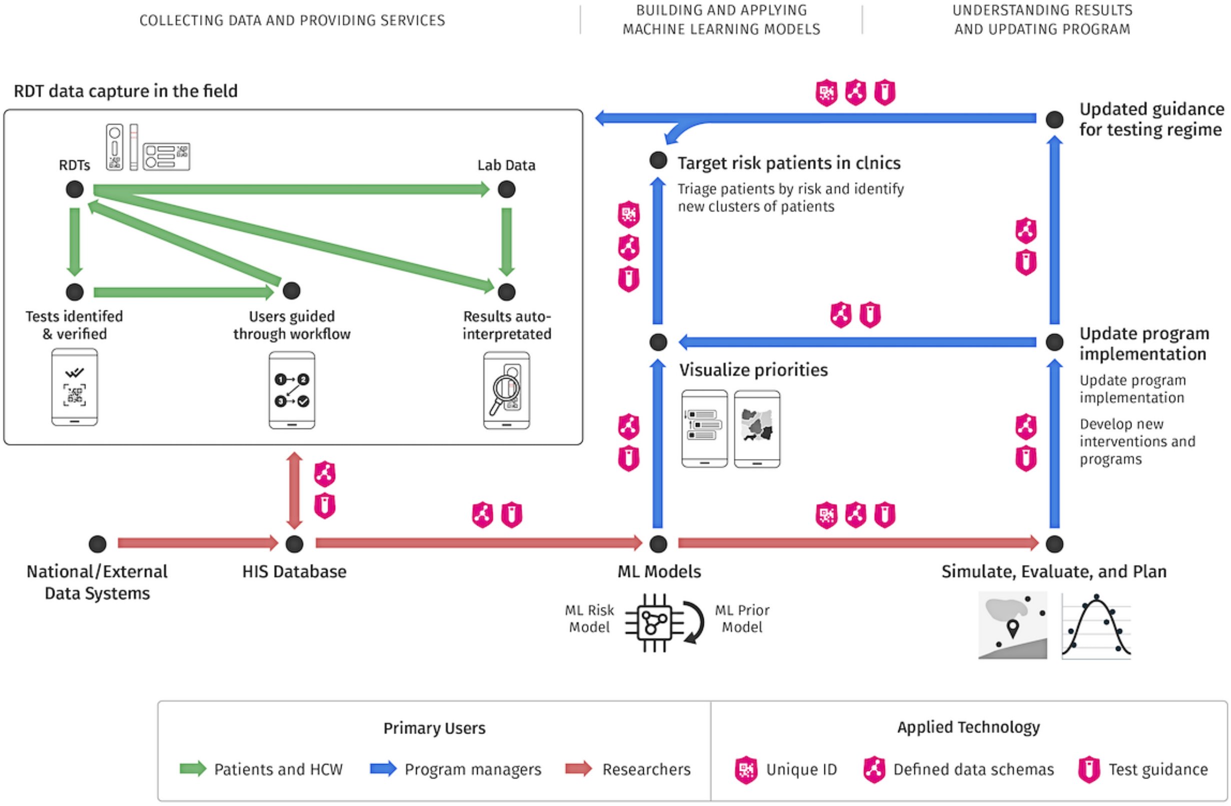
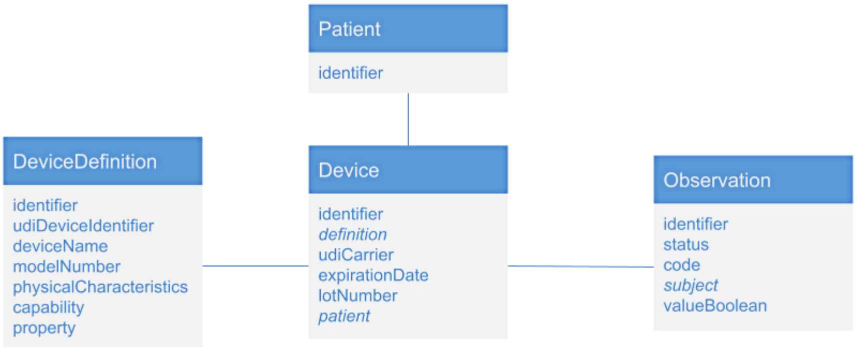

COLLECTING DATA AND PROVIDING SERVICES
BUILDING AND APPLYING MACHINE LEARNING MODELS
UNDERSTANDING RESULTS AND UPDATING PROGRAM
RDT data capture in the field
RDTs
Lab Data
Tests identifed & verified
Users guided through workflow
Results auto-interpretated
Updated guidance for testing regime
Target risk patients in clnics
Triage patients by risk and identify new clusters of patients
Update program implementation
Visualize priorities
Update program implementation
Develop new interventions and programs
National/External Data Systems
HIS Database
ML Models
ML Risk Model
ML Prior Model
Simulate, Evaluate, and Plan
Primary Users
Patients and HCW
Program managers
Researchers
Applied Technology
Unique ID
Defined data schemas
Test guidance

Figure 7. A Fast Healthcare Interoperability Resources–based workflow using the Device Definition resource for Open Guidelines based rapid diagnostic tests connected to Device, Observation, and Patient resources. Medical devices are defined using the DeviceDefinition resource (to specify their physical characteristics and links to external information systems). Each rapid diagnostic test used corresponds to a Device resource linked to the appropriate DeviceDefinition resource, as well as to Patient and Observation resources that store patient information and test results, respectively.



Diagnostics

Overview

It is useful to review the RDT-OG system with related hardware, software, and standard-based approaches to field diagnostics, and to note limitations and next steps to integrate the RDT-OGs into the digital health ecosystem.

Integrated Hardware-Based Field Diagnostics

Hardware-based field diagnostics require reusable equipment to function and connect to software systems. For example, the DekiReader is a portable device that guides users through a malaria RDT, reads, and automatically interprets test results; it is notable that its results are not significantly different from human readings [6,15]. Similarly, NutriPhone pairs a lateral flow cassette with a hardware device and app to guide users and process images of test results to measure vitamin B_{12} levels. It has not been tested at scale; however, in a sample of 12 participants, there was a correlation of 0.93 with the results from an immunoassay [16].

Software-Based Field Diagnostics

In contrast, software-based field diagnostics do not require additional hardware to function, while having accuracy comparable to human interpretations of RDTs or images of used tests [17]. Dell and Borriello [18] made use of the pre-existing Open Data Kit to read various cassette-based RDTs using only smartphones and 3D printable stands for consistent image capture. Similarly, Ozkan and Kayhan [19] developed an RDT holder that clips onto smartphones to improve image consistency and data interpretation. Demonstrating the utility beyond cassette-based RDT formats, Ra et al [20] combined a urine test strip with color calibration markers and a smartphone app to improve automated urinalysis accuracy across various lighting conditions. There are also proprietary RDT-reading platforms, including BBI Solutions' Novarum Smartphone Reader and Abingdon Health's AppDx Smartphone Reader, that integrate on-cassette QR codes but with limited information (such as RDT type) [21,22].

Though these approaches integrate modern software, their generalizability is limited by having been designed in the absence of guidelines that standardize their solutions, and therefore do not adhere to PPH axioms 2 and 3. Vashist et al [23] reviewed how smartphone-based health care apps and devices, including related medical, privacy, and data standards, remain fractured without guidelines or standards. A recent review [24] further extended the number of diagnostic devices and companies involved, and again concluded there is a lack of unification.

General Challenges in Field Diagnostics

Yager et al [25] describe the biomedical engineering community as historically focused on laboratory-based diagnostics and highlight the work needed to adapt tools for settings in low- and middle-income countries. Improvement in test instructions, health worker training, and performance monitoring all correlate with reduced preanalytical errors, improved test performance, and increased result reliability [7,26,27].

There is a growing consensus that point-of-care diagnostics and smartphones equipped with digital health solutions are converging and that this advancement may significantly expand self-managed care [28,29]. However, we currently lack digital health interventions with diagnostics linked to clinical care pathways and infectious disease surveillance systems [30], as well as solutions to the privacy and data stewardship challenges necessary for large-scale deployment [28]. Despite these challenges, researchers have outlined numerous promising future point-of-care linkages: handheld ultrasounds and software platforms with standardized databases integrating artificial intelligence [31], telecytology platforms for test-and-not-treat strategies [32], and accurate diagnoses of oral cancer using convolutional neural networks [33].

Related Regulations and Standards

Existing standards for diagnostics encompass RDTs, including several related standards in health technology, medical devices, and precision medicine. For example, given that most rapid diagnostic tests have maximal temperature limits for storage and use, temperature exposure monitors such as those used on vaccine vials would be warranted based on performance degradation from heat exposure [34]. In addition, use of the Fast Healthcare Interoperability Resources (FHIR) [35] can define clinical data as a graph of well-defined fields and data types (Figure 7). A number of current digital lab data platforms and application programming interfaces already handle related diagnostic and laboratory information management with FHIR as a common standard [36].

Regulators, such as the US Food and Drug Administration, define pathways to classify novel medical devices, including communication-enabled RDTs, for which there are no similar existing devices [37]. Similarly, the Medical Device Communications Testing Project from the National Institute of Standards and Technology is relevant to any form of medical device communication and applicable to RDTs which communicate via radio frequency or electrochemical means [38]. These regulatory bodies also promote innovation (eg, the National Institute of Standards and Technology Text Retrieval Conference, where annual precision medicine competitions model the most effective treatments) exemplifying how communities can benefit from well-structured data [39].

Limitations

This study has some limitations. First, the response rate from the survey was 40% (33/81), and the survey results may have benefited from additional feedback from a broader group. Nevertheless, there was strong thematic concordance between core responses from the survey and findings from the literature review. Second, after designing the open guidelines, we did not solicit additional feedback from the same and similar groups of persons who were contacted for the survey. This additional step would serve to validate the utility of the open guidelines. We note that our goal was to collect feedback from RDT producers and users of RDT-OGs. Third, we limited ourselves to information usage issues and solutions for cassette-based rapid tests and did not include the simpler dipstick type strips. However, the same concepts would apply and would need to be implemented in a way that is compatible with the lower space and cost profile of such tests [40]. Despite these caveats, the proposed RDT-OG approach is clearly applicable to the majority of RDTs currently deployed globally, and to those likely to be produced in the future as multiplex and complex tests become the norm.

Conclusions and Policy Recommendations

In response to the goals and ambitions of the RDT community, we defined PPH axioms and derived RDT-OGs. The recommended modular foundation is designed to accelerate current RDT development, fieldwork, and successfully translate RDTs into effective field evaluations and deployments at scale. These guidelines thus confer functionality to diagnostic devices, the smartphone apps interpreting them, and the health information system analyzing them. For example, temperature sensors may be essential to assure proper storage and quality of some rapid diagnostic tests [34], and the modularity of open guidelines can accommodate this need. Although modifying supply chains may be infeasible in areas with rigid logistics or fixed asset costs, the RDT-OGs gives the community a pathway to extend the functionality of pre-existing field-based diagnostics through advances in machine learning, which do not require RDT modifications.

National and global policy makers have shown a willingness and ability to convene communities around guidelines that benefit RDT stakeholders; for example, the WHO prequalification of medicines program, FHIR, SNOMED, and LOINC. As the WHO, the Global Fund, Foundation for Innovative New Diagnostics, and others continue this work, there is ample opportunity to adopt formal guidelines around RDTs and their usage. For example, the WHO's role in creating and promoting prequalified malaria RDTs has incentivized manufacturers to increase low-cost RDT production [41,42]. A similar approach to incentivize machine-readable RDT identifiers, and data schemas to interpret them, would likely address challenges currently faced.

Thus, by providing guidance for RDT hardware, software, and data interoperability, standards-setting organizations can transform RDTs into a formidable public health tool for disease prevention and treatment, in addition to diagnosis. These innovations can accelerate long-term disease control efforts, such as for malaria, which is responsible for 7.8% of the annual deaths of children under 5 years old (20,000 children worldwide [43]). Furthermore, these innovations can accelerate rapidly evolving disease control efforts, such as for COVID-19, where serological or antigen detection investigations face challenges in obtaining case information; these challenges are expected to further increase as testing efforts continue to scale up, and with transition from mitigation to containment [44]. Therefore, in both routine and emergency scenarios, adopting RDT-OGs would apply key advances in information technology to close the critical gap between diagnostics and public health interventions, and enable a new era of precision public health.

Acknowledgments

This research was supported by the Bill and Melinda Gates Foundation where Arunan Skandarajah provided helpful insights, discussion, and review of these ideas during his tenure. The technical and implementation teams at the Summit Institute for Development, the Kenya Medical Research Institute, and Ona assisted in the operationalization of this work, results of which will be published elsewhere; Craig Appl, Samuel Githengi, and Vincent Karuri provided insightful commentary on applying Fast Healthcare Interoperability Resources (FHIR) and software engineering techniques to this work. We thank Ted Prusik and Mohannad Abdo of Temptime Corp. and Zebra Technologies Corp. for helpful guidance and support.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Survey instrument.
[PDF File (Adobe PDF File), 96 KB - biomedeng_v7i2e26800_app1.pdf]

References

1. Tay A, Pavesi A, Yazdi SR, Lim CT, Warkiani ME. Advances in microfluidics in combating infectious diseases. Biotechnology Advances 2016 Jul;34(4):404-421. [doi: 10.1016/j.biotechadv.2016.02.002]
2. The Academy of Medical Sciences; 2016. Improving the Development and Deployment of Rapid Diagnostic Tests in LMICs. URL: https://acmedsci.ac.uk/file-download/21094033 [accessed 2020-12-31]
3. Mehl G, Labrique A. Prioritizing integrated mHealth strategies for universal health coverage. Science 2014 Sep 12;345(6202):1284-1287. [doi: 10.1126/science.1258926] [Medline: 25214614]
4. Chhetri A, Iversen M, Kaasbøll J, Kanjo C. Evaluating mHealth Apps Using Affordances: Case of CommCare Versus DHIS2 Tracker. In: Nielsen P, Kimaro HC. eds. Information and Communication Technologies for Development. Strengthening Southern-Driven Cooperation as a Catalyst for ICT4D. Vol 551. IFIP Advances in Information and Communication Technology. Springer International Publishing; 2019:619-632.
5. Dameff C, Clay B, Longhurst C. Personal Health Records: More Promising in the Smartphone Era? JAMA. JAMA 2019;321(4):339. [doi: 10.1001/jama.2018.20434] [Medline: 30633300]
6. Oyet C, Roh ME, Kiwanuka GN, Orikiriza P, Wade M, Parikh S, et al. Evaluation of the Deki Reader™, an automated RDT reader and data management device, in a household survey setting in low malaria endemic southwestern Uganda. Malar J 2017 Nov 07;16(1):449-449 [FREE Full text] [doi: 10.1186/s12936-017-2094-3] [Medline: 29115991]
7. Gillet P, Maltha J, Hermans V, Ravinetto R, Bruggeman C, Jacobs J. Malaria rapid diagnostic kits: quality of packaging, design and labelling of boxes and components and readability and accuracy of information inserts. Malar J 2011 Feb 13;10(1):39 [FREE Full text] [doi: 10.1186/1475-2875-10-39] [Medline: 21314992]
8. GSMA Intelligence; 2019. The Mobile Economy 2019. URL: https://www.gsmaintelligence.com/research/?file=b9a6e6202ee1d5f787cfebb95d3639c5&download [accessed 2020-12-31]
9. Hilbert M, López P. The world's technological capacity to store, communicate, and compute information. Science 2011 Apr 1;332(6025):60-65 [FREE Full text] [doi: 10.1126/science.1200970] [Medline: 21310967]

10. Kumar A, Maskara S. Coping up with the Information Overload in the Medical Profession. JBM 2015;03(11):124-127. [doi: 10.4236/jbm.2015.311016]
11. Fricke S. Semantic Scholar. jmla 2018 Jan 12;106(1):145-147. [doi: 10.5195/jmla.2018.280]
12. Malaria rapid diagnostic test performance. Results of WHO product testing of malaria RDTs: Round 7 (2015-2016). New Delhi: Asia Pacific Observatory on Health Systems and Policies, WHO Regional Office for South-East Asia; 2017.
13. Prazuck T, Colin M, Giachè S, Gubavu C, Seve A, Rzepecki V, et al. Evaluation of performance of two SARS-CoV-2 Rapid IgM-IgG combined antibody tests on capillary whole blood samples from the fingertip. PLoS One 2020;15(9):e0237694 [FREE Full text] [doi: 10.1371/journal.pone.0237694] [Medline: 32941461]
14. Agustina R, Syam AF, Wirawan F, Widyahening IS, Rahyussalim AJ, Yusra Y, et al. Integration of symptomatic, demographical and diet-related comorbidities data with SARS-CoV-2 antibody rapid diagnostic tests during epidemiological surveillance: a cross-sectional study in Jakarta, Indonesia. BMJ Open 2021 Aug 10;11(8):e047763 [FREE Full text] [doi: 10.1136/bmjopen-2020-047763] [Medline: 34376448]
15. Kalinga AK, Mwanziva C, Chiduo S, Mswanya C, Ishengoma DI, Francis F, et al. Comparison of visual and automated Deki Reader interpretation of malaria rapid diagnostic tests in rural Tanzanian military health facilities. Malar J 2018 May 29;17(1):214 [FREE Full text] [doi: 10.1186/s12936-018-2363-9] [Medline: 29843721]
16. Lee S, O'Dell D, Hohenstein J, Colt S, Mehta S, Erickson D. NutriPhone: a mobile platform for low-cost point-of-care quantification of vitamin B12 concentrations. Sci Rep 2016 Jun 15;6(1):28237 [FREE Full text] [doi: 10.1038/srep28237] [Medline: 27301282]
17. Poostchi M, Silamut K, Maude RJ, Jaeger S, Thoma G. Image analysis and machine learning for detecting malaria. Transl Res 2018 Apr;194:36-55 [FREE Full text] [doi: 10.1016/j.trsl.2017.12.004] [Medline: 29360430]
18. Dell N, Borriello G. Mobile tools for point-of-care diagnostics in the developing world. : Association for Computing Machinery; 2013 Presented at: ACM DEV '13; 2013; Bangalore, India. [doi: 10.1145/2442882.2442894]
19. Ozkan H, Kayhan OS. A Novel Automatic Rapid Diagnostic Test Reader Platform. Comput Math Methods Med 2016;2016:7498217-7498210 [FREE Full text] [doi: 10.1155/2016/7498217] [Medline: 27190549]
20. Ra M, Muhammad MS, Lim C, Han S, Jung C, Kim W. Smartphone-Based Point-of-Care Urinalysis Under Variable Illumination. IEEE J Transl Eng Health Med 2018;6:2800111 [FREE Full text] [doi: 10.1109/JTEHM.2017.2765631] [Medline: 29333352]
21. Polwart, Neil, Tyrie, Graham, Ashbrook, Anthony P. Testing apparatus for performing an assay and software for portable device. 2018. URL: https://patents.google.com/patent/PT2646809T/en [accessed 2020-12-31]
22. Abingdon H. Flexible Lateral Flow Smartphone Reader Customization. URL: https://www.abingdonhealth.com/contract-services/lateral-flow-reader/appdx/ [accessed 2020-12-31]
23. Vashist SK, Schneider EM, Luong JHT. Commercial Smartphone-Based Devices and Smart Applications for Personalized Healthcare Monitoring and Management. Diagnostics (Basel) 2014 Aug 18;4(3):104-128 [FREE Full text] [doi: 10.3390/diagnostics4030104] [Medline: 26852680]
24. Urusov, Zherdev, Dzantiev. Towards Lateral Flow Quantitative Assays: Detection Approaches. Biosensors (Basel) 2019 Jul 17;9(3):89 [FREE Full text] [doi: 10.3390/bios9030089] [Medline: 31319629]
25. Yager P, Domingo GJ, Gerdes J. Point-of-care diagnostics for global health. Annu Rev Biomed Eng 2008;10:107-144. [doi: 10.1146/annurev.bioeng.10.061807.160524] [Medline: 18358075]
26. Rennie W, Phetsouvanh R, Lupisan S, Vanisaveth V, Hongvanthong B, Phompida S, et al. Minimising human error in malaria rapid diagnosis: clarity of written instructions and health worker performance. Trans R Soc Trop Med Hyg 2007 Jan;101(1):9-18. [doi: 10.1016/j.trstmh.2006.03.011] [Medline: 17049572]
27. Majors CE, Smith CA, Natoli ME, Kundrod KA, Richards-Kortum R. Point-of-care diagnostics to improve maternal and neonatal health in low-resource settings. Lab Chip 2017 Oct 11;17(20):3351-3387 [FREE Full text] [doi: 10.1039/c7lc00374a] [Medline: 28832061]
28. Vashist S. Point-of-Care Diagnostics: Recent Advances and Trends. Biosensors (Basel) 2017 Dec 18;7(4):62 [FREE Full text] [doi: 10.3390/bios7040062] [Medline: 29258285]
29. Perez MV, Mahaffey KW, Hedlin H, Rumsfeld JS, Garcia A, Ferris T, Apple Heart Study Investigators. Large-Scale Assessment of a Smartwatch to Identify Atrial Fibrillation. N Engl J Med 2019 Nov 14;381(20):1909-1917 [FREE Full text] [doi: 10.1056/NEJMoa1901183] [Medline: 31722151]
30. Wood CS, Thomas MR, Budd J, Mashamba-Thompson TP, Herbst K, Pillay D, et al. Taking connected mobile-health diagnostics of infectious diseases to the field. Nature 2019 Feb;566(7745):467-474 [FREE Full text] [doi: 10.1038/s41586-019-0956-2] [Medline: 30814711]
31. Interagency WGOMI. Roadmap for Medical Imaging Research and Development.: Interagency Working Group on Medical Imaging National Science and Technology Council (U.S.) URL: https://www.whitehouse.gov/wp-content/uploads/2017/12/Roadmap-for-Medical-Imaging-Research-and-Development-2017.pdf [accessed 2020-12-31]
32. Kamgno J, Pion SD, Chesnais CB, Bakalar MH, D'Ambrosio MV, Mackenzie CD, et al. A Test-and-Not-Treat Strategy for Onchocerciasis in Loa loa-Endemic Areas. N Engl J Med 2017 Nov 23;377(21):2044-2052 [FREE Full text] [doi: 10.1056/NEJMoa1705026] [Medline: 29116890]

33. Sunny S, Baby A, James BL, Balaji D, Rana MH, Gurpur P, et al. A smart tele-cytology point-of-care platform for oral cancer screening. PLoS One 2019 Nov 15;14(11):e0224885 [FREE Full text] [doi: 10.1371/journal.pone.0224885] [Medline: 31730638]
34. Haage V, Ferreira de Oliveira-Filho E, Moreira-Soto A, Kühne A, Fischer C, Sacks JA, et al. Impaired performance of SARS-CoV-2 antigen-detecting rapid diagnostic tests at elevated and low temperatures. J Clin Virol 2021 May;138:104796 [FREE Full text] [doi: 10.1016/j.jcv.2021.104796] [Medline: 33773413]
35. Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. J Am Med Inform Assoc 2016 Sep 17;23(5):899-908 [FREE Full text] [doi: 10.1093/jamia/ocv189] [Medline: 26911829]
36. Syzdykova A, Malta A, Zolfo M, Diro E, Oliveira JL. Open-Source Electronic Health Record Systems for Low-Resource Settings: Systematic Review. JMIR Med Inform 2017 Nov 13;5(4):e44 [FREE Full text] [doi: 10.2196/medinform.8131] [Medline: 29133283]
37. U.S. FDA. De Novo Classification Request.: U.S. Federal Drug Administration; 2019. URL: https://www.fda.gov/medical-devices/premarket-submissions/de-novo-classification-request [accessed 2020-12-31]
38. Garguilo JJ, Martinez S, Cherkaoui M. Medical device interoperability a standards-based testing approach. Biomed Instrum Technol 2011;45(3):249-255. [doi: 10.2345/0899-8205-45.3.249] [Medline: 21639776]
39. Roberts K, Demner-Fushman D, Voorhees EM, Hersh WR, Bedrick S, Lazar AJ, et al. Overview of the TREC 2019 Precision Medicine Track. 2019 Nov Presented at: Text Retrieval Conference; 2019; Washington D.C URL: http://europepmc.org/abstract/MED/34849512
40. Park C, Ngo H, Lavitt LR, Karuri V, Bhatt S, Lubell-Doughtie P, et al. The Design and Evaluation of a Mobile System for Rapid Diagnostic Test Interpretation. 2021 Mar 19 Presented at: Proc ACM Interact Mob Wearable Ubiquitous Technol; March 2021; Virtual p. 1-26. [doi: 10.1145/3448106]
41. Lissfelt, Jennifer, Pasquier, Julie. WHO Diagnostics Prequalification Project (DxPQ) and WHO Medicines Prequalification Project (MPQ) Mid-Term Evaluation.: Euro Health Group, UNITAID; 2016. URL: https://unitaid.org/assets/Prequalification-DxPQ-and-MPQ-MTE-Detailed-Report-1-June-2016._final.pdf [accessed 2020-12-31]
42. Murray CJL, Rosenfeld LC, Lim SS, Andrews KG, Foreman KJ, Haring D, et al. Global malaria mortality between 1980 and 2010: a systematic analysis. The Lancet 2012 Feb 04;379(9814):413-431. [doi: 10.1016/S0140-6736(12)60034-8] [Medline: 22305225]
43. Perin J, Mulick A, Yeung D, Villavicencio F, Lopez G, Strong KL, et al. Global, regional, and national causes of under-5 mortality in 2000–19: an updated systematic analysis with implications for the Sustainable Development Goals. The Lancet Child & Adolescent Health 2022 Feb;6(2):106-115 [FREE Full text] [doi: 10.1016/S2352-4642(21)00311-4] [Medline: 34800370]
44. Yong SEF, Anderson DE, Wei WE, Pang J, Chia WN, Tan CW, et al. Connecting clusters of COVID-19: an epidemiological and serological investigation. Lancet Infect Dis 2020 Jul;20(7):809-815 [FREE Full text] [doi: 10.1016/S1473-3099(20)30273-5] [Medline: 32330439]

Abbreviations

FHIR: Fast Healthcare Interoperability Resources
IUI: information utilization index
OG: open guideline
PPH: precision public health
RDT: rapid diagnostic test
WHO: World Health Organization

Edited by T Leung; submitted 31.12.20; peer-reviewed by H Mehdizadeh; comments to author 26.02.21; revised version received 24.07.21; accepted 14.04.22; published 29.07.22.

Please cite as:
Lubell-Doughtie P, Bhatt S, Wong R, Shankar AH
Transforming Rapid Diagnostic Tests for Precision Public Health: Open Guidelines for Manufacturers and Users
JMIR Biomed Eng 2022;7(2):e26800
URL: https://biomedeng.jmir.org/2022/2/e26800
doi:10.2196/26800
PMID:38875688

Original Paper

Detection of Mental Fatigue in the General Population: Feasibility Study of Keystroke Dynamics as a Real-world Biomarker

Alejandro Acien[1,2], PhD; Aythami Morales[2], PhD; Ruben Vera-Rodriguez[2], PhD; Julian Fierrez[2], PhD; Ijah Mondesire-Crump[1], MD; Teresa Arroyo-Gallego[1], PhD

[1]nQ Medical Inc, Cambridge, MA, United States

[2]School of Engineering, Universidad Autonoma de Madrid, Madrid, Spain

Corresponding Author:
Teresa Arroyo-Gallego, PhD
nQ Medical Inc
245 Main Street
Cambridge, MA, 02142
United States
Phone: 1 857 526 281
Email: gallego@nq-medical.com

Abstract

Background: Mental fatigue is a common and potentially debilitating state that can affect individuals' health and quality of life. In some cases, its manifestation can precede or mask early signs of other serious mental or physiological conditions. Detecting and assessing mental fatigue can be challenging nowadays as it relies on self-evaluation and rating questionnaires, which are highly influenced by subjective bias. Introducing more objective, quantitative, and sensitive methods to characterize mental fatigue could be critical to improve its management and the understanding of its connection to other clinical conditions.

Objective: This paper aimed to study the feasibility of using keystroke biometrics for mental fatigue detection during natural typing. As typing involves multiple motor and cognitive processes that are affected by mental fatigue, our hypothesis was that the information captured in keystroke dynamics can offer an interesting mean to characterize users' mental fatigue in a real-world setting.

Methods: We apply domain transformation techniques to adapt and transform TypeNet, a state-of-the-art deep neural network, originally intended for user authentication, to generate a network optimized for the fatigue detection task. All experiments were conducted using 3 keystroke databases that comprise different contexts and data collection protocols.

Results: Our preliminary results showed area under the curve performances ranging between 72.2% and 80% for fatigue versus rested sample classification, which is aligned with previously published models on daily alertness and circadian cycles. This demonstrates the potential of our proposed system to characterize mental fatigue fluctuations via natural typing patterns. Finally, we studied the performance of an active detection approach that leverages the continuous nature of keystroke biometric patterns for the assessment of users' fatigue in real time.

Conclusions: Our results suggest that the psychomotor patterns that characterize mental fatigue manifest during natural typing, which can be quantified via automated analysis of users' daily interaction with their device. These findings represent a step towards the development of a more objective, accessible, and transparent solution to monitor mental fatigue in a real-world environment.

(JMIR Biomed Eng 2022;7(2):e41003) doi:10.2196/41003

KEYWORDS

fatigue; keystroke; biometrics; digital biomarker; TypeNet; domain adaptation; fatigue detection; typing patterns; circadian cycles; mental fatigue; psychomotor patterns; monitoring; mental health; keystroke dynamics

Introduction

Background

Mental fatigue is a state of brain exhaustion caused by long periods of cognitive activity, lack of sleep, or stress. According to Tanaka et al [1], mental fatigue may lead to overactivation of the visual cortex in the occipital lobe, which has been linked to cognitive impairment and low psychomotor performance. Patients experiencing this condition usually report, among other symptoms, a reduction of their concentration capacity, headaches, dizziness, and slowed reflexes and responses [2]. From a clinical point of view, these psychomotor impairments induced by mental fatigue could be a sign of other emerging diseases, including neurodegenerative or cardiovascular conditions [3,4]. As an example, patients with Parkinson disease have been reported to show higher level of physical and mental fatigue in early stages of the disease than healthy participants [5]. Fatigue has been reported to be one of the major causes of disability for up to half of the patients with Parkinson disease [6], limiting their ability to participate in daily routines or social activities [7,8].

Although multiple tools exist for the assessment of fatigue, there is no clinical standard that enables an objective and complete evaluation of people's state in this domain. The most accepted method is the Fatigue Assessment Scale, a patient-reported outcome composed of 10 items that evaluate physical and physiological aspects of fatigue [9]. The subjective and episodic nature of these tools makes it difficult to detect and evaluate fatigue in daily practice and in the context of clinical trials. There is a clinical and research need to develop more accessible, accurate, and specific biomarkers to monitor fatigue and its clinical causes [10,11].

Keystroke dynamics is a biometric trait commonly used to authenticate users based on their typing patterns [12,13]. The speed of pressing and releasing keys [14] or the pressure exerted when pressing a key [15] are some of the typing features used by keystroke biometric algorithms for user authentication. Finger kinematics during typing are fine motor skills ruled by the neuromotor cortex and have also been presented as a powerful biomarker in the diagnosis and monitoring of different neurodegenerative diseases, including Parkinson disease [16-18], multiple sclerosis [19], and Alzheimer disease [20]. A recent meta-analysis carried out by Alfalahi et al [21] demonstrates the promising performance of keystroke dynamics–based models for the diagnosis of fine motor impairment in Parkinson disease and mild cognitive impairment diseases. However, the authors show caution in the transition from a controlled assessment in the clinic to the unsupervised remote diagnosis and monitoring, owing to the sparsity and unpredictable nature of typing activity in the real-world context. Continuous keystroke data are easy to gather via commodity hardware (eg, phones and laptops) without requiring the use of proprietary devices. Furthermore, remote data collection can avoid intrusive visits to the clinic, which enhances the patient's quality of life. Other prior state-of-the-art works have studied how mental fatigue affects typing activity. As an example, Ulinskas et al [22] conducted a study with 53 participants typing a fixed password. They achieved up to 91% of accuracy detecting a state of increasing fatigue between 2 consecutive keystroke sessions by using k-nearest neighbors (k-NNs) classifiers and statistical keystroke features. In contrast, in the study by Slooten et al [23], the authors study which keystroke features are influenced by mental fatigue. They suggest that the addition of keystroke dynamics features to sleep-related markers does not improve mental fatigue detection. However, they point to the subjectivity of the questionnaires used to label the fatigue keystroke data as a limitation to their study.

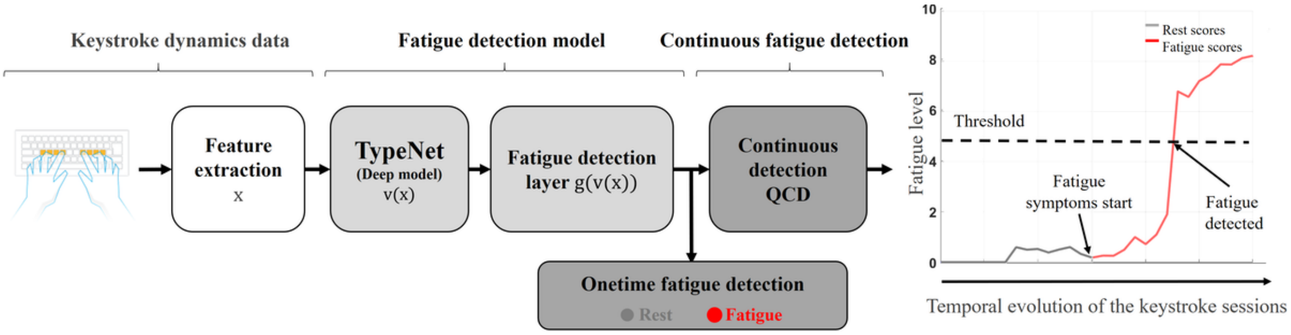
Objectives

In this paper, we study the applicability of keystroke dynamics as a potential biomarker of mental fatigue, going a step forward in the state-of-the-art characterization of this psychomotor condition by proposing a new active fatigue detection (AFD) framework based on deep neuronal networks (DNNs). To develop this, we will use TypeNet [24], a state-of-the-art DNN originally designed to model identity via typing patterns at large scale (approximately 100,000 users). The main idea behind this work is to leverage the keystroke dynamics patterns learnt by TypeNet for user recognition and to reoptimize this network for the fatigue detection task.

A schema of the proposed system is shown in Figure 1. The system is composed of 3 main elements: the input layer, the fatigue detection model, and the postprocessing module for active detection. The input layer ingests keystroke session data and generates a predefined feature vector that is then fed to the fatigue detection model. The fatigue detection model is created by connecting the output of the TypeNet network to a fatigue detection layer, which optimizes the original authentication model for fatigue identification. Finally, the postprocessing module for active detection ingests the temporal sequences of fatigue detection scores to produce users' calibrated fatigue level on the basis of their baseline or previous fatigue states. This block enables real-time monitoring of on-off fatigue fluctuations over consecutive keystroke sessions. In this work, we evaluate the proposed system in a controlled data context to test the performance of the fatigue detection model to discriminate between labeled rest and mental fatigue sessions. In addition to this, we present a real-world application of the system applied to natural typing data to evaluate its suitability to identify daily fatigue cycles in a healthy population.

The main contributions of this work are 4-fold: (1) we develop a deep neural network able to identify mental fatigue symptoms through keystroke patterns, (2) we analyze the ability of the proposed model to detect small variations in fatigue levels between different keystroke sessions, (3) we propose an AFD algorithm that continuously monitors users' keystroke session sequences to detect longitudinal variations in their fatigue state, and (4) we evaluate the applicability of the proposed system to detect fatigue trends in real-world user data.

Figure 1. Block diagram of the entire system proposed. The fatigue detection layer adapts the capacity of TypeNet to model user behavior through keystroke patterns for the fatigue detection task. This information is taken by the active detection algorithm to detect changes in users' fatigue level over consecutive keystroke sessions. QCD: quick change detection.



Methods

Keystroke Data Sets

In this section, we analyze in more detail the 3 keystroke databases (summarized in Table 1) used in this work to train and evaluate our proposed system.

- First, the Aalto database [25] was used to train the TypeNet model that we used as keystroke embedding feature extractor in our fatigue detection model. This database is composed of 168,000 participants with 15 keystroke sessions per participant. The database was acquired using a web-based questionnaire under an uncontrolled environment where each user used their own physical keyboard. All users were initially informed of the acquisition of their press (key down) and release (key up) event timings during the completion of the questionnaire. The questionnaire required users (1) to memorize an English sentence randomly chosen from a pool of 1525 sentences of the Enron mobile email and Gigaword Newswire corpus (these sentences contained a minimum of 3 words and a maximum of 70 characters) and (2) to type the memorized sentence as quickly and accurately as they could. All participants in the database completed 15 sessions (ie, one sentence for each session) on either a desktop or a laptop physical keyboard. The authors of the database reported demographic statistics of the users: 72% of the participants took a typing course, 218 countries were involved, and 85% of them had English as native language. The richness of the Aalto database resides not only in the huge amount of participants acquired but also in the diversity of ethnicities, countries, and different typing skill levels of the participants enrolled allowing TypeNet to authenticate users through keystroke dynamics at internet scale with a high performance [26].
- Second, the neuroQWERTY Sleep Inertia (nQSI) database [27] was designed to detect psychomotor impairment by waking up the participants during the night, thus inducing a sleep inertia status (a mental fatigue condition produced by lack of sleep). The database comprises 14 healthy participants with 4 keystroke sessions per participant of 15-minute duration collected in mechanical keyboards. Two of the keystroke sessions were captured during the day, whenever the participant felt well rested, labeling them as rest state (no fatigue). The other 2 keystroke sessions labeled as the fatigue ones were captured at midnight, when the participants woke up during the phase III and IV of the sleep cycle [28] to capture the keystroke sessions, thereby inducing the sleep inertia state. The acquisition process was monitored by the owners of the database to ensure the quality of the keystroke data captured in both rest and fatigue states (supervised scenario). We used this database to train and test our proposed system for the mental fatigue detection task through keystroke dynamics.
- Finally, the neuroQWERTY Crowdsource (nQCS) database [29] is composed of >800 participants from a healthy control group and group of patients with self-reported neurodegenerative diseases or other conditions (eg, Parkinson disease, Alzheimer disease, multiple sclerosis, or rheumatoid arthritis) typing on mechanical keyboards during a time span of 9 months. An enormous challenge for exploiting this data set is that the keystroke data captured were acquired passively, in a total transparent way for the participant, without any type of supervision or labeled data. In the context of this work, this database was used to study whether our proposed system was able to detect trends in the fatigue levels during the daily typing habits of the healthy participant subset (a total of 251 healthy participants).

Table 1. List of keystroke data sets used in this study.

Database	Subjects, n	Sessions, n	Session size	Supervised	Context
Aalto [25]	168,000	15	Approximately 70 keys	No	Development of TypeNet for general user typing model
neuroQWERTY Sleep Inertia [27]	14	4	15 minutes	Yes	Development and evaluation of the fatigue detection system
neuroQWERTY Crowdsource [29]	251	Approximately 1000	Approximately 3 minutes	No	Evaluation of fatigue detection in a real-world environment

Ethics Approval

Participants in the nQSI study provided informed consent before experiments, and experimental procedures were approved by the Committee On the Use of Humans as Experimental Subjects at the Massachusetts Institute of Technology (protocol number 1311).

Participants in the nQCS study provided informed consent before experiments, and experimental procedures were approved by the Committee On the Use of Humans as Experimental Subjects at the Massachusetts Institute of Technology (protocol number 1504007090).

Data Preprocessing and Feature Extraction

The raw data captured in all 3 keystroke databases are time series of 3 dimensions: press times, release times, and the keycode of each key. Owing to privacy concerns, the keycode was discarded, and the keystroke features computed for each keystroke session were based only on the press and release key time events. These timestamps were in coordinated universal time format but with different time resolution depending on the acquisition protocol and device used in each keystroke database. To normalize the keystroke data of the 3 databases, all timestamps were converted to seconds while ensuring that all keystroke features computed later are close to 1. This normalization step is necessary to avoid saturation of the neurons in the recurrent layers of our system.

The keystroke features vector is extracted at key level and is composed by (1) hold times (ie, the elapsed time between press and release a key), (2) flight times (ie, the elapsed time between 2 consecutive press events), (3) interkey latency (ie, the elapsed time between release a key and press the next key), and (4) interrelease latency (ie, the elapsed time between 2 consecutive release events). According to this, the keystroke feature vector x used as input of our model has a dimension of 150 × 4 (150 keystrokes by 4 features). If the keystroke sequence is lower than 150 keys, we compute zero padding to fill with zeros up to reach such length; otherwise, we truncate the keystroke sequence taking the first 150 keys.

The reason why we chose these keystroke features is because we wanted to ensure to keep the same feature set as the one used to evaluate the TypeNet DNN model in previous works [24,26,30] with the Aalto database. Remember that the TypeNet model is part of our fatigue detection system that we adapt for the fatigue detection task with transfer learning techniques, and therefore, the keystroke features set used to feed the TypeNet model (ie, the input of our fatigue detection model) must be the same.

System Design

The fatigue detection model is trained and tested with the labeled keystroke data from the nQSI database. As depicted in Figure 2A, the input of the TypeNet network is a keystroke feature vector x extracted from the raw keystroke data in the nQSI database. The output of TypeNet is a 1 × 128–dimensional embedding feature vector v(x) that authenticates users by applying a distance metric learning (DML) method [31]. TypeNet was originally trained to model the typing patterns of 100,000 users. The training process of TypeNet was aimed to generate a 128-dimensional feature space where keystroke events generated by the same user tend to cluster in a closer region of the feature space, whereas events from different users are projected in different areas of the same feature space. In this work, we use the nQSI data set to adapt the transformed authentication feature space to the fatigue detection task. We apply domain adaptation techniques [32] based on the addition of a fatigue detection layer that is trained to transform the authentication-based feature vectors, v(x), into fatigue detection feature vectors with the same dimension, g(v(x)), as shown in Figure 2B. The fatigue detection layer is optimized using a DML approach and a leave-one-out (LOO) cross-validation protocol. Figure 3 presents some examples showing the results of the transformation in the nQSI data set. The hypothesis underlying the method is that the features learned to model the typing patterns x of 100,000 users contain useful information to characterize users’ fatigue patterns. The fatigue detection layer serves as a nonlinear transformation g(.) to reveal such patterns in the learned space v(x). The fatigue score is computed at the output of the fatigue detection model as the Euclidean distance between pairs of fatigue detection feature vectors (equation 2).

Figure 2. Overview of the fatigue detection model design. (A) The fatigue detection model is trained with the labeled keystroke data from the neuroQWERTY Sleep Inertia database. At the output, the model separates the fatigue embedding vectors g(v(x)) that correspond to each of the 2 user's states under study (ie, fatigue or rest) while favoring proximity between the embedding vectors that belong to the same class. (B) An example of the transformation from the embedding vectors generated by TypeNet v(x) at the embedding output of the proposed model g(v(x)). The sample output shown in this figure applies t-distributed stochastic neighbor embedding (t-SNE) to generate a 2D projection of the 1×128 output.

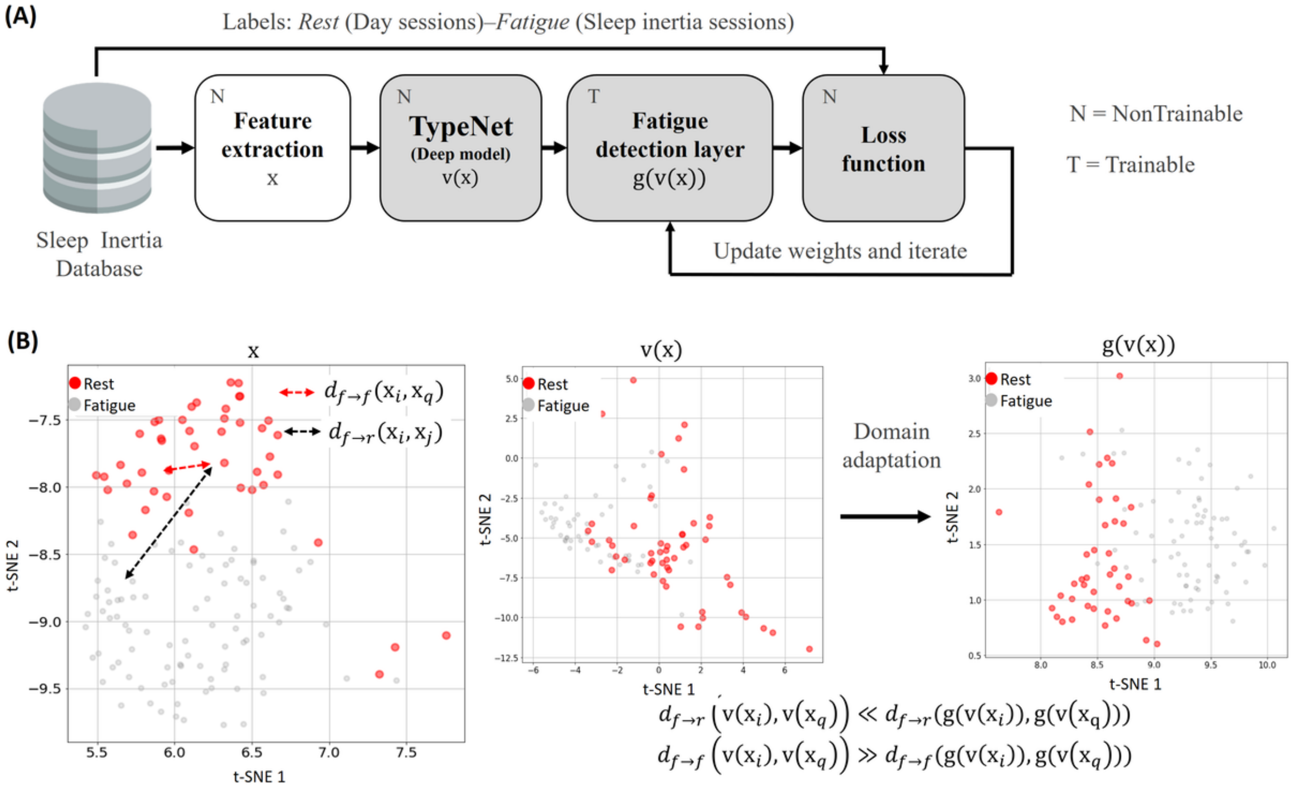
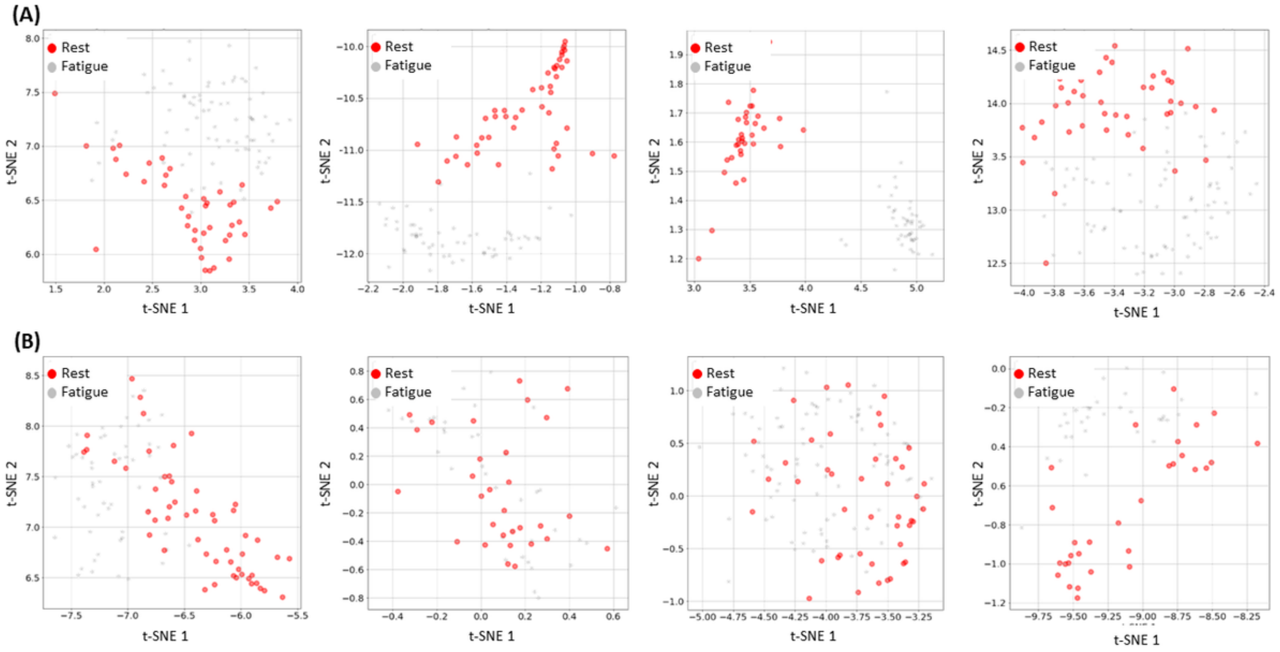

(A)
Labels: *Rest* (Day sessions)–*Fatigue* (Sleep inertia sessions)
Sleep Inertia Database
N Feature extraction x
N TypeNet (Deep model) v(x)
T Fatigue detection layer g(v(x))
N Loss function
Update weights and iterate
N = NonTrainable
T = Trainable
(B)
x
v(x)
g(v(x))
Rest
Fatigue
$d_{f\to f}(x_i, x_q)$
$d_{f\to r}(x_i, x_j)$
Domain adaptation
t-SNE 1
t-SNE 2
$d_{f\to r}\left(v(x_i), v(x_q)\right) \ll d_{f\to r}(g(v(x_i)), g(v(x_q)))$
$d_{f\to f}\left(v(x_i), v(x_q)\right) \gg d_{f\to f}(g(v(x_i)), g(v(x_q)))$

Figure 3. Intrauser variation of the embedding fatigue vectors g(v(x)). We observe how the fatigue detection model presents varying performance depending on the user. Row (A) shows examples of fatigue embedding vectors for those participants where we observe a good separation between fatigue and rest embedding vectors, whereas for participants in the row (B), the separation is not as clear. This user-dependent performance could be a result of the varying levels of intrauser fluctuations observed during natural typing [33]. t-SNE: t-distributed stochastic neighbor embedding.



TypeNet Architecture and Domain Adaptation

The TypeNet architecture proposed in the study by Acien et al [26] is composed of 2 long short-term memory layers of 128 neurons. Long short-term memory layers are a special type of recurrent neural network layers specifically designed to be sensitive to temporal changes in the input sequences, which we think could be well suited to detect relevant changes in the typing behavior of the participant when they are fatigued. In addition, each recurrent layer has a recurrent dropout of 0.2 and a dropout layer of 0.5 between them to avoid overfitting during training. The input of the TypeNet architecture has a masking layer to avoid the computation of error gradients for those zeros (ie, zeros generated when zero padding is needed for keystroke sequences lower than 150 keys) and do not contribute to the loss function during training (more details of TypeNet architecture and evaluation are provided in the study by Acien et al [26]). Finally, the output of the TypeNet architecture is an embedding feature vector v(x) of size 128 × 1.

In this work, we transform this embedding feature vector v(x) (originally used for keystroke user authentication at large scale) into a new embedding vector g(v(x)) of the same size that is better suited for the fatigue detection task. To do this, we use domain adaptation techniques [32], in which the model learns a new task (ie, the keystroke fatigue detection task) via knowledge transfer from a previously learnt task (ie, keystroke user authentication). In Figure 1, an overview of the entire transfer learning process is depicted. The output of the TypeNet model is connected to the fatigue detection layer, which is composed of a multilayer perceptron layer of 128 neurons with *relu* activation. During the training process, the keystroke feature vector x extracted from keystroke sessions of the sleep inertia database is used to feed the TypeNet network, which is frozen during the entire training process so the weights of this network are not altered. Then, TypeNet computes the embedding features vector v(x) that are optimized for keystroke user authentication, thanks to the previous training with the Aalto database in Acien et al [26]. Finally, the fatigue detection layer is fed with this embedding feature vector and learns to transform these embedding features into a new feature embedding vector g(v(x)) optimized for the fatigue detection task, thanks to the labeled data of the nQSI sleep inertia database.

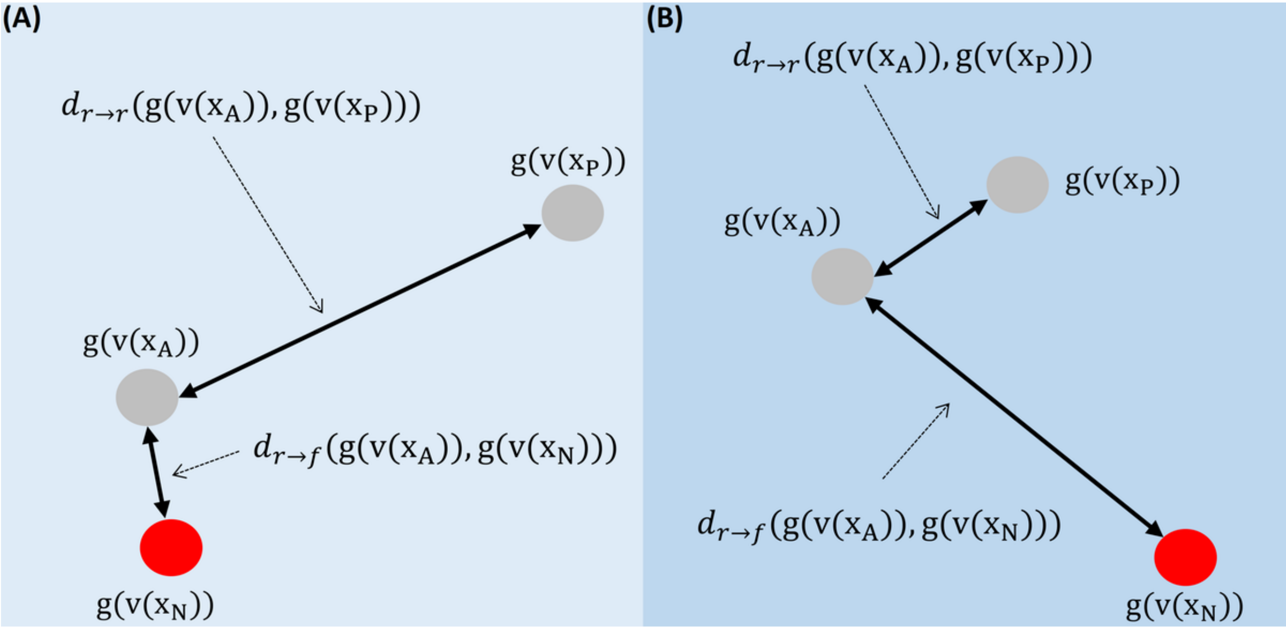
This type of domain adaptation process is also referred to as fine tuning, where the part of the TypeNet architecture that has the knowledge of typing patterns from thousands of users of the Aalto database is frozen, and therefore, we only need to train the last layer (the fatigue detection layer) to adapt these typing patterns for the fatigue detection task with the sleep inertia database. The main reason why we use transfer learning with fine-tuning techniques is because to train a DNN model from the scratch for the fatigue detection task, we will need thousands of participants with labeled keystroke data to make the model robust, generalizable, and accurate. This technique allows us to overcome this issue, taking advantage of other DNN models previously trained with thousands of participants for a similar task like TypeNet, and adapt it for the fatigue detection task using only 16 participants of the sleep inertia database. Fine-tuning techniques have been broadly used in state-of-the-art works [34-36], where the databases used are not large enough to train a DNN model from scratch.

Finally, to train the fatigue detection model successfully, we use the triplet loss function. This loss function is well suited for DML approaches where the output of the model to train is an embedding feature vector instead of a single score. A triplet is composed by 3 different samples from 2 different classes: Anchor (A) and Positive (P) are different keystroke sequences from the same class (fatigue or rest), and Negative (N) is a keystroke sequence from the other class. The triplet loss function is defined as follows:

where α is a margin between positive and negative pairs and *d* is the Euclidean distance calculated as follow:

This learning process minimizes the distance between embedding vectors from the same class (d(g(v(x_A)), g(v(x_P)))), and maximizes it for embeddings from different classes (d(g(v(x_A)), g(v(x_N)))). Note that all 3 samples x_A, x_P, and x_N belong to the same participant to avoid intrauser variations as much as possible. An example of how the triplet loss function works is depicted in Figure 4A, where g(v(x_P)) and g(v(x_A)) are 2 feature embedding vectors (ie, the output of the fatigue detection model when fed with x_P and x_A samples, respectively) that belong to the same class, whereas g(v(x_N)) belongs to the opposite class. During the training process (Figure 4B), the triplet loss function will make g(v(x_P)) and g(v(x_A)) get closer at the same time they get far from g(v(x_N)). Remember that we only train the fatigue detection layer because the TypeNet network is frozen during training (fine tuning), thereby this entire process is learnt by the fatigue detection layer. The unique purpose of this layer is to separate in the latent space the feature embedding vectors that belong to the rest state from those that belong to the fatigue state. Examples of the final results are shown in Figure 3 by applying dimensional reduction to the embedding feature vectors for 2D visualization. Regarding experimental protocol details, we follow a LOO cross-validation strategy by using all participants but one of the sleep inertia database to train the proposed system and testing with the remaining participant. This means that we have 16 different fatigue detection models (one for each test participant). Regarding training details, the hyperparameters remain the same as those used to train TypeNet in the study by Acien et al [24]: learning rate of 0.005 and Adam optimizer with β_1=0.9, β_2=0.999, and $\varepsilon=10^{-8}$. The models were trained for 30 epochs with 100 batches per epoch and 64 triplets in each batch.

Figure 4. Example of how triplet loss works. 2D representation of the embedding feature fatigue vectors g(v(x)) before (A) and after (B) the triplet loss training. The embedding vectors that belong to the same class ($g(v(x_P))$ and $g(v(x_A))$) get closer; meanwhile, they get far from the embedding of the opposite class ($g(v(x_P))$ and $g(v(x_N))$).


(A)
$d_{r\rightarrow r}(g(v(x_A)), g(v(x_P)))$
$g(v(x_P))$
$g(v(x_A))$
$d_{r\rightarrow f}(g(v(x_A)), g(v(x_N)))$
$g(v(x_N))$
(B)
$d_{r\rightarrow r}(g(v(x_A)), g(v(x_P)))$
$g(v(x_P))$
$g(v(x_A))$
$d_{r\rightarrow f}(g(v(x_A)), g(v(x_N)))$
$g(v(x_N))$

Quick Change Detection Algorithm

The AFD algorithm is based on the quick change detection algorithm proposed in the study by Perera et al [37] for intrusion detection based on mobile behavior biometrics. In this work, the algorithm is redesigned for the AFD task. The algorithm is based on calculating a new score from the cumulative sum of previous events (keystroke sessions). If the participant is in a rested state (gray lines in Figure 5), the cumulative sum will be almost 0. At the moment the mental state of the participant changes into fatigue during typing, this score will tend to increase until reaching a certain threshold, in which we detect the fatigue symptoms. This module can be interpreted as a postprocessing step connected at the output of the fatigue detection model to increase the reliability of the system and to account for the relevance of participants' preceding states when computing their current fatigue score.

To evaluate the AFD algorithm, we will use the precomputed fatigue detection scores resulting from the LOO framework that optimized the fatigue detection model. This ensures that the condition of independence between training and testing sets is carried over in this new experiment. In this context, for a given participant, the cumulative sum is calculated as follows:

where j means the actual keystroke session and is the previous cumulative score. L_j is the contribution of the actual event calculated as the log-likelihood ratio between score distributions:

where $score_j$ is the fatigue score of the participant's current event, and f_R, f_F are, respectively, the probability density estimators of the participant's remaining rest and fatigue scores. Note that the output of the fatigue detection model is an embedding feature vector g(v(x)) of size 1×128, so we compute t-distributed stochastic neighbor embedding for dimensional reduction to one dimension (ie, we reduce the size of the embedding vector to one) to obtain a single fatigue score $score_j$. According to equation 4, the log-likelihood ratio L_j will be negative if $score_j$ belongs to rested keystroke session and positive in the opposite case, and therefore, multiple consecutive keystroke sessions of the fatigued participant will increase the cumulative sum $score_j^{AFD}$. Figure 6 depicts an example of the entire AFD algorithm pipeline for a single participant. The fatigue detection model computes the embedding feature vector g(v(x)) when fed with a keystroke session. Then, we compute t-distributed stochastic neighbor embedding for dimensional reduction to obtain the $score_{j.}$ Finally, we upgrade $score_j^{AFD}$ by computing the L_j with the new score according to equation 3, which will increase up to reach the fatigue detection threshold in case the participant is fatigued.

Figure 5. Active fatigue detection (AFD) curves. (A), (B), and (C) are 3 different use cases of the AFD algorithm where the threshold chosen affects the performance. (D) It shows the probability of false detection (PFD) versus the probability of nondetection (PND) and PFD versus average detection delay (ADD) curves as a result of moving the threshold; the value chosen for the threshold is the point where both PFD and PND values are equal, called equal error rate (EER).

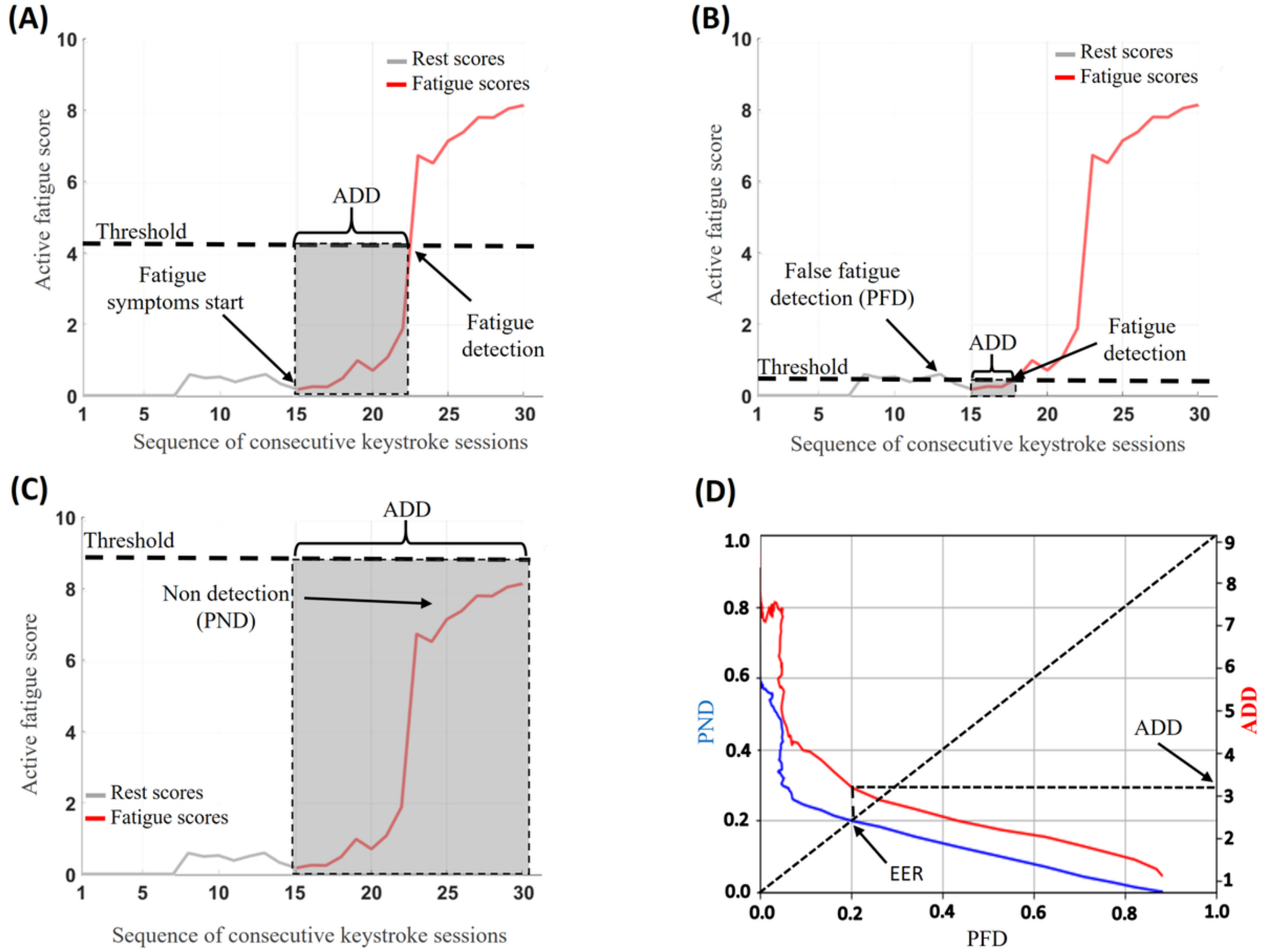
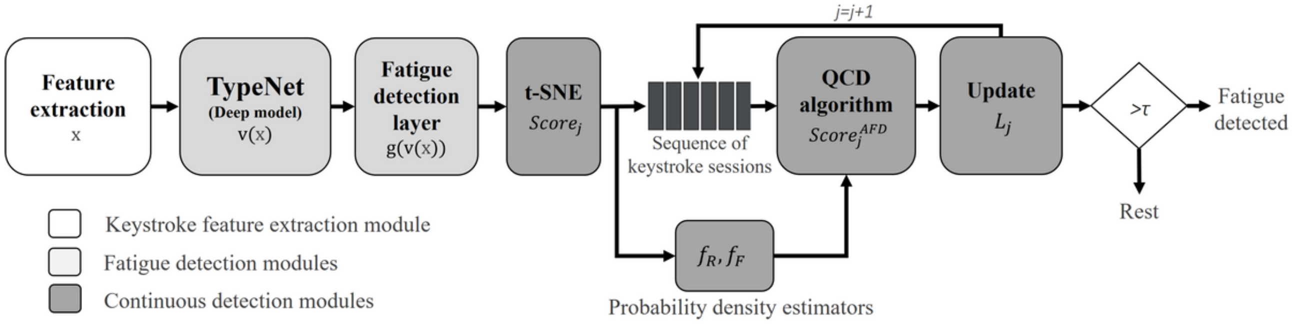


Figure 6. The entire pipeline of the active detection algorithm. The score_j is computed by performing t-distributed stochastic neighbor embedding (t-SNE) for dimensional reduction to the embedding fatigue vector g(v(x)). Then, score^{AFD} is obtained by comparing score_j with the distributions obtained from the neuroQWERTY Sleep Inertia (nQSI) database. Finally, a threshold τ is used to detect the Fatigue states. QCD: quick change detection.



Results

Onetime Fatigue Detection Approach

To evaluate the performance of the fatigue detection model, we use the nQSI data set to generate a pool of intrauser keystroke sample pairs. We contemplate a binary classification framework based on 2 scenarios: (1) no change—when the 2 samples belong to the same class (fatigue→fatigue or rest→rest) and (2) change—when the 2 samples belong 2 different classes (fatigue→rest or rest→fatigue). In Figure 2, we can observe the distances for 2 examples: $d_{f\to f}(x_i, x_j)$ is the distance between 2 fatigue samples (no change, distance between 2 red dots) and $d_{f\to r}(x_i, x_q)$ is the distance between the fatigue and rest sample (change, distance between a red and a gray dot). The distance between samples is directly compared with a predefined threshold. A fatigue score superior to the threshold reveals a change in the keystroke patterns, whereas a value below the threshold implies no change. We compare the performance of the fatigue detection model based on DML with different statistical classification algorithms trained with the feature vectors x: random forest (RF), support vector machine (SVM) with Gaussian Kernel, and k-NN. In addition, we also compare with the proposed fatigue detection model but replacing the DML approach by a softmax activation layer trained as a binary classification model using binary cross entropy loss. This provides a reference deep learning model used as baseline to compare with our DML approach. Figure 7 presents the receiver operating characteristic analysis comparison in 2 different setups. In the first one, we limit the input size to 150 keystrokes per sample. This input format was defined in accordance to the design of the pretrained TypeNet architecture. In this scenario, the best performance is achieved by the proposed fatigue detection model that achieves an area under the curve (AUC) of 72.1%, followed by the RF classifier with AUC of 68.4%. The worst performance is observed in the softmax-based variation of the proposed fatigue detection model.

In the second set-up, we increase the input size to 5-minute long keystroke sessions (ie, an average of approximately 1100 keys per sample) for the RF, SVM, and k-NN classifiers, while keeping the original 150-keystroke long inputs for the proposed fatigue detection methods and its softmax variation (owing to the limitation of 150 keys as the input size of the TypeNet model). In this case, the DML approach is slightly outperformed by the RF and SVM classifiers that present AUCs of 77.8% and 74.4%, respectively, in exchange of larger input data. Finally, we summarize the performance metrics for the 2 setups proposed in Table 2. We can observe that our DML approach achieved the highest F_1-score, a measure of the test accuracy, in both scenarios. Sensitivity and specificity values are estimated using the closest-to-(0,1) corner in the receiver operating characteristic plane to define the cutoff point. Performance metrics are computed by pooling the cross-validated scores into a single set of predictions used to generate an overall metric estimate for the whole system.

Figure 7. Receiver operating characteristic (ROC) analysis for fatigue detection. Area under the curve (AUC) scores computed with keystroke sample pairs of length 150 keys (A) and 5-minute duration (B). The ROC curves were calculated independently for each participant, and the ROCs showed are the average of all of them. DML: distance metric learning; k-NN: k-nearest neighbor; RF: random forest; SVM: support vector machine.

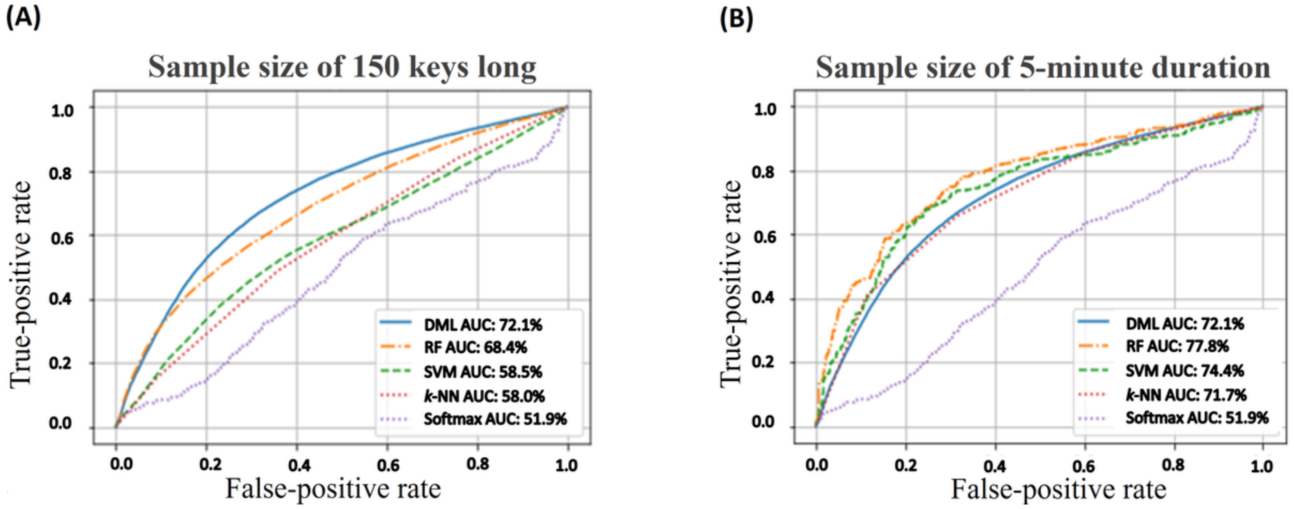


Table 2. Performance metrics of the onetime fatigue detection approach.

Set up	System	AUC[a] (%)	P value	Specificity (%)	Sensitivity (%)	Precision (%)	F_1-score (%)
150 keys	Fatigue (DML[b])	72.1	<.001	73	69	67	72.2
150 keys	Random forest	68.4	<.001	68	63	64.6	70.3
150 keys	Support vector machine	58.5	<.001	58	58	57.9	65.2
150 keys	k-nearest neighbor	58	<.001	77	51	64.6	70.3
150 keys	Fatigue (Softmax)	51.9	<.001	50	52	48	49.1
5 minutes	Fatigue (DML)	72.1	<.001	73	69	67	72.2
5 minutes	Random forest	77.8	<.001	70	76	66.3	71
5 minutes	Support vector machine	74.4	<.001	70	73	65.9	70.7
5 minutes	k-nearest neighbor	71.7	<.001	76	65	64.7	67.6
5 minutes	Fatigue (Softmax)	51.9	<.001	50	52	48	49.1

[a] AUC: area under the curve.

[b] DML: distance metric learning.

Continuous Fatigue Detection Approach

In this experiment, we consider the quick change detection algorithm [37] that dynamically updates a confidence fatigue score by calculating a cumulative sum from previously measured fatigue states. The purpose of this algorithm is to adapt the fatigue detection method to the needs posed by real-time evaluation of fatigue in a real-world environment.

In Figure 5A, we show an example of the application of this algorithm at the output of the fatigue detection model. The example uses a simulated sequence of keystroke sessions generated by concatenating 15 rest and 15 fatigue keystroke samples from a user in the nQSI database. As the simulated sequence starts in a rest state, the initial fatigue scores are lower and close to 0 during the first 15 evaluation intervals (ie, the 15 user keystroke sessions labeled as rest in the nQSI database). As the simulated sequence starts introducing fatigue samples (from the remaining 15 user keystroke sessions labeled as fatigue), the AFD score tends to increase until it reaches a certain threshold that would indicate there has been a fatigue state change. The number of keystroke sessions elapsed since the models start getting fatigue samples until the AFD algorithm reaches the fatigue threshold is called average detection delay (ADD). This parameter measures the number of keystroke sessions required to detect fatigue since the symptoms start.

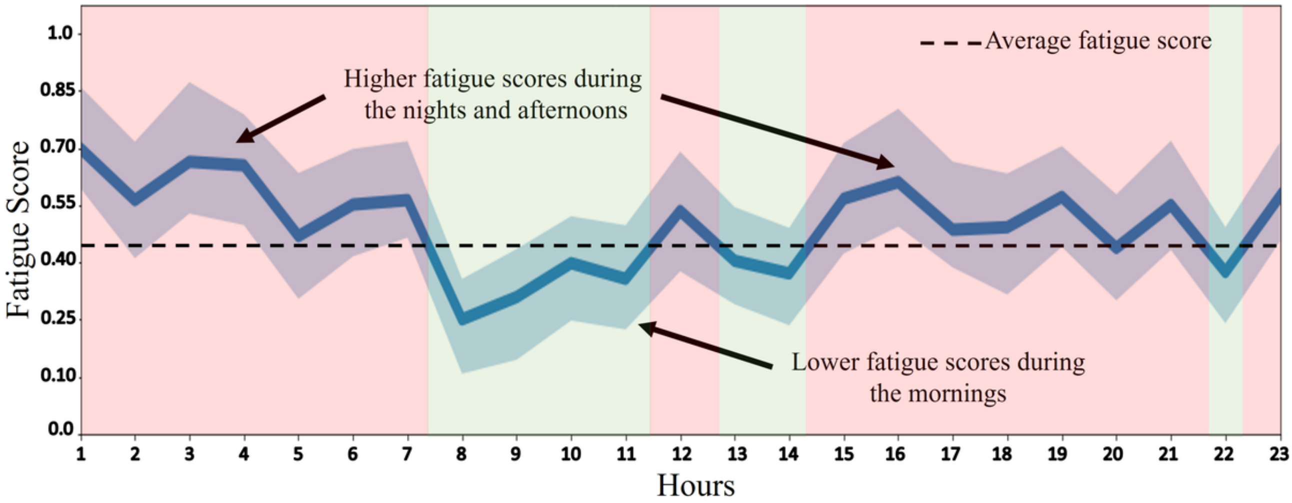
The configuration of the threshold in the AFD score is crucial for the performance of the algorithm. As shown in Figure 5B, as we lower the threshold, we reduce the ADD from 7 (Figure 5A) to 3 keystroke sessions, in exchange of a higher risk of false positives. This value is called probability of false detection (PFD) and measures the probability of false fatigue detection (similar to the false match rate). In contrast, increasing the threshold controls the PFD at the cost of increasing the ADD as well as the probability of nondetection (PND). PND measures the probability of the active fatigue score never reaching the threshold over a sequence of keystroke sessions in a fatigued interval (Figure 5C).

According to this, there is always a trade-off between the PND and PFD values as we move the threshold. Figure 5D shows the PND (left y-axis) versus PFD and ADD (right y-axis) versus PFD. To optimize both specificity and sensitivity metrics at the same time, we have the point equal error rate (EER). The EER value is the point where the blue curve (ie, PND vs PFD) crosses the diagonal (the dotted black line) and is equal to 20%. This would be equivalent to an AUC=100−EER=80%. Finally, based on the configuration of the threshold, we can infer the number of fatigue keystroke sessions required, according to our results, to reach the threshold (ie, the ADD value). Once we have calculated the EER that minimizes both PND and PFD values, the red curve (PFD vs ADD) in Figure 5D indicates the number of keystroke sessions required (ADD) for the chosen PFD, which is slightly above 3.

Independent Evaluation in Real-world Environment

As mentioned above, the nQSI database used to evaluate our system was acquired under supervised conditions with labeled keystroke sessions. To evaluate the behavior of the proposed method in the context of its intended use, we applied the resulting model to the nQCS database. As a reminder, this database includes keystroke data from a group of healthy volunteers that was captured during their daily use of the device, without any supervision or prompt to stimulate typing activity. We compute the fatigue scores measured on each pair of consecutive keystroke sessions for each user typing stream. Each user typing stream is composed of multiple keystroke sessions generated over varying observation periods and activity levels. We only take into account the fatigue scores obtained between keystroke sessions with elapsed time of <2 hours within the same day to avoid long pauses between sessions that may introduce artifacts in the resulting fatigue signal. In Figure 8, we present the aggregate trends of the fatigue score levels versus the time of the day. The results suggest lower fatigue levels during the morning and midday hours. Higher fatigue scores are observed during the afternoon hours and overnight. Note that this figure was obtained by averaging the scores from all 251 volunteers in the nQCS database, and therefore, there is an equalization effect caused by different user's habits.

Figure 8. Fatigue score analysis in the neuroQWERTY Crowdsource (nQCS) database. The fatigue scores are calculated, at the user level, between consecutive sessions over their daily typing activity. The graph presents the nQCS population aggregate average and CIs of the resulting fatigue score daily sequences.



Discussion

Principal Findings

Using domain adaptation techniques, we leveraged an algorithm built for user authentication to detect signs of fatigue via natural typing. The resulting classifier was then adapted for real-time fatigue monitoring by appending an active detection algorithm that compares successive user states. This allows for background evaluation of users' fatigue state in an objective and real-world environment. The proposed classifier was able to differentiate intrapatient fatigue versus rested states with an AUC of 72.1% in onetime detection set-up. When simulating a continuous detection set-up by concatenating consecutive keystroke sessions, the proposed model is able to detect early fatigue symptoms after 4 keystroke sessions with an AUC of 80%. A preliminary application of the fatigue classifier combined with active detection showcased its applicability to real-world data in a crowdsource data set. Given that this method relies on data collected passively from a user's daily interactions with their computer, the proposed pipeline operates unobtrusively with low burden and allows for a background, objective evaluation of a user's fatigue state in the real-world environment.

Relying on machine learning techniques, we were able to liaise a large data set created to study typing behaviors in the general population with the information gathered in a limited size data set built specifically to characterize fatigue through the analysis of keystroke dynamics. This approach allowed us to apply a deep learning architecture in the absence of a high-dimensional data set specifically characterized for the phenomenon under study, quantification of daily fatigue levels in users' keystroke patterns. Our work exhibits the potential of domain adaptation techniques to minimize the complexity of gathering large and curated data repositories required to train deep learning models by taking advantage of open-source unsupervised data sets in combination with much smaller supervised data sets. In our case, the Aalto database supplies the high volume of data required to build a network optimized for user authentication that is then fine-tuned using the sleep inertia data set to solve the fatigue detection task. Another novel technical contribution of this work is the addition of an active detection algorithm that adapts the classifier for its application in real-time fatigue detection. This dynamic adaptation of the fatigue score threshold turns users into their own controls over time by carrying information from previous estimates to generate the present score. It is one of the main differentiators of this work from prior state-of-the-art approaches to this problem, which generally use a cross-sectional design to evaluate fatigue at a given time point [38-40]. As an example, Ulinskas et al [38] assign fatigue levels to users' data based on the time of the day. Morning, afternoon, and evening data generated by the same user are treated as independent samples in a multivariate classification framework that ignores the sequential relation between fatigue states over daily cycles. The classification results in the controlled experiment (ie, accuracy in the separation of rested vs fatigue samples in the sleep inertia data set) are worse than the ones presented in previous work completed using the same data set [27]. However, this approach reduces significantly the size of the input sample, 150 keystroke sequences (<1 minute at average speed) in comparison with the 15-minute long typing samples used in the study by Giancardo et al [27]. The value of this parameter is critical for the applicability of fatigue detection via keystroke monitoring in a real-world setting, as users are unlikely to generate continuous 15-minute long typing samples on their daily use of computer keyboards. When applied on an independent data set comprised by natural typing data collected in a real-world environment, the population-level results align with the results presented in previous studies on daily sleep and alertness cycles [41], which suggest high alertness during daytime, peaking a few hours after awakening, and higher sleepiness during nighttime.

In general, our results suggest that users are usually more awake and active during the mornings. Fatigue appears generally during the afternoon and increases as the day gets closer to regular sleep times. The daily averaged scores suggest a subtle fatigue peak after midday that could be associated with what has been referred in the literature as postlunch dip in performance [42]. This consensus with sleep and performance studies supports our hypothesis that keystroke dynamics can be used to quantify daily fatigue in computer users in an objective and unobtrusive manner. However, it is important to note that these daily cycle results have been analyzed at a population level and are not considering the variability in participants' personal routines and schedules. Future studies pairing keystroke with other high-frequency fatigue–related data (eg, sleep and activity) could help us better assess the performance of the proposed method at user level.

As for its clinical application, fatigue is a common symptom that can precede or reflect the presence of a more serious mental or physical condition. The current standard to clinically assess fatigue relies on patient-reported outcomes through standardized questionnaires, such as the Fatigue Severity Scale [43]. To identify fatigue as a symptom, patients must first identify unusually excessive fatigue patterns and then alert their physician before it can be further investigated. This leaves fatigue as a commonly overlooked or unrecognized predictor of other emerging disorders [3,4]. Fatigue has also been reported as a frequent side effect of disease treatment [44] and long-term sequel of conditions such as COVID-19 [45].

The proposed methodology is designed to validate an approach for objective and passive fatigue monitoring. Leveraging the widespread use of PCs, this framework presents an opportunity to provide more visibility and accurate tracking of fatigue and its clinical implications. As it runs in the background of users' computers, this approach could potentially be used to alert patients and health care professionals of early signs of abnormal fatigue to uncover progressive disease or the presence of underlying conditions. In the context of clinical trials or during disease management, this method could also be used to enable objective and real-world evaluation of the impact of newly developed or existing treatment regimens on a patient's fatigue state.

As a major limitation of this work, the fatigue detection model performs better for some participants than others because of the intrauser variations when typing [33]. In users who show little variation between resting and fatigue states, the model does not effectively classify performance. An example of this is shown in Figure 3, where we can observe a clear separation between the rest keystroke sessions and the fatigue ones for the participants (Figure 3A), meanwhile the fatigue detection model struggles when trying to separate the keystroke sessions for the participants of the Figure 3B with poor results.

Conclusions

This work presents a step toward the development of a real-world fatigue monitoring tool that operates passively by leveraging users' natural interaction with their PCs. It is important to note that the data set used in these analyses is composed solely of healthy controls; future work should evaluate the performance of the proposed method in a cohort that includes participants with conditions affecting psychomotor health that may mask or be confounded by fatigue symptoms. Another limitation and potential line for future research is that this work has been tested using mechanical keyboard data, thus future applications of this specific methodology require users who type frequently on mechanical keyboard devices. Adapting this framework to include touchscreen devices would expand the population that could benefit from this method. Given that typing kinematics vary significantly between mechanical and touchscreen devices, this adaptation would require additional studies. The limited dimension of the sleep inertia database is another aspect to take into account in future studies. Although the use of domain adaptation techniques reduces the need for larger supervised data sets, increasing the size of the controlled cohort would allow for optimization of the target task layer and independent validation of the fatigue detection classifier. Finally, although the crowdsource results are similar to previously published studies on daily alertness, full validation would require a labeled real-world data set to test the generalizability of the proposed framework for its application in the real work setting. Additional validation in specific use case scenarios would pave the way for use of this method as an objective, high-resolution, and quasicontinuous way to monitor users' fatigue with minimal burden on their daily routine.

Acknowledgments

This work is a collaboration between nQ Medical Inc and Biometrics and Data Pattern Analytics Lab group and has been supported by the following projects: Training in Secure and Privacy-Preserving Biometrics-European Training Network (MSCA-ITN-2019-860813), INTER-ACTION (PID2021-126521OB-I00 MICINN/FEDER), BBforTAI (PID2021-127641OB-I00 MICINN/FEDER), IDEA-FAST (H2020-IMI2-2018-15-853981), edBB (UAM), and Instituto de Ingenieria del Conocimiento. AA is supported by an FPI fellowship from the Spanish MINECO. AM is supported by “Programa de Excelencia del Profesorado Universitario” from CAM. The authors would like to thank the neuroQWERTY team for their work in creating the neuroQWERTY Crowdsource and neuroQWERTY Sleep Inertia data sets. The authors would like to thank the nQ Medical's team members for their feedback in the preparation of this manuscript.

Authors' Contributions

AA, AM, and TA-G conceived the experiments; AA and TA-G conducted the experiments; and TA-G, RV-R, and JF analyzed the results. All authors reviewed the manuscript.

Conflicts of Interest

The authors declare no competing nonfinancial interests but the following competing financial interests: AA, IM-C, and TA-G are employees at nQ Medical Inc and received a regular salary while contributing to this work.

References

1. Tanaka M. Effects of mental fatigue on brain activity and cognitive performance: a magnetoencephalography study. Anat Physiol 2015;s4:5. [doi: 10.4172/2161-0940.S4-002]
2. Johansson B, Starmark A, Berglund P, Rödholm M, Rönnbäck L. A self-assessment questionnaire for mental fatigue and related symptoms after neurological disorders and injuries. Brain Inj 2010 Jan;24(1):2-12. [doi: 10.3109/02699050903452961] [Medline: 20001478]
3. O'Keefe-McCarthy S, McGillion MH, Victor JC, Jones J, McFetridge-Durdle J. Prodromal symptoms associated with acute coronary syndrome acute symptom presentation. Eur J Cardiovasc Nurs 2016 Apr;15(3):e52-e59. [doi: 10.1177/1474515115580910] [Medline: 25851233]
4. Goldman J, Postuma R. Premotor and nonmotor features of Parkinson's disease. Curr Opin Neurol 2014 Aug;27(4):434-441 [FREE Full text] [doi: 10.1097/WCO.0000000000000112] [Medline: 24978368]
5. Lou J, Kearns G, Oken B, Sexton G, Nutt J. Exacerbated physical fatigue and mental fatigue in Parkinson's disease. Mov Disord 2001 Mar;16(2):190-196. [doi: 10.1002/mds.1042] [Medline: 11295769]
6. Kluger B, Herlofson K, Chou KL, Lou JS, Goetz CG, Lang AE, et al. Parkinson's disease-related fatigue: a case definition and recommendations for clinical research. Mov Disord 2016 May;31(5):625-631 [FREE Full text] [doi: 10.1002/mds.26511] [Medline: 26879133]

7. Friedman J, Brown RG, Comella C, Garber CE, Krupp LB, Lou JS, Working Group on Fatigue in Parkinson's Disease. Fatigue in Parkinson's disease: a review. Mov Disord 2007 Feb 15;22(3):297-308. [doi: 10.1002/mds.21240] [Medline: 17133511]
8. Zesiewicz TA, Patel-Larson A, Hauser RA, Sullivan KL. Social security disability insurance (SSDI) in Parkinson's disease. Disabil Rehabil 2007 Dec 30;29(24):1934-1936. [doi: 10.1080/09638280701257247] [Medline: 17852221]
9. De Vries J, Michielsen HJ, Van Heck GL. Assessment of fatigue among working people: a comparison of six questionnaires. Occup Environ Med 2003 Jun;60 Suppl 1:i10-i15 [FREE Full text] [doi: 10.1136/oem.60.suppl_1.i10] [Medline: 12782741]
10. Rahimian Aghdam S, Alizadeh SS, Rasoulzadeh Y, Safaiyan A. Fatigue assessment scales: a comprehensive literature review. Arch Hyg Sci 2019 Oct 01;8(3):145-153. [doi: 10.29252/ArchHygSci.8.3.145]
11. Michielsen H, De Vries J, Van Heck GL, Van de Vijver FJ, Sijtsma K. Examination of the dimensionality of fatigue. Eur J Psychol Assess 2004 Jan;20(1):39-48. [doi: 10.1027/1015-5759.20.1.39]
12. Kim J, Kang P. Freely typed keystroke dynamics-based user authentication for mobile devices based on heterogeneous features. Pattern Recognit 2020 Dec;108:107556. [doi: 10.1016/j.patcog.2020.107556]
13. Morales A, Fierrez J, Tolosana R, Ortega-Garcia J, Galbally J, Gomez-Barrero M, et al. Keystroke biometrics ongoing competition. IEEE Access 2016;4:7736-7746. [doi: 10.1109/ACCESS.2016.2626718]
14. Banerjee S, Woodard D. Biometric authentication and identification using keystroke dynamics: a survey. J Pattern Recognit Res 2012;7(1):116-139. [doi: 10.13176/11.427]
15. Acien A, Morales A, Vera-Rodriguez R, Fierrez J. Keystroke mobile authentication: performance of long-term approaches and fusion with behavioral profiling. In: Proceedings of the Pattern Recognition and Image Analysis: 9th Iberian Conference. 2019 Presented at: IbPRIA '19; Jul 1–4, 2019; Madrid, Spain.
16. Giancardo L, Sánchez-Ferro A, Arroyo-Gallego T, Butterworth I, Mendoza CS, Montero P, et al. Computer keyboard interaction as an indicator of early Parkinson's disease. Sci Rep 2016 Oct 05;6:34468 [FREE Full text] [doi: 10.1038/srep34468] [Medline: 27703257]
17. Arroyo-Gallego T, Ledesma-Carbayo MJ, Butterworth I, Matarazzo M, Montero-Escribano P, Puertas-Martín V, et al. Detecting motor impairment in early Parkinson's disease via natural typing interaction with keyboards: validation of the neuroQWERTY approach in an uncontrolled at-home setting. J Med Internet Res 2018 Mar 26;20(3):e89 [FREE Full text] [doi: 10.2196/jmir.9462] [Medline: 29581092]
18. Papadopoulos A, Iakovakis D, Klingelhoefer L, Bostantjopoulou S, Chaudhuri KR, Kyritsis K, et al. Unobtrusive detection of Parkinson's disease from multi-modal and in-the-wild sensor data using deep learning techniques. Sci Rep 2020 Dec 07;10(1):21370 [FREE Full text] [doi: 10.1038/s41598-020-78418-8] [Medline: 33288807]
19. Lam K, Meijer KA, Loonstra FC, 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;27(9):1421-1431 [FREE Full text] [doi: 10.1177/1352458520968797] [Medline: 33150823]
20. Van Waes L, Leijten M, Mariën P, Engelborghs S. Typing competencies in Alzheimer's disease: an exploration of copy tasks. Comput Human Behav 2017 Aug;73:311-319. [doi: 10.1016/j.chb.2017.03.050]
21. Alfalahi H, Khandoker AH, Chowdhury N, Iakovakis D, Dias SB, Chaudhuri KR, et al. Diagnostic accuracy of keystroke dynamics as digital biomarkers for fine motor decline in neuropsychiatric disorders: a systematic review and meta-analysis. Sci Rep 2022 May 11;12(1):7690 [FREE Full text] [doi: 10.1038/s41598-022-11865-7] [Medline: 35546606]
22. Ulinskas M, Woźniak M, Damaševičius R. Analysis of keystroke dynamics for fatigue recognition. In: Proceedings of the 17th International Conference on Computational Science and Its Applications. Cham, Switzerland: Springer; 2017 Presented at: ICCSA '17; July 3-6, 2017; Trieste, Italy p. 235-247.
23. Slooten V. , Mirna, M. In: , Smolders, I., & Kort, I. Identifying fatigue using keystroke dynamics. Eindhoven: Eindhoven University of Technology; Jan 22, 2021:1-82.
24. Acien A, Morales A, Vera-Rodriguez R, Fierrez J, Monaco JV. TypeNet: scaling up keystroke biometrics. In: Proceedings of the 2020 IEEE International Joint Conference on Biometrics. 2020 Presented at: IJCB '20; September 28-October 01, 2020; Houston, TX, USA p. 1-7.
25. Dhakal V, Feit A, Kristensson PO, Oulasvirta A. Observations on typing from 136 million keystrokes. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. 2018 Presented at: CHI '18; April 21-26, 2018; Montreal, Canada p. 1-12.
26. Acien A, Morales A, Monaco JV, Vera-Rodriguez R, Fierrez J. TypeNet: deep learning keystroke biometrics. IEEE Trans Biom Behav Identity Sci 2022 Jan;4(1):57-70. [doi: 10.1109/TBIOM.2021.3112540]
27. Giancardo L, Sánchez-Ferro A, Butterworth I, Mendoza CS, Hooker JM. Psychomotor impairment detection via finger interactions with a computer keyboard during natural typing. Sci Rep 2015 Apr 16;5:9678 [FREE Full text] [doi: 10.1038/srep09678] [Medline: 25882641]
28. Carskadon MA, Dement WC. Chapter 2 - Normal human sleep: an overview. In: Kryger MH, Roth T, Dement EC, editors. Principles and Practice of Sleep Medicine. 4th edition. Philadelphia, PA, USA: Saunders; 2005:13-23.
29. Tripathi S, Arroyo-Gallego T, Giancardo L. Keystroke-dynamics for Parkinson's disease signs detection in an at-home uncontrolled population: a new benchmark and method. IEEE Trans Biomed Eng (forthcoming) 2022 Jun 29;PP. [doi: 10.1109/TBME.2022.3187309] [Medline: 35767495]

30. Morales A, Fierrez J, Acien A, Tolosana R, Serna I. SetMargin loss applied to deep keystroke biometrics with circle packing interpretation. Pattern Recognit 2022 Feb;122:108283. [doi: 10.1016/j.patcog.2021.108283]
31. Hadsell R, Chopra S, LeCun Y. Dimensionality reduction by learning an invariant mapping. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2006 Presented at: CVPR '06; June 17-22, 2006; New York, NY, USA.
32. Singh R, Vatsa M, Patel VM, Ratha N. Domain Adaptation for Visual Understanding. Cham, Switzerland: Springer; 2020.
33. Acien A, Hernandez-Ortega J, Morales A, Fierrez J, Vera-Rodriguez R, Ortega-Garcia J. On the analysis of keystroke recognition performance based on proprietary passwords. In: Proceedings of the 8th International Conference of Pattern Recognition Systems. 2017 Presented at: ICPRS '17; July 11-13, 2017; Madrid, Spain.
34. Creagh A, Lipsmeier F, Lindemann M, Vos MD. Interpretable deep learning for the remote characterisation of ambulation in multiple sclerosis using smartphones. Sci Rep 2021 Jul 12;11(1):14301 [FREE Full text] [doi: 10.1038/s41598-021-92776-x] [Medline: 34253769]
35. Fawaz H, Forestier G, Weber J, Idoumghar L, Muller PA. Transfer learning for time series classification. arXiv 2018.
36. Phan H, Chen OY, Koch P, Lu Z, McLoughlin I, Mertins A, et al. Towards more accurate automatic sleep staging via deep transfer learning. IEEE Trans Biomed Eng 2021 Jun;68(6):1787-1798. [doi: 10.1109/TBME.2020.3020381] [Medline: 32866092]
37. Perera P, Fierrez J, Patel V. Quickest intruder detection for multiple user active authentication. In: Proceedings of the 2020 IEEE International Conference on Image Processing. 2020 Presented at: ICIP '20; October 25-28, 2020; Abu Dhabi, United Arab Emirates.
38. Ulinskas M, Damaševičius R, Maskeliūnas R, Woźniak M. Recognition of human daytime fatigue using keystroke data. Procedia Comput Sci 2018;130:947-952. [doi: 10.1016/j.procs.2018.04.094]
39. de Jong M, Bonvanie AM, Jolij J, Lorist MM. Dynamics in typewriting performance reflect mental fatigue during real-life office work. PLoS One 2020 Oct 6;15(10):e0239984 [FREE Full text] [doi: 10.1371/journal.pone.0239984] [Medline: 33022017]
40. Al-Libawy H, Al-Ataby A, Al-Nuaimy W, Al-Taee MA, Al-Jubouri Q. Fatigue detection method based on smartphone text entry performance metrics. In: Proceedings of the 2016 9th International Conference on Developments in eSystems Engineering. 2016 Presented at: DeSE '16; August 31-September 02, 2016; Liverpool, UK. [doi: 10.1109/dese.2016.9]
41. Kryger M, Roth T, Dement WC. Principles and Practice of Sleep Medicine. 6th edition. Amsterdam, The Netherlands: Elsevier; 2011.
42. Monk TH. The post-lunch dip in performance. Clin Sports Med 2005 Apr;24(2):e15-xii. [doi: 10.1016/j.csm.2004.12.002] [Medline: 15892914]
43. Krupp LB, LaRocca NG, Muir-Nash J, Steinberg AD. The fatigue severity scale. Application to patients with multiple sclerosis and systemic lupus erythematosus. Arch Neurol 1989 Oct;46(10):1121-1123. [doi: 10.1001/archneur.1989.00520460115022] [Medline: 2803071]
44. Morrow GR, Andrews PL, Hickok JT, Roscoe JA, Matteson S. Fatigue associated with cancer and its treatment. Support Care Cancer 2002 Jul;10(5):389-398. [doi: 10.1007/s005200100293] [Medline: 12136222]
45. Wostyn P. COVID-19 and chronic fatigue syndrome: is the worst yet to come? Med Hypotheses 2021 Jan;146:110469 [FREE Full text] [doi: 10.1016/j.mehy.2020.110469] [Medline: 33401106]

Abbreviations

ADD: average detection delay
AFD: active fatigue detection
AUC: area under the curve
DML: distance metric learning
DNN: deep neuronal network
EER: equal error rate
k-NN: k-nearest neighbor
LOO: leave-one-out
nQCS: neuroQWERTY Crowdsource
nQSI: neuroQWERTY Sleep Inertia
PFD: probability of false detection
PND: probability of nondetection
RF: random forest
SVM: support vector machine

Edited by A Mavragani; submitted 12.07.22; peer-reviewed by A Papadopoulos, A Finny; comments to author 23.08.22; revised version received 07.10.22; accepted 19.10.22; published 21.11.22.

Please cite as:
Acien A, Morales A, Vera-Rodriguez R, Fierrez J, Mondesire-Crump I, Arroyo-Gallego T
Detection of Mental Fatigue in the General Population: Feasibility Study of Keystroke Dynamics as a Real-world Biomarker
JMIR Biomed Eng 2022;7(2):e41003
URL: https://biomedeng.jmir.org/2022/2/e41003
doi:10.2196/41003
PMID:38875698

Original Paper

Noncontact Longitudinal Respiratory Rate Measurements in Healthy Adults Using Radar-Based Sleep Monitor (Somnofy): Validation Study

Ståle Toften[1], BSc, MSc; Jonas T Kjellstadli[1], BSc, MSc, PhD; Ole Kristian Forstrønen Thu[2,3], MD, PhD; Ole-Johan Ellingsen[2], MSc

[1]Department of Data Science and Research, VitalThings AS, Tønsberg, Norway

[2]VitalThings AS, Tønsberg, Norway

[3]Department of Anesthesiology and Intensive Care Medicine, Levanger Hospital, Nord-Trøndelag Hospital Trust, Levanger, Norway

Corresponding Author:
Ståle Toften, BSc, MSc
Department of Data Science and Research
VitalThings AS
Jarlsøveien 48
Tønsberg, 3124
Norway
Phone: 47 47899717
Email: st@vitalthings.com

Abstract

Background: Respiratory rate (RR) is arguably the most important vital sign to detect clinical deterioration. Change in RR can also, for example, be associated with the onset of different diseases, opioid overdoses, intense workouts, or mood. However, unlike for most other vital parameters, an easy and accurate measuring method is lacking.

Objective: This study aims to validate the radar-based sleep monitor, Somnofy, for measuring RRs and investigate whether events affecting RR can be detected from personalized baselines calculated from nightly averages.

Methods: First, RRs from Somnofy for 37 healthy adults during full nights of sleep were extensively validated against respiratory inductance plethysmography. Then, the night-to-night consistency of a proposed filtered average RR was analyzed for 6 healthy participants in a pilot study in which they used Somnofy at home for 3 months.

Results: Somnofy measured RR 84% of the time, with mean absolute error of 0.18 (SD 0.05) respirations per minute, and Bland-Altman 95% limits of agreement adjusted for repeated measurements ranged from –0.99 to 0.85. The accuracy and coverage were substantially higher in deep and light sleep than in rapid eye movement sleep and wake. The results were independent of age, sex, and BMI, but dependent on supine sleeping position for some radar orientations. For nightly filtered averages, the 95% limits of agreement ranged from −0.07 to −0.04 respirations per minute. In the longitudinal part of the study, the nightly average was consistent from night to night, and all substantial deviations coincided with self-reported illnesses.

Conclusions: RRs from Somnofy were more accurate than those from any other alternative method suitable for longitudinal measurements. Moreover, the nightly averages were consistent from night to night. Thus, several factors affecting RR should be detectable as anomalies from personalized baselines, enabling a range of applications. More studies are necessary to investigate its potential in children and older adults or in a clinical setting.

(JMIR Biomed Eng 2022;7(2):e36618) doi:10.2196/36618

KEYWORDS

noncontact; monitoring; radar technology; respiratory rate; Somnofy; validation

Introduction

Background

Respiratory rate (RR) is arguably the most valuable parameter to detect clinical deterioration in hospital wards [1-3], and it is an important measure of health and wellness. Substantial change in RR can be associated with lower respiratory tract infections [4], fever [5,6], acute asthma [7], acute brain damage [8], opioid overdose [9], or exacerbation of chronic obstructive pulmonary disease (COPD) [10,11]. Other factors such as intense workouts [12], emotions or anxiety [13], and menstrual cycle [14] have also been shown to affect RR. A solution that can automatically, conveniently, and continuously monitor RR can have a range of applications.

In recent years, many new methods and devices have been developed to measure RR [15], but an accurate and simple method is still lacking [16]. Capnography is sometimes regarded as the gold standard [15,17], but hospitals still use manual counting of breaths [18], even though chest patches and under-the-mattress sensors are also available [19]. Chest patches derive RR from electrical cardiography by analyzing respiration-induced modulations on the heart signals, a technique also used in photoplethysmography in consumer wearables [20]. Studies on sleep use wearables such as thermistors, nasal pressure, and respiratory inductance plethysmography (RIP) to measure respiration. Although some of these technologies are accurate, they are unfortunately not suitable for longitudinal studies. It would be preferable for such a device to be noncontact and mobile and not need recharging or maintenance. As RRs vary during the day, measurements are often performed when the person is resting to obtain consistent measurements. To increase consistency, it can be advantageous to measure during nighttime, when the person sleeps and is unable to affect the measurements intentionally or unintentionally. Nocturnal RRs have also been shown to be an independent predictor of long-term mortality (RRs >16 respirations [breaths] per minute [RPM]) [21], and nocturnal hyperpnea is shown to be an indicator of periodic limb movement disorders [22]. Different under-the-mattress sensors have been investigated for this purpose [23-25], but radar technology is also an alternative.

Radar technology has been extensively studied for measuring RR [26-29] and even for detecting apnea [26,30-32]. However, most studies have measured RR only during optimized conditions where the participant is asked to sit or lie still [26] or during natural movements or sleep, but only for short periods [28,29]. A recent study validated RRs during full nights of sleep in both healthy individuals and patients with sleep apnea, but their study included only 6 healthy participants and the precision decreased significantly for participants with sleep apnea [27]. Moreover, their study did not analyze factors that may affect precision, such as body position (prone, supine, and side), sleep stage, or BMI or for how much of the night the radar was able to measure RR. There is still a need for more validation of radar technology for continuous monitoring during sleep. Furthermore, most studies dealing with RRs compare spot measurements with aggregate statistics that combine measurements from different people [33,34], even though there are large variations between individuals within the normal range [35,36]. Thus, the technology will be more useful if it can also be used to establish meaningful personalized baselines. The development and use of such baselines must be carefully considered, as RRs vary extensively even throughout the night. Thus, a person can easily seem ill in one moment and healthy in the next, when using standard spot measurements. For the technology to be able to reliably detect events affecting RR, it is vital that both the normal variations and measurement error for RR are significantly smaller than the effect of the event.

Objectives

The aim of this study was to investigate whether a commercially available radar-based sleep monitor, Somnofy (VitalThings), can be used as a longitudinal RR monitor. The first objective was to extensively benchmark RR measurements from Somnofy against those derived from RIP. This objective included both instantaneous measurements and filtered nightly averages, which are proposed as robust metrics in longitudinal monitoring of RR. The second objective was to analyze the night-to-night consistency of this nightly average in a pilot study over 3 months and investigate whether it can be possible to reliably detect events affecting RR as deviations from personalized baselines with this type of technology.

Methods

Participants and Data Sample

For the first part of the study, 55 volunteers from Norway were recruited to sleep 1 night at a sleep laboratory. The participants were recruited directly or through social media. The inclusion criterion was healthy adults aged >18 years. In total, 33% (18/55) of the individuals were later removed from the data set. Of these 18 individuals, 15 (83%) were excluded owing to indications of sleep-related disorders that can influence RR (sleep apnea and periodic limb movement disorder), whereas 3 (17%) were excluded owing to initial recording problems (the recordings lacked >2 hours of data). Thus, the first data set contained data from 1 night of sleep from 67% (37/55) of the healthy adults (21/37, 57% were women). The average age was 32.6 (SD 10.6) years, and they had an average BMI of 23.3 (SD 2.9) kg/m^2.

For the second part of the study, 6 Norwegian individuals (aged 11-81 years; n=2, 33% were women) were recruited to use Somnofy at home for 3 months. The inclusion criterion was the participants to be in good health, which meant that they did not have any disease that increased the possibility for hospitalization, causing discontinuity in the measurements.

Ethical Considerations

As this was not a clinical study and included only healthy participants, it was exempted from review in accordance with the Norwegian National Research Ethics Committee (reference number: "2019/995 A", June, 2019). Written informed consent was obtained from all participants in accordance with the principles embodied in the Declaration of Helsinki. All methods were performed in accordance with relevant guidelines and regulations.

Procedure

The participants in the first phase slept 1 night at the sleep laboratory in the Colosseum Clinic in Oslo, Norway. They were not allowed to consume alcohol or other drugs 48 hours before the assessments and could not smoke during the assessments. Full polysomnography (PSG) was performed to detect possible sleep disorders. The PSG data were also used to derive RRs, which were later compared with the output from Somnofy. In total, 2 Somnofy units recorded each night. One unit was placed on a nightstand to the left of the participant and the other unit was placed on the wall above the participant's head. Both units were aimed at the participant's chest. Overall, 4 participants lacked data from one of the sensors. Consequently, data from only 1 randomly selected sensor were used per participant (nightstand: 20/37, 54% and wall: 17/37, 46%). However, for the analyses specifically investigating the difference between the 2 sensor locations, both sensors were used, and the 4 participants with only 1 sensor location were dropped.

In the second part of the study, the participants were each given 1 Somnofy unit to take home, and they were instructed to place it on their nightstand. The adult participants (3/6, 50%) shared beds with their spouses. For these participants, the Somnofy distance parameter was set to a distance at the midpoint between the 2 individuals' thoraces in a normal sleeping position. No problems were detected with this setup, and no data were removed owing to disruption by the spouse. All participants were encouraged to live normally. During the study, 67% (4/6) of the participants experienced periods of self-reported illness. They did not recall the exact start or end times of these illnesses. As no illness occurred in the first half of the period, it was used to calculate personalized RR baselines for all participants (n=40 nights). Baselines were calculated as the average RR over the period, and 95% CIs were calculated as the baseline–1.96 × SD to baseline+1.96 × SD.

Somnofy

Somnofy (version 0.7; VitalThings) was used in this study. Somnofy uses an impulse radio ultrawideband radar with an average sampling rate of 23.8 GHz, which, through configuration, is sampled into a 3-m–long frame of 5-cm bins updated with a frequency of approximately 17 Hz. Somnofy measures humans by emitting signals that are reflected by the human body. If the body moves, it will affect the signals that are returned to the radar. The RR is further derived by using the Doppler effect and signal processing techniques, primarily the Fast Fourier transform (FT), to analyze periodic movements caused by the chest wall. The Fast FT is calculated every second for the last 20 seconds of the data, using a Hann window and 19-second overlap. Artifacts and harmonics are automatically removed by Somnofy; therefore, it provides only estimates that it is confident in. In this study, Somnofy was configured to provide RRs between 8 and 30 RPM. Movement is derived by analyzing the changes in the received radar signal over the last 6 seconds. The operating frequency enables the radar signal to travel through bedsheets and clothes before being reflected on the human body. More information on the principles of radar technology is available in previous studies [26].

Somnofy also calculates the nightly average RR. When calculating nightly averages, it is not necessary to use every instantaneous RR throughout the night. Using only selected measurements and filtering outliers can increase the night-to-night consistency, because the average does not depend on, for example, the amount of movement during the night. Therefore, Somnofy considers only periods without movement and rapid eye movement (REM) sleep, where RR tends to vary, when calculating nightly averages. In addition, outliers defined as >0.675 SD away from the mean are disregarded.

Somnofy is certified according to the Federal Communication Commission and "Conformité Européene" and harmless to human beings. Somnofy is installed by simply placing the unit on a nightstand or mounting it on a wall. Somnofy measures RR for 1 person and can do so despite the presence of 2 individuals in a bed, for which it measures only the nearest person if the person lies 5 cm closer to the radar than the other person. However, when 2 individuals are sharing a bed, the distance parameter in Somnofy should be set to a distance between the 2 individuals' thoraces to prevent the unit from starting to measure the other person when the intended participant exits the bed. Somnofy also collects additional information about the sleeping environment and scores sleep stages (accuracy to detect sleep=0.97 and accuracy to detect wake=0.72 in epoch-by-epoch analyses) [37]. For more details about Somnofy, refer to the validation study on sleep stage classification [37]. Currently, Somnofy is not a Food and Drug Administration–approved medical device.

PSG Recordings

PSG was performed using SOMNOscreen plus (SOMNOmedics) by sleep specialists following the guidelines of the American Academy of Sleep Medicine [38]. RRs were derived from 32 Hz RIP using the short-time FT (STFT) as implemented by the Python library *SciPy* (version 1.4.1). The STFT was calculated with a 20-second Hann window and a 19-second overlap, providing 1 measurement per second. For minimal noise, the RIP belt (thorax or abdomen) with the highest signal quality was used for each 20-second window to derive RR.

However, the RRs derived from RIP were still noisy when the input data quality was low. To remove this noise, all measurements that were more than double or less than half of the measurement in the previous second were disregarded. The RRs were filtered further by removing outliers, which were defined as measurements >1.96 SDs away from the mean of a 15-minute interval around the measurement.

Nightly average RRs for RIP were calculated using the same time stamps as those used by Somnofy. To synchronize the clocks in Somnofy and PSG, the cross-correlation between movement from Somnofy and PSG was maximized. Time in bed was defined as the time from lights out to lights on.

Statistical Analysis

As RRs were measured continuously during the night, the Bland-Altman method [39] was chosen as the main statistical tool to validate Somnofy against RIP [40]. In contrast to techniques that predefine an acceptable error margin, the

Bland-Altman limit of agreement can be considered across several applications, as the difference is quantified. For instantaneous measurements, the Bland-Altman limits of agreement were calculated per night, and on the combined data set, adjusting for multiple measurements per participant [41]. In addition, the mean absolute error (MAE) was used to measure the absolute deviance, the coefficient of determination (R^2) was calculated to measure the correlation between RIP and Somnofy during the night, and coverage (percentage of the time Somnofy provided measurements regardless of whether RIP provided measurements) and gap (longest time that passed without a Somnofy measurement) were analyzed to investigate the reliability and robustness of the device.

MAE and coverage for instantaneous RRs were analyzed across age, sex, BMI, sensor location, and sleeping position for significant differences. The null hypothesis was that there was no difference with α set to .05. Age (young adult or adult), sex (male or female), and BMI (normal weight or overweight) were analyzed using a 2-tailed, 2-sample, unpaired *t* test, as the sample sizes were <30. To avoid bias toward individual participants, analyses were performed on average values for each night, disregarding waking periods. Calculations were performed using Python (version 3.6.8) and the *SciPy* (version 1.4.1) library.

From the radars' point of view, sleeping position depended on sensor location. A total of 8 different position parameters were established as combinations of the four sleeping positions (supine, prone, left, and right) and the two sensor locations (nightstand and wall). For each night, combinations with <300 measurements were disregarded. Consequently, not all combinations were available for every night, because not all the participants slept in every sleeping position. As the data set was paired, had >2 levels, and was unbalanced, a linear mixed effects model was chosen to analyze statistical significance. For these analyses, only measurements taken during Somnofy-defined sleep were used. The output model was analyzed using the Tukey method to investigate individual pairwise relationships. The analyses were performed in R (version 3.6.3), using the *"lme4"* (version 1.1-23) and *"multcomp"* (version 1.4-13) packages.

Results

Data Statistics

Table 1 shows the age, sex, and BMI distribution of the participants in the validation part of the study, and Table 2 shows the relevant sleep and respiratory parameters. On average, the participants spent 8 (SD 0.7) hours in bed, with PSG-defined sleep efficiency of 85.3% (SD 8.3%). The average RIP RR ranged from 11 to 21.4 RPM. On average, 5.8% (SD 1.9%) of the data were removed per night owing to filtering of RIP noise. Of the 61,775 data points removed, most data were removed from PSG-defined wake (n=31,063, 50.28%), light sleep (n=19,137, 30.98%), and REM sleep (n=7,561, 12.24%). Figure 1 displays the noise filtering for RIP for the nights with the least, average, and most noise. The filter removed most outliers but did not remove the natural variations in RR that occur during wake and REM sleep. While removing time stamps with PSG artifacts, 3.42% (30,294/886,512) of the Somnofy data were removed.

Table 1. Age, sex, and BMI of the participants in the validation study (N=37).

Categories	Participants, n (%)	Female participants, n (%)	Age (years), mean (SD); range	BMI (kg/m^2), mean (SD); range
All	37 (100)	21 (57)	32.6 (10.6); 20-62	23.3 (2.9); 18.5-28.7
Normal weight[a]	25 (68)	15 (41)	31.8 (10.7); 22-62	21.7 (2); 18.5-24.7
Overweight[b]	12 (32)	6 (16)	34.4 (10.8); 20-55	26.6 (1.1); 25.2-28.7
Young adult[c]	22 (59)	13 (35)	25.6 (2.6); 20-29	22.3 (2.9); 18.5-28.7
Adult[d]	15 (41)	8 (22)	42.9 (9.6); 31-62	24.6 (2.3); 20.8-27.8

[a] 18.5≤BMI<25.

[b] BMI≥25.

[c] Aged <30 years.

[d] Aged ≥30 years.

Table 2. Sleep and respiratory parameters for the validation study (N=37).

Parameters	Values, mean (SD)	Values, range
PSG[a]—time in bed (hours)	8 (0.7)	6.8-10
PSG—sleep efficiency (%)	85.3 (8.3)	58.9-95.4
PSG—wake after sleep onset (minutes)	44.9 (33.8)	4-143.9
Somnofy—total sleep time (hours)	6.9 (0.9)	4.1-8.2
RIP[b]—noise removed (%)	5.8 (1.9)	3.1-12.5
RIP—average respiratory rate (RPM[c])	15.5 (2.1)	11-21.4
AHI[d]	1.1 (1.1)	0-3.8
PLMI[e]	1.3 (2.8)	0-13.8
ArI[f]	9.2 (3)	2.6-18.29

[a] PSG: polysomnography.

[b] RIP: respiratory inductance plethysmography.

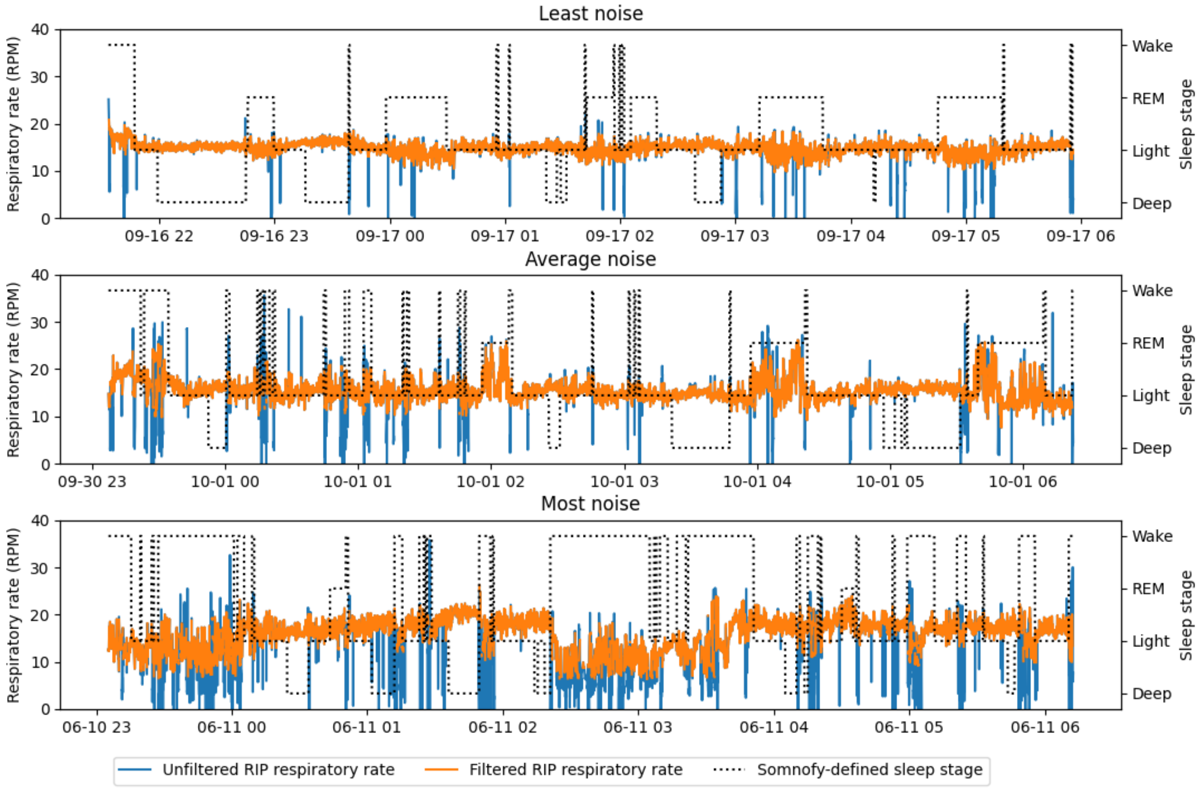
[c] RPM: respirations per minute.

[d] AHI: apneas and hypopneas per hour of sleep.

[e] PLMI: periodic limb movements per hour of sleep.

[f] ArI: arousals per hour of sleep.

Figure 1. Nights with the least, average, and most noise removed from respiratory rates derived from respiratory inductance plethysmography (RIP). The respiratory rates as respirations per minute (RPM) are displayed on the y-axis, and the date and time (mm-dd HH) are displayed on the x-axis. The filter removes obvious outliers in the respiratory rate without removing normal variations during wake and rapid eye movement (REM) sleep.



Instantaneous RR

The results of the measurements of instantaneous RR are displayed in Table 3. During time in bed, Somnofy managed to measure RR 84% (SD 6%) of the time (coverage) and the MAE of these measurements was 0.18 (SD 0.05) RPM compared with RIP. On average, the 95% limits of agreement with RIP ranged from −0.94 (SD 0.35) to 0.80 (SD 0.32) RPM, with bias of −0.07 (SD 0.02) RPM. After adjusting for repeated measurements, the limits of agreement on the whole data set ranged from −0.99 to 0.85. Figure 2 shows the Bland-Altman plot for this scenario. The orange regression line (slope=−0.0057; R^2=0.0059) indicates that Somnofy tends to underestimate RR compared with RIP and that Somnofy underestimates more for high RRs. During Somnofy-defined sleep, the coverage was 90% (SD 3.7%) and the limits of agreement ranged from −0.83 (SD 0.28) to 0.69 (SD 0.25), on average. In particular, the worst nights were improved by removing wake data, indicating that these nights had high amount of wake, which was difficult for Somnofy to measure.

Table 4 shows the coverage and accuracy across the different Somnofy-defined sleep stages. Somnofy was most accurate during deep sleep (non-REM 3) and light sleep (non-REM 1 or non-REM 2), whereas accuracy and coverage were substantially lower during wake and REM sleep than during other sleep stages. The results across PSG-defined sleep stages were similar (Multimedia Appendix 1).

Table 3. Results for instantaneous respiratory rate (N=37).

Parameters	During time in bed, mean (SD); range	During Somnofy-defined sleep, mean (SD); range
Number of measurements[a] (1000s)	28.66 (2.66); 24.48 to 36.07	24.69 (3.26); 14.88 to 29.67
Coverage[b] (%)	83.5 (6); 69 to 93.3	89.5 (3.7); 80.2 to 97.3
Longest gap[c] (minimum)	4.97 (3.80); 1.27 to 16.85	2.42 (1.77); 0.55 to 10
Number of common measurements[d] (1000s)	23.54 (2.90); 17.21 to 30.35	21.81 (3.14); 13 to 27.23
R^2[e]	0.89 (0.15); 0.03 to 0.96	0.90 (0.06); 0.73 to 0.97
MAE[f]	0.18 (0.04); 0.10 to 0.28	0.17 (0.04); 0.10 to 0.24
Bias	−0.07 (0.02); −0.14 to −0.04	−0.07 (0.02); −0.12 to −0.04
LoA[g]—low	−0.94 (0.35); −2.19 to −0.43	−0.83 (0.28); −1.57 to −0.43
LoA—high	0.80 (0.32); 0.32 to 1.90	0.69 (0.25); 0.32 to 1.36

[a] Number of instantaneous respiratory rate measurements in 1000s.

[b] Percentage of the time Somnofy provided respiratory rate measurements.

[c] Highest number of minutes between 2 Somnofy measurements per night.

[d] Number of times both Somnofy and noise-filtered respiratory inductance plethysmography provided measurement.

[e] R^2: coefficient of determination.

[f] MAE: mean absolute error.

[g] Bland-Altman 95% limits of agreement, calculated as bias – 1.96 × SD to bias + 1.96 × SD.

Figure 2. Bland-Altman plot for instantaneous respiratory rates. The y-axis displays the disagreement between Somnofy and respiratory inductance plethysmography (RIP; N=871,072), whereas the x-axis displays the average of Somnofy and RIP measurements. All values are presented as respirations per minute (RPM). Measurements from the same night are visualized with the same color. For overlapping measurements, the top measurements were picked randomly. The y-axis is limited to between –1.2 and 1.2.

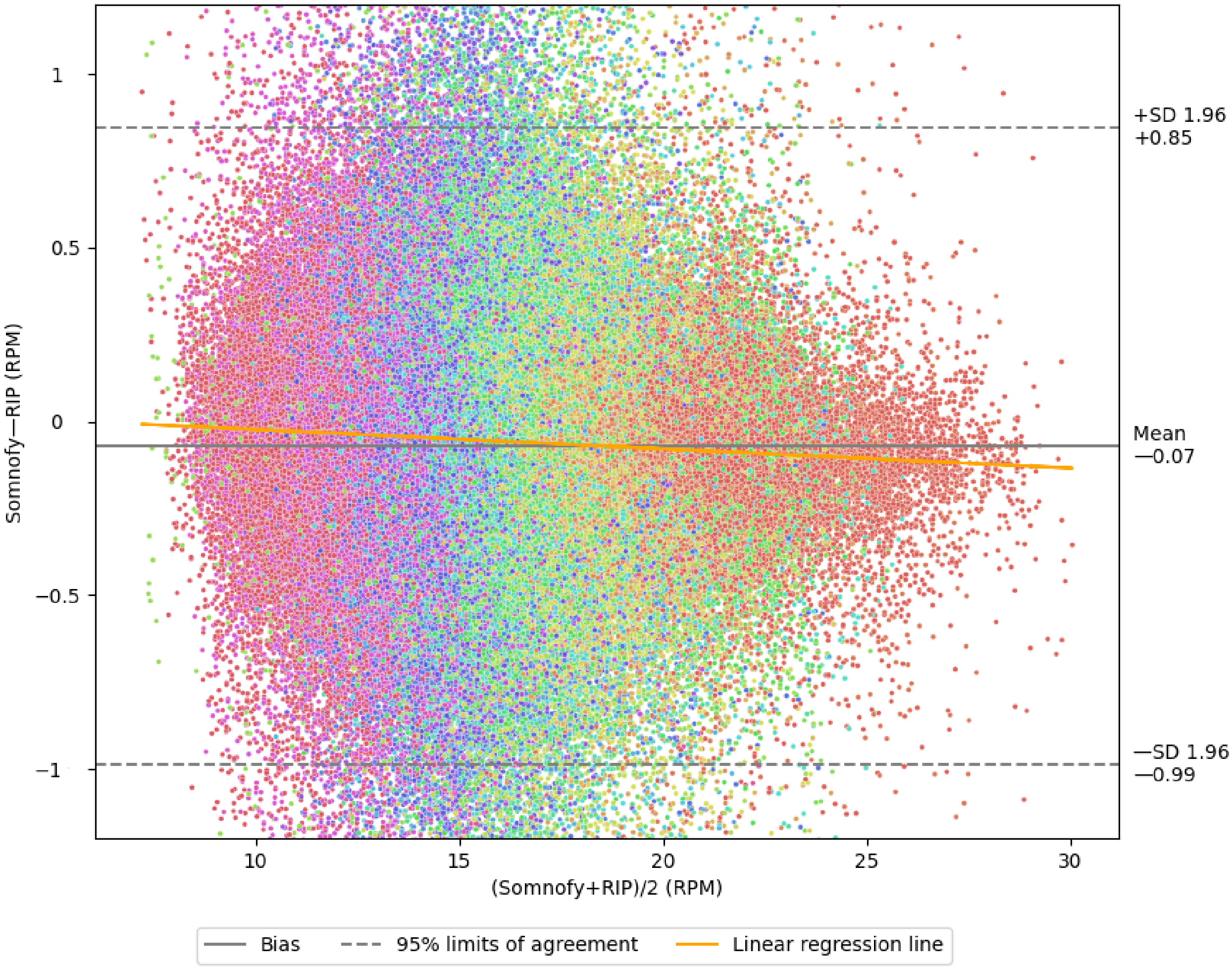


Table 4. Results for instantaneous respiratory rate for Somnofy-defined sleep stages.

Parameters	Wake	Light	Deep	Rapid eye movement
Number of measurements[a]	141,320	539,229	195,462	178,862
RIP[b]—average respiratory rate (RPM[c])	16.9	15.3	15.7	16.1
Coverage[d] (%)	47.1	91.3	98.1	75.3
Number of common measurements[e]	62,981	485,315	189,369	132,109
MAE[f]	0.33	0.15	0.11	0.30
Bias	−0.12	−0.06	−0.05	−0.12
LoA[g]—low	−1.97	−0.66	−0.38	−1.67
LoA—high	1.72	0.55	0.28	1.42

[a] Number of instantaneous respiratory rate measurements.

[b] RIP: respiratory inductance plethysmography.

[c] RPM: respirations per minute.

[d] Percentage of the time Somnofy provided respiratory rate measurements.

[e] Number of times both Somnofy and noise-filtered RIP provided measurement.

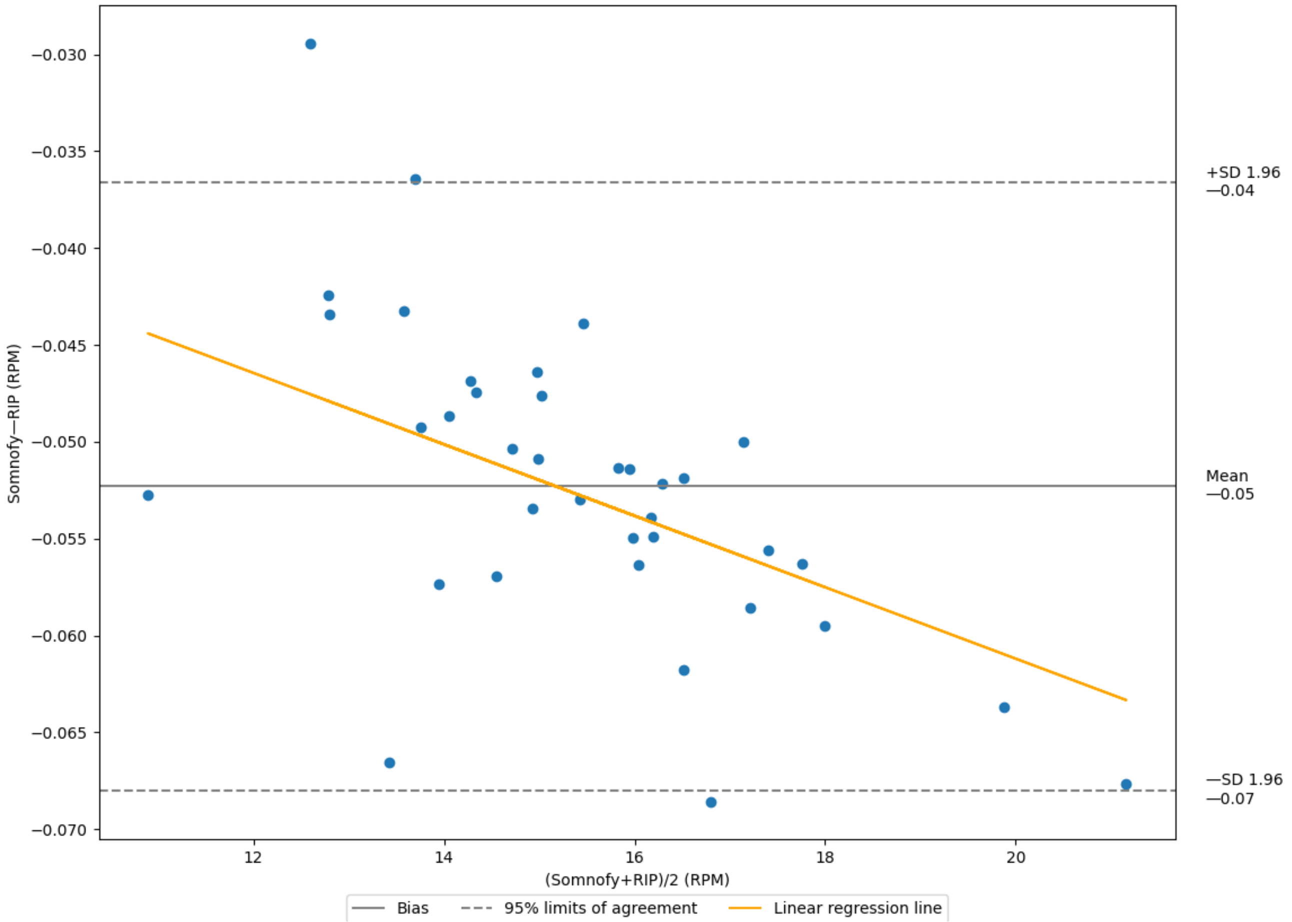
[f] MAE: mean absolute error.

[g] Bland-Altman 95% limits of agreement, adjusted for repeated measurements and calculated as bias – 1.96 × SD to bias + 1.96 × SD.

Nightly Average RR

For the nightly average RRs, MAE was 0.052 (SD 0.008) RPM. Figure 3 displays the Bland-Altman plot for these averages. The Bland-Altman limits of agreement show that 95% of the nightly averages are expected to have a disagreement with RIP between −0.07 and −0.04 RPM. As indicated by the orange line (slope=−0.0018; R^2=0.343), there seems to be a trend, where Somnofy underestimates more for high RRs. This trend is a similar to that for the instantaneous RR measurements shown in Figure 2.

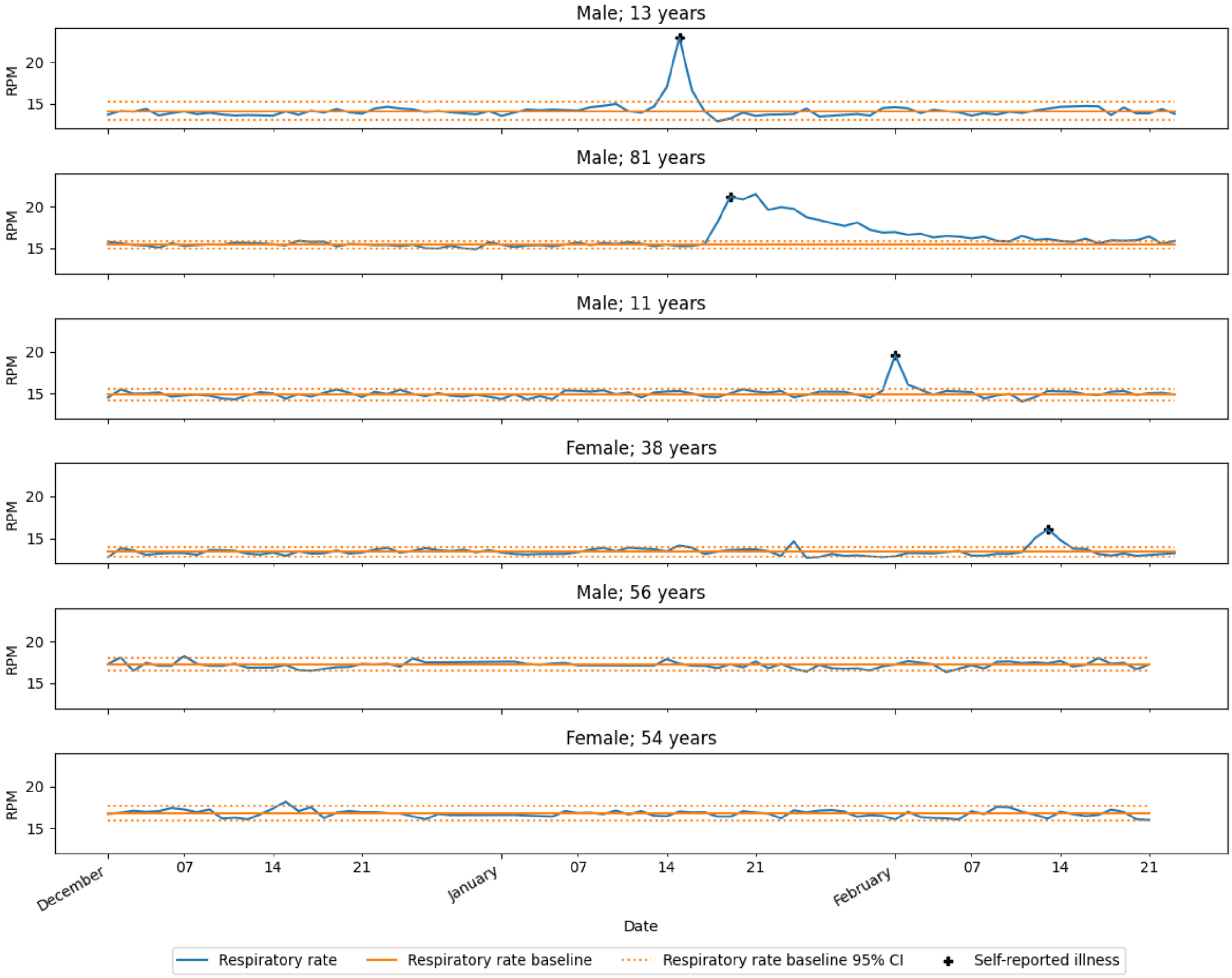
Figure 3. Bland-Altman analysis for nightly filtered respiratory rates. The y-axis displays the disagreement between Somnofy and respiratory inductance plethysmography (RIP; N=37), whereas the x-axis displays the average of Somnofy and RIP measurements. All values are presented as respirations per minute (RPM). The nightly averages are visualized as blue dots.



Night-to-Night Consistency of Nightly Average RR

The results from the longitudinal pilot study are shown in Figure 4. The RRs were fairly consistent from night to night for all participants (6/6, 100%), and most values were within the 95% CIs around the baselines. Moreover, periods with self-reported illness substantially deviated from their respective baselines. In total, 67% (4/6) of the participants reported 1 illness each. The participants who were aged 13, 11, and 38 years, respectively, reported a cold, and the participant aged 81 years reported an infection that was treated with antibiotics.

Figure 4. Nightly filtered average respiratory rates over 3 months. A total of 6 individual participants are labeled with sex and age above the corresponding graph. The y-axis displays the average filtered respiratory rates as respirations per minute (RPM), whereas the x-axis displays the date at wake up. Self-reported illness is tagged on the first peak of the respiratory rate during the illness. All nights substantially outside the CI were related to the self-reported illnesses.



Other Analyses

In this study, the null hypothesis that MAE and coverage were independent of age (MAE: *P*=.97 and coverage: *P*=.63), sex (MAE: *P*=.28 and coverage: *P*=.73), and BMI (MAE: *P*=.99 and coverage: *P*=.43) could not be rejected. In contrast, for sleeping position and sensor location, the null hypothesis was rejected. Some combinations, all including supine sleeping position, showed statistically significantly higher MAE (mean 0.045 RPM, SD 0.01) and statistically significantly lower coverage (mean 5.0%, SD 0.9%) than other radar or sleeping positions. The mean difference and *P* values for all the significantly different combinations are shown in Table 5.

Table 5. Statistically significant differences for sleeping position and sensor location[a].

Significantly different combinations	MAE[b] (RPM[c])		Coverage[d] (%)	
	Mean difference	*P* value	Mean difference	*P* value
Supine nightstand—left wall	0.048	<.001	−6.21	<.001
Supine nightstand—right wall	0.046	<.001	−5.99	<.001
Supine nightstand—left nightstand	0.052	.002	−4.78	<.001
Supine nightstand—right nightstand	0.042	<.001	−4.17	.002
Supine nightstand—prone wall	0.053	.02	N/A[e]	N/A
Supine wall—right wall	0.027	.04	−4.15	.002
Supine wall—left wall	N/A	N/A	−4.37	.001

[a] The table shows combinations of sleeping position (right, left, supine, and prone) and sensor location (nightstand and wall) that were statistically significantly different using Tukey method on a linear mixed effects model (33/37, 89%).

[b] MAE: mean absolute error.

[c] RPM: respirations per minute.

[d] Percentage of the time Somnofy provided respiratory rate measurements.

[e] N/A: not applicable.

Discussion

Principal Findings

This study demonstrated that Somnofy can accurately detect instantaneous RRs during time in bed for healthy adults. On average, Somnofy was able to measure RR 84% (SD 6%) of the time with MAE of 0.18 (SD 0.04) RPM. The Bland-Altman 95% limits of agreement ranged from −0.99 to 0.85 RPM. The accuracy and coverage varied significantly according to the sleep stage, where deep sleep (MAE=0.11; coverage=98%) was the most accurate, followed by light sleep (MAE=0.15; coverage=91%), REM sleep (MAE=0.30; coverage=75%), and wake (MAE=0.33; coverage=47%). For filtered nightly averages, the measurements from RIP and Somnofy were almost identical with Bland-Altman 95% limits of agreement, ranging from −0.07 to −0.04 RPM. Overall, Somnofy tended to slightly underestimate RR. Results were independent of age, BMI, and sex but were slightly worse for supine sleeping position.

The longitudinal part of the study showed that the nightly RRs seem fairly consistent from night to night. Most nightly averages were within the 95% CIs of the personalized baselines. Moreover, the CIs were smaller than, for example, the normal effect of 1-degree increase in body temperature on RR (eg, associated with fever) [5,6]. All substantial deviations coincided with self-reported illness, which were expected to increase the RR. The small night-to-night variations can be caused by other factors that affect RR, such as increased RRs after intense workouts [12] and RRs varying with emotions or anxiety [13] and owing to menstrual cycle [14]. A large study investigating these factors is necessary to understand the usefulness of investigating small deviations from baseline.

The coverage and accuracy varied substantially according to the sleep stage. It is probably easier for Somnofy to measure RR during light and deep sleep, as RR during these periods is more stable than that during wake or REM sleep. Coverage was particularly low during wake, when more movement likely resulted in more noise. Somnofy takes advantage of this by using values only from light and deep sleep to calculate the nightly average, during which the accuracy is higher. This has the additional benefit that the average is independent of the amount of wake and REM sleep during the night, during which RR varies more and tends to be higher. Thus, using values only from light and deep sleep should also improve the night-to-night consistency.

On average, 5.8% (SD 1.9%) of the PSG data had to be filtered for noise. Another reference device or signal processing technique could have been used instead of RIP with STFT. Unfortunately, there is no gold standard for longitudinal measurements of RR, and changing the reference is unlikely to affect the results significantly. RIP was used in one study [23], and RIP and STFT were used in another study [27]. One study used a nasal flow sensor, but had to remove 10 out of 40 nights owing to missing or unusable flow data [24]. Another study used both flow and effort signals with a peak detection algorithm, but based on the number of epochs they analyzed divided by the average time in bed in their study, they removed approximately twice as much noise as removed in this study [25].

Compared with wearables, radar technology has the benefit that nothing must be attached to the body, which can negatively affect sleep. Wearables also must be charged, and the user must remember to wear the device in bed. According to Fitbit, only half of its users wear their armbands at night [42], indicating that compliance can be a problem. Furthermore, neither wearables nor under-the-mattress sensors have been shown to measure sleep as reliably as Somnofy [37], information that can be used for more consistent nightly averages.

As Somnofy measures RRs from a distance, it was especially interesting to analyze the results across sleeping positions and sensor locations. All significant differences were found for supine sleeping position, indicating that this position may be more difficult to assess. For the nightstand sensor, the supine position may be more difficult, because the respiration movement is mainly perpendicular to the transmitted radar signals. This is also true for the prone position, but in this position, some of the respiration movements may be pushed in different directions by the bed. The wall sensor may have difficulties in assessing the supine position owing to more movement. Here, the body is free to move, and the sensor has good view of the movements. Interestingly, Somnofy seems to measure RR equally well when the chest or the back is aimed toward the sensor. In addition, there was no statistically significant difference according to age, sex, or BMI, which are factors that can affect both how respiration is visible on the body surface and actual RRs. However, the study had few participants with high BMI; therefore, this should be investigated further.

Previous studies have reported that active measurements of RRs can be imprecise if the user is aware of being measured, and therefore, consciously affects breathing [43]. As Somnofy measures RRs from a distance, it should be possible to do so without affecting the user. Moreover, measuring during sleep enforces measurements to occur during rest. Noncontact measurements of RRs should also have other benefits such as user-friendliness and less administration.

Comparison With Previous Studies

Few commercially available technologies suitable for longitudinal studies have been validated for RR measurements. Comparison between studies is always difficult, because the sleep data, reference device or signal processing, and performance metrics are different. However, to the best of our knowledge, the results of this study are significantly better than those of other technologies [23-25,44-46] including radar technology [27,29]. The validated measurements are also more instantaneous, as previous studies have averaged RRs over epochs of data. Moreover, previous studies of noncontact RR measurements during sleep have not investigated the effect of all the factors that can affect the results, such as sleep stage, age, BMI, and sleeping position [23-25,27,46]. This is also the first study that explicitly analyses coverage and measures gaps in continuous RR measurements during sleep using radar technology.

To the best of our knowledge, the accuracy in this study is also significantly higher than that of the measuring methods used in hospitals, such as manual counting of breaths [47] and chest patches [19]. However, these studies were performed on different populations and in different settings, which could have made measurements more difficult.

Previous studies have not validated filtered nightly average, as proposed in this study. This filtered average was substantially more accurate than the standard nightly averages reported in other studies [24,25]. In theory, it should also be more suitable for longitudinal studies, as the night-to-night variability should be lower.

No other study has analyzed whether anomaly detection from personalized baselines is a sound application of longitudinal RR monitoring. For the approach to be sensible, RRs need to be sufficiently consistent from night to night and the measurement error needs to be sufficiently small, depending on the application. However, a pilot study investigated a specific use case for patients with COPD [48]. Their intention was to detect exacerbation of COPD by comparing the median RR of one night with those of the previous nights. They concluded that RR can be obtained using radar technology and that RR may be an indicator of change in clinical status. Another study found that the use of both instantaneous and previous RRs improved precision when detecting clinical deterioration [49].

Limitations

The study investigated measurements only during rest, and the results cannot be automatically applied to general situations where the individual is awake. Furthermore, the study was limited to healthy adults. More studies are necessary to validate Somnofy for people with different illnesses, children, and older adults.

The second part of the study included few participants. A large population should be analyzed to investigate the night-to-night consistency of average nocturnal RRs in a general population. Moreover, the participants' illnesses were self-reported. No physicians were consulted for diagnosis, and body temperatures were not measured. Further studies should investigate which types of disease can be detected with this technology and how early in the development of these diseases the RR changes.

This study analyzed only RR. Although RR can be valuable to measure longitudinally, more value can be added by measuring other biomarkers such as temperature, blood pressure, and heart rate simultaneously. Heart rate is often measured using comparable technology [23-25], and radar technology has previously also been validated for measuring heart rate during sleep [27]. Heart rate measurement was not available from Somnofy at the time of this study.

Conclusions

This study shows that Somnofy accurately measures RR during sleep in healthy adults. To the best of our knowledge, Somnofy has higher precision than any other noncontact device suitable for longitudinal monitoring, especially for nightly averages. Moreover, measuring RRs during sleep seems to be a sound option for consistent longitudinal measurements. Several events that affect the RR should be detectable as deviations from personalized nocturnal baselines, making the device suitable for a broad range of applications. Further studies are necessary to validate the use of Somnofy for children and older adults or to use this device in clinical settings.

Acknowledgments

The authors wish to thank Tony Wader and Nikita Zhitniy at the Colosseum Clinic in Oslo for their contributions to polysomnography measurements and scoring.

Conflicts of Interest

Data collection was funded by VitalThings, the company that owns Somnofy. All authors work for VitalThings. OJE owns shares in VitalThings.

Multimedia Appendix 1
Instantaneous respiratory rates across PSG-defined sleep stages.
[DOCX File , 15 KB - biomedeng_v7i2e36618_app1.docx]

References

1. Churpek MM, Adhikari R, Edelson DP. The value of vital sign trends for detecting clinical deterioration on the wards. Resuscitation 2016 May;102:1-5 [FREE Full text] [doi: 10.1016/j.resuscitation.2016.02.005] [Medline: 26898412]
2. Churpek MM, Yuen TC, Huber MT, Park SY, Hall JB, Edelson DP. Predicting cardiac arrest on the wards: a nested case-control study. Chest 2012 May;141(5):1170-1176 [FREE Full text] [doi: 10.1378/chest.11-1301] [Medline: 22052772]
3. Mochizuki K, Shintani R, Mori K, Sato T, Sakaguchi O, Takeshige K, et al. Importance of respiratory rate for the prediction of clinical deterioration after emergency department discharge: a single-center, case-control study. Acute Med Surg 2017 Apr 10;4(2):172-178 [FREE Full text] [doi: 10.1002/ams2.252] [Medline: 29123857]
4. McFadden JP, Price RC, Eastwood HD, Briggs RS. Raised respiratory rate in elderly patients: a valuable physical sign. Br Med J (Clin Res Ed) 1982 Mar 27;284(6316):626-627 [FREE Full text] [doi: 10.1136/bmj.284.6316.626] [Medline: 6802262]
5. Nijman RG, Thompson M, van Veen M, Perera R, Moll HA, Oostenbrink R. Derivation and validation of age and temperature specific reference values and centile charts to predict lower respiratory tract infection in children with fever: prospective observational study. BMJ 2012 Jul 03;345(jul03 1):e4224 [FREE Full text] [doi: 10.1136/bmj.e4224] [Medline: 22761088]
6. Davies P, Maconochie I. The relationship between body temperature, heart rate and respiratory rate in children. Emerg Med J 2009 Sep 21;26(9):641-643. [doi: 10.1136/emj.2008.061598] [Medline: 19700579]
7. Papiris S, Kotanidou A, Malagari K, Roussos C. Clinical review: severe asthma. Crit Care 2002 Mar;6(1):30-44 [FREE Full text] [doi: 10.1186/cc1451] [Medline: 11940264]
8. North JB, Jennett S. Abnormal breathing patterns associated with acute brain damage. Arch Neurol 1974 Nov 01;31(5):338-344. [doi: 10.1001/archneur.1974.00490410086010] [Medline: 4411797]
9. Schiller EY, Goyal A, Mechanic OJ. Opioid Overdose. In: StatPearls. Treasure Island (FL): StatPearls Publishing; 2022.
10. Franciosi LG, Page CP, Celli BR, Cazzola M, Walker MJ, Danhof M, et al. Markers of exacerbation severity in chronic obstructive pulmonary disease. Respir Res 2006 May 10;7(1):74 [FREE Full text] [doi: 10.1186/1465-9921-7-74] [Medline: 16686949]
11. Yañez AM, Guerrero D, Pérez de Alejo R, Garcia-Rio F, Alvarez-Sala JL, Calle-Rubio M, et al. Monitoring breathing rate at home allows early identification of COPD exacerbations. Chest 2012 Dec;142(6):1524-1529. [doi: 10.1378/chest.11-2728] [Medline: 22797131]
12. Børsheim E, Bahr R. Effect of exercise intensity, duration and mode on post-exercise oxygen consumption. Sports Med 2003;33(14):1037-1060. [doi: 10.2165/00007256-200333140-00002] [Medline: 14599232]
13. Boiten FA, Frijda NH, Wientjes CJ. Emotions and respiratory patterns: review and critical analysis. Int J Psychophysiol 1994 Jul;17(2):103-128. [doi: 10.1016/0167-8760(94)90027-2]
14. N R H. Effects of different phases of menstrual cycle on respiration. J Evol Med Dent Sci 2014 Nov 26;3(65):14183-14188. [doi: 10.14260/jemds/2014/3899]
15. Liu H, Allen J, Zheng D, Chen F. Recent development of respiratory rate measurement technologies. Physiol Meas 2019 Aug 02;40(7):07TR01. [doi: 10.1088/1361-6579/ab299e] [Medline: 31195383]
16. Marjanovic N, Mimoz O, Guenezan J. An easy and accurate respiratory rate monitor is necessary. J Clin Monit Comput 2020 Apr 24;34(2):221-222. [doi: 10.1007/s10877-019-00357-1] [Medline: 31342305]
17. Frasca D, Geraud L, Charriere J, Debaene B, Mimoz O. Comparison of acoustic and impedance methods with mask capnometry to assess respiration rate in obese patients recovering from general anaesthesia. Anaesthesia 2015 Jan 10;70(1):26-31 [FREE Full text] [doi: 10.1111/anae.12799] [Medline: 25040754]
18. Badawy J, Nguyen OK, Clark C, Halm EA, Makam AN. Is everyone really breathing 20 times a minute? Assessing epidemiology and variation in recorded respiratory rate in hospitalised adults. BMJ Qual Saf 2017 Oct 26;26(10):832-836 [FREE Full text] [doi: 10.1136/bmjqs-2017-006671] [Medline: 28652259]
19. Breteler MJ, KleinJan EJ, Dohmen DA, Leenen LP, van Hillegersberg R, Ruurda JP, et al. Vital signs monitoring with wearable sensors in high-risk surgical patients: a clinical validation study. Anesthesiology 2020 Mar;132(3):424-439 [FREE Full text] [doi: 10.1097/ALN.0000000000003029] [Medline: 31743149]
20. Charlton PH, Birrenkott DA, Bonnici T, Pimentel MA, Johnson AE, Alastruey J, et al. Breathing rate estimation from the electrocardiogram and photoplethysmogram: a review. IEEE Rev Biomed Eng 2018;11:2-20. [doi: 10.1109/rbme.2017.2763681]

21. Baumert M, Linz D, Stone K, McEvoy RD, Cummings S, Redline S, et al. Mean nocturnal respiratory rate predicts cardiovascular and all-cause mortality in community-dwelling older men and women. Eur Respir J 2019 Jul 31;54(1):1802175 [FREE Full text] [doi: 10.1183/13993003.02175-2018] [Medline: 31151958]
22. Light M, Schmickl C, Owens RL, Pogach M, Thomas R. 1022 the “respiratory signature” of periodic leg movements – a potential way to track individual therapy response objectively. Sleep 2019 Apr;42(Supplement_1):A411. [doi: 10.1093/sleep/zsz069.1019]
23. Ranta J, Aittokoski T, Tenhunen M, Alasaukko-oja M. EMFIT QS heart rate and respiration rate validation. Biomed Phys Eng Express 2019 Jan 15;5(2):025016. [doi: 10.1088/2057-1976/aafbc8]
24. Yang R, Bendjoudi A, Buard N, Boutouyrie P. Pneumatic sensor for cardiorespiratory monitoring during sleep. Biomed Phys Eng Express 2019 Aug 23;5(5):055014. [doi: 10.1088/2057-1976/ab3ac9]
25. Siyahjani F, Garcia Molina G, Barr S, Mushtaq F. Performance evaluation of a smart bed technology against polysomnography. Sensors (Basel) 2022 Mar 29;22(7):2605 [FREE Full text] [doi: 10.3390/s22072605] [Medline: 35408220]
26. Khan F, Ghaffar A, Khan N, Cho SH. An overview of signal processing techniques for remote health monitoring using impulse radio UWB transceiver. Sensors (Basel) 2020 Apr 27;20(9):2479 [FREE Full text] [doi: 10.3390/s20092479] [Medline: 32349382]
27. Kang S, Lee Y, Lim Y, Park H, Cho SH, Cho SH. Validation of noncontact cardiorespiratory monitoring using impulse-radio ultra-wideband radar against nocturnal polysomnography. Sleep Breath 2020 Sep 10;24(3):841-848. [doi: 10.1007/s11325-019-01908-1] [Medline: 31401735]
28. Diraco G, Leone A, Siciliano P. A radar-based smart sensor for unobtrusive elderly monitoring in ambient assisted living applications. Biosensors (Basel) 2017 Nov 24;7(4):55 [FREE Full text] [doi: 10.3390/bios7040055] [Medline: 29186786]
29. Lauteslager T, Maslik M, Siddiqui F, Marfani S, Leschziner GD, Williams AJ. Validation of a new contactless and continuous respiratory rate monitoring device based on ultra-wideband radar technology. Sensors (Basel) 2021 Jun 11;21(12):4027 [FREE Full text] [doi: 10.3390/s21124027] [Medline: 34207961]
30. Zhou Y, Shu D, Xu H, Qiu Y, Zhou P, Ruan W, et al. Validation of novel automatic ultra-wideband radar for sleep apnea detection. J Thorac Dis 2020 Apr;12(4):1286-1295 [FREE Full text] [doi: 10.21037/jtd.2020.02.59] [Medline: 32395265]
31. Kang S, Kim D, Lee Y, Lim Y, Park H, Cho SH, et al. Non-contact diagnosis of obstructive sleep apnea using impulse-radio ultra-wideband radar. Sci Rep 2020 Mar 24;10(1):5261 [FREE Full text] [doi: 10.1038/s41598-020-62061-4] [Medline: 32210266]
32. Toften S, Kjellstadli JT, Tyvold SS, Moxness MH. A pilot study of detecting individual sleep apnea events using noncontact radar technology, pulse oximetry, and machine learning. J Sensors 2021 Jul 15;2021:1-9. [doi: 10.1155/2021/2998202]
33. Bleyer AJ, Vidya S, Russell GB, Jones CM, Sujata L, Daeihagh P, et al. Longitudinal analysis of one million vital signs in patients in an academic medical center. Resuscitation 2011 Nov;82(11):1387-1392. [doi: 10.1016/j.resuscitation.2011.06.033] [Medline: 21756971]
34. Rodríguez-Molinero A, Narvaiza L, Ruiz J, Gálvez-Barrón C. Normal respiratory rate and peripheral blood oxygen saturation in the elderly population. J Am Geriatr Soc 2013 Dec 12;61(12):2238-2240. [doi: 10.1111/jgs.12580] [Medline: 24329828]
35. Wallis LA, Healy M, Undy MB, Maconochie I. Age related reference ranges for respiration rate and heart rate from 4 to 16 years. Arch Dis Child 2005 Nov 01;90(11):1117-1121 [FREE Full text] [doi: 10.1136/adc.2004.068718] [Medline: 16049061]
36. Fleming S, Thompson M, Stevens R, Heneghan C, Plüddemann A, Maconochie I, et al. Normal ranges of heart rate and respiratory rate in children from birth to 18 years of age: a systematic review of observational studies. Lancet 2011 Mar 19;377(9770):1011-1018 [FREE Full text] [doi: 10.1016/S0140-6736(10)62226-X] [Medline: 21411136]
37. Toften S, Pallesen S, Hrozanova M, Moen F, Grønli J. Validation of sleep stage classification using non-contact radar technology and machine learning (Somnofy®). Sleep Med 2020 Nov;75:54-61. [doi: 10.1016/j.sleep.2020.02.022] [Medline: 32853919]
38. Berry RB, Brooks R, Gamaldo CE, Harding SM, Lloyd RM, Marcus CL, et al. The AASM Manual for the Scoring of Sleep and Associated Events Rules, Terminology and Technical Specifications. Darien, Illinois: American Academy of Sleep Medicine; 2015.
39. Martin Bland J, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986 Feb;327(8476):307-310. [doi: 10.1016/s0140-6736(86)90837-8]
40. Ranganathan P, Pramesh CS, Aggarwal R. Common pitfalls in statistical analysis: measures of agreement. Perspect Clin Res 2017;8(4):187-191 [FREE Full text] [doi: 10.4103/picr.PICR_123_17] [Medline: 29109937]
41. Bland JM, Altman DG. Agreement between methods of measurement with multiple observations per individual. J Biopharm Stat 2007;17(4):571-582. [doi: 10.1080/10543400701329422] [Medline: 17613642]
42. Goode L. Will Fitbit's sleep apnea tracking actually work? The Verge. 2017 Aug 30. URL: https://www.theverge.com/2017/8/30/16227040/fitbit-sleep-apnea-tracking-ionic-smartwatch-sensors [accessed 2021-08-10]
43. Hill A, Kelly E, Horswill MS, Watson MO. The effects of awareness and count duration on adult respiratory rate measurements: an experimental study. J Clin Nurs 2018 Mar 28;27(3-4):546-554. [doi: 10.1111/jocn.13861] [Medline: 28426897]

44. Sen-Gupta E, Wright D, Caccese J, Wright JA, Jortberg E, Bhatkar V, et al. A pivotal study to validate the performance of a novel wearable sensor and system for biometric monitoring in clinical and remote environments. Digit Biomark 2019 Mar 1;3(1):1-13 [FREE Full text] [doi: 10.1159/000493642] [Medline: 32095764]
45. Berryhill S, Morton C, Dean A, Berryhill A, Provencio-Dean N, Patel S, et al. Effect of wearables on sleep in healthy individuals: a randomized crossover trial and validation study. J Clin Sleep Med 2020 May 15;16(5):775-783 [FREE Full text] [doi: 10.5664/jcsm.8356] [Medline: 32043961]
46. Ben-Ari J, Zimlichman E, Adi N, Sorkine P. Contactless respiratory and heart rate monitoring: validation of an innovative tool. J Med Eng Technol 2010;34(7-8):393-398. [doi: 10.3109/03091902.2010.503308] [Medline: 20698739]
47. Latten GH, Spek M, Muris JW, Cals JW, Stassen PM. Accuracy and interobserver-agreement of respiratory rate measurements by healthcare professionals, and its effect on the outcomes of clinical prediction/diagnostic rules. PLoS One 2019;14(10):e0223155 [FREE Full text] [doi: 10.1371/journal.pone.0223155] [Medline: 31581207]
48. Ballal T, Heneghan C, Zaffaroni A, Boyle P, de Chazal P, Shouldice R, et al. A pilot study of the nocturnal respiration rates in COPD patients in the home environment using a non-contact biomotion sensor. Physiol Meas 2014 Dec 17;35(12):2513-2527. [doi: 10.1088/0967-3334/35/12/2513] [Medline: 25402668]
49. Akel M, Carey K, Winslow C, Churpek M, Edelson D. Less is more: detecting clinical deterioration in the hospital with machine learning using only age, heart rate, and respiratory rate. Resuscitation 2021 Nov;168:6-10 [FREE Full text] [doi: 10.1016/j.resuscitation.2021.08.024] [Medline: 34437996]

Abbreviations

COPD: chronic obstructive pulmonary disease
FT: Fourier transform
MAE: mean absolute error
PSG: polysomnography
REM: rapid eye movement
RIP: respiratory inductance plethysmography
RPM: respirations per minute
RR: respiratory rate
STFT: short-time Fourier transform

Edited by A Mavragani; submitted 20.01.22; peer-reviewed by S Akbas, F Siyahjani; comments to author 30.05.22; revised version received 21.06.22; accepted 23.07.22; published 12.08.22.

Please cite as:
Toften S, Kjellstadli JT, Thu OKF, Ellingsen OJ
Noncontact Longitudinal Respiratory Rate Measurements in Healthy Adults Using Radar-Based Sleep Monitor (Somnofy): Validation Study
JMIR Biomed Eng 2022;7(2):e36618
URL: https://biomedeng.jmir.org/2022/2/e36618
doi:10.2196/36618
PMID:38875674

Original Paper

High-Dimensional Analysis of Finger Motion and Screening of Cervical Myelopathy With a Noncontact Sensor: Diagnostic Case-Control Study

Takafumi Koyama[1*], MD, PhD; Ryota Matsui[2*], BEng; Akiko Yamamoto[1], MD; Eriku Yamada[1], MD; Mio Norose[1], MD; Takuya Ibara[3], PhD; Hidetoshi Kaburagi[1], MD, PhD; Akimoto Nimura[3], MD, PhD; Yuta Sugiura[2], PhD; Hideo Saito[2], PhD; Atsushi Okawa[1], MD, PhD; Koji Fujita[3], MD, PhD

[1]Department of Orthopaedic and Spinal Surgery, Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University, Tokyo, Japan
[2]School of Science for Open and Environmental Systems, Graduate School of Science and Technology, Keio University, Kanagawa, Japan
[3]Department of Functional Joint Anatomy, Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University, Tokyo, Japan
[*]these authors contributed equally

Corresponding Author:
Koji Fujita, MD, PhD
Department of Functional Joint Anatomy
Graduate School of Medical and Dental Sciences
Tokyo Medical and Dental University
1-4-5, Yushima, Bunkyo-ku
Tokyo, 113-8519
Japan
Phone: 81 3 5803 5279
Fax: 81 3 5803 5281
Email: fujiorth@tmd.ac.jp

Abstract

Background: Cervical myelopathy (CM) causes several symptoms such as clumsiness of the hands and often requires surgery. Screening and early diagnosis of CM are important because some patients are unaware of their early symptoms and consult a surgeon only after their condition has become severe. The 10-second hand grip and release test is commonly used to check for the presence of CM. The test is simple but would be more useful for screening if it could objectively evaluate the changes in movement specific to CM. A previous study analyzed finger movements in the 10-second hand grip and release test using the Leap Motion, a noncontact sensor, and a system was developed that can diagnose CM with high sensitivity and specificity using machine learning. However, the previous study had limitations in that the system recorded few parameters and did not differentiate CM from other hand disorders.

Objective: This study aims to develop a system that can diagnose CM with higher sensitivity and specificity, and distinguish CM from carpal tunnel syndrome (CTS), a common hand disorder. We then validated the system with a modified Leap Motion that can record the joints of each finger.

Methods: In total, 31, 27, and 29 participants were recruited into the CM, CTS, and control groups, respectively. We developed a system using Leap Motion that recorded 229 parameters of finger movements while participants gripped and released their fingers as rapidly as possible. A support vector machine was used for machine learning to develop the binary classification model and calculated the sensitivity, specificity, and area under the curve (AUC). We developed two models, one to diagnose CM among the CM and control groups (CM/control model), and the other to diagnose CM among the CM and non-CM groups (CM/non-CM model).

Results: The CM/control model indexes were as follows: sensitivity 74.2%, specificity 89.7%, and AUC 0.82. The CM/non-CM model indexes were as follows: sensitivity 71%, specificity 72.87%, and AUC 0.74.

Conclusions: We developed a screening system capable of diagnosing CM with higher sensitivity and specificity. This system can differentiate patients with CM from patients with CTS as well as healthy patients and has the potential to screen for CM in a variety of patients.

(JMIR Biomed Eng 2022;7(2):e41327) doi:10.2196/41327

KEYWORDS

cervical myelopathy; myelopathy; spinal cord disease; spinal cord disorder; nervous system disorder; nervous system disease; clumsiness; screening; 10-second hand grip and release test; machine learning; carpal tunnel syndrome; Leap Motion; clinical informatics; system validation; screening system; sensor; model; diagnosis; diagnostic; high-dimensional analysis; motion sensor; motion detection; high-dimensional data analysis; high-dimensional statistics

Introduction

Cervical myelopathy (CM) occurs in patients with cervical spondylotic myelopathy, ossification of the posterior longitudinal ligament, or cervical disk herniation [1-3]. CM causes symptoms such as clumsiness of the hands, numbness of the extremities and trunk, and gait disturbance, and often requires surgery. The longer the duration and the more severe the disease, the worse the postoperative outcome [4-6]. However, some patients with CM are unaware of their early symptoms and consult a spine surgeon only after their condition has become severe [7]. Therefore, screening and early diagnosis of CM are important for symptom monitoring and to determine the optimum time for surgery [8].

Clumsiness of hands is a characteristic and important symptom of CM and is referred to as myelopathy hand [9]. The 10-second hand grip and release (10-s) test is commonly used to check for the presence of myelopathy hand [9,10]. In the 10-s test, patients repeatedly grip and release their hand as fast as possible for 10 seconds; if the number of repetitions is less than 20, a myelopathy hand is suspected. The 10-s test is simple but would be more useful for screening if it could objectively evaluate not only the number of repetitions but also the changes in movement specific to myelopathy hand.

Nowadays, the latest commercial sensors and devices using virtual reality have been developed and are being used in the medical field [11]. Some studies have reported using smartphones and stylus pens to analyze hand movements and diagnose diseases [12-14]. In the field of cervical spine, there have been reports of diagnosis, surgery, and rehabilitation using virtual reality [15-17]. Several studies have also been conducted to analyze the movement of the myelopathy hand using sensors [18-22]. Most of these studies used wearable sensors such as motion capture systems, strain sensors, gyro sensors, and bend sensors, which are complicated.

For a simpler test, we analyzed hand and finger movements in the 10-s test using Leap Motion (Leap Motion) in a previous study [23]. Leap Motion is a noncontact sensor consisting of infrared cameras and LEDs, and captures hand and finger movements in real time [24,25]. Furthermore, we applied a machine learning algorithm to the obtained data to create a binary classification model to classify CM with 84% sensitivity, 60.7% specificity, and 0.85 area under the curve (AUC). However, because of the limitations of the system, only fingertip movements, not all joint movements, were recorded. Moreover, because only patients with CM and healthy participants were compared, it was not clear whether our model could differentiate CM from other hand disorders such as carpal tunnel syndrome (CTS).

To solve these problems, we improved the system so that the joints of each finger can also be recorded by Leap Motion and aimed to develop a system capable of diagnosing CM with higher sensitivity and specificity. Furthermore, we included patients with CTS, a common hand disorder, to verify if it is possible to distinguish CM from CTS.

Methods

Ethics Approval

This study was approved by the Institutional Review Board of Tokyo Medical and Dental University (M2019-047). Written informed consent was provided by all participants.

Recruitment

We included preoperative patients with CM (CM group), preoperative patients with CTS (CTS group), and volunteers (control group) between February 2020 and July 2021. Experienced spine surgeons diagnosed CM based on symptoms, physical and neurological findings, and magnetic resonance imaging (MRI) or computed tomography myelogram. Experienced hand surgeons diagnosed CTS based on symptoms, physical findings such as the Tinel sign and Phalen test, and nerve conduction studies (NCSs) measured by Neuropack X1 (Nihon Kohden). Volunteers were recruited from patients who had undergone total hip arthroplasty.

In all groups, participants with a history of other upper extremity disease, injury, or surgery; those with neurological diseases such as stroke, brain tumor, and traumatic brain injury; those with inflammatory diseases such as rheumatoid arthritis; those with dementia or psychiatric disease; and those who refused to participate were excluded. Moreover, spine surgeons also examined participants in the CTS and control groups, and excluded those with symptoms or physical findings suggestive of CM from the CTS and control groups. Similarly, hand surgeons examined participants in the CM and control groups, and excluded those with symptoms or physical findings suggestive of CTS from the CM and control groups.

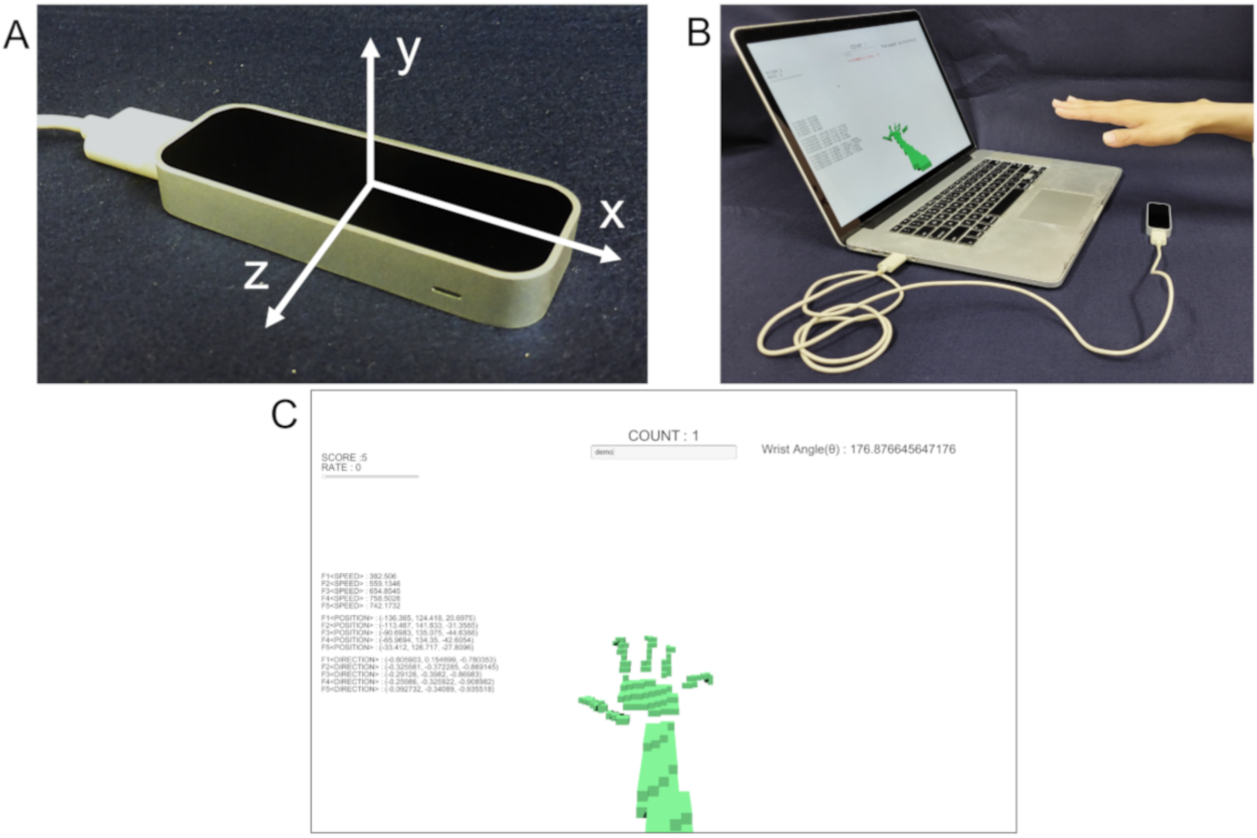
In the CM group, primary diseases causing CM were recorded. The maximally compressed levels of the spinal cord were also recorded from the sagittal and axial images of the preoperative T2-weighted MRI. In the CTS group, Bland classifications were recorded as severity based on NCSs [26]. Finally, the CTS and control groups were combined to create a non-CM group.

Measurements With Leap Motion

Before the measurement, the procedure and a short demonstration were provided to the participants. The protocol of the measurement with Leap Motion was based on a previous study and was performed as follows: participants sat in front of

Leap Motion placed in front of a laptop computer and connected by USB, extended the elbow on the side to be measured, placed the hand 10 cm above Leap Motion in a pronated position, and gripped and released the fingers as rapidly and as fully as possible 20 times after seeing the sign to start the examination (Figure 1) [23]. During the measurement, we confirmed that the system could correctly capture participant hand movements by watching the 3D hand model displayed on the screen in real time. All participants completed both hand measurements twice. A total of 229 parameters, listed in Table 1, were measured as waveform data (60 frames per second).

Figure 1. Images of the measurement with Leap Motion. Leap Motion and the three axes measured by Leap Motion (A). Participants placed their hand above Leap Motion, connected to a laptop computer via USB (B). During the measurement, a 3D hand model was displayed in real time on the screen of the laptop computer (C).



```
SCORE :5
RATE : 0
Wrist Angle(θ) : 176.876645647176

F1<SPEED> : 382.506
F2<SPEED> : 559.1346
F3<SPEED> : 654.8545
F4<SPEED> : 758.5026
F5<SPEED> : 742.1732

F1<POSITION> : (-136.365, 124.418, 20.6975)
F2<POSITION> : (-113.467, 141.833, -31.3585)
F3<POSITION> : (-90.6983, 135.075, -44.6388)
F4<POSITION> : (-65.9694, 134.35, -42.6054)
F5<POSITION> : (-33.412, 126.717, -27.8096)

F1<DIRECTION> : (-0.605903, 0.154699, -0.780353)
F2<DIRECTION> : (-0.325581, -0.372285, -0.869145)
F3<DIRECTION> : (-0.29126, -0.3982, -0.86983)
F4<DIRECTION> : (-0.25086, -0.325922, -0.908982)
F5<DIRECTION> : (-0.092732, -0.34089, -0.935518)
```

Table 1. Parameters measured by Leap Motion.

Parameters	Values, n	Total (N=229), n
Extended fingers (n)	1	1
Position of palm	3 dimensions[a]	3
Direction of palm	3 dimensions	3
Angle of wrist extension	1	1
Position of wrist	3 dimensions	3
Direction of forearm	3 dimensions	3
Speed of fingertip	5 fingers	5
Position of fingertip	5 fingers × 3 dimensions	15
Direction of fingertip	5 fingers × 3 dimensions	15
Position of distal end of bone	5 fingers × 4 bones[b] × 3 dimensions	60
Position of center of bone	5 fingers × 4 bones × 3 dimensions	60
Direction of bone	5 fingers × 4 bones × 3 dimensions	60

[a] Dimensions consist of x, y, and z coordinates.

[b] Bones consist of distal phalanx, middle phalanx, proximal phalanx, and metacarpus. For convenience, bones of the thumb were assumed to consist of distal phalanx, proximal phalanx, metacarpus, and carpal bones.

Statistical Analysis

Characteristics of Participants

The characteristics of participants were assessed using Student *t* test for age, chi-square test for sex and measured side of the hand, and Fisher exact test for hand dominance. A *P* value <.05 was considered statistically significant.

Binary Classification Model

We aimed to create two models, one to diagnose CM among the CM and control groups (CM/control model), and the other to diagnose CM among the CM and non-CM groups (CM/non-CM model).

Preprocessing of the data was performed prior to the application of machine learning. First, each waveform data was divided into 15 segments of 64 frames each while allowing for overlap because each participant took different frames to perform 20 grips and releases. These segments (64 frames) were linearly detrended and multiplied by the Hanning window function [27]. The processed segments were converted to frequency domain data using fast Fourier transform. The subwaveforms (64 frames) were converted into frequency domain data, selecting only the lower 16 frequencies. Finally, a 54,960-dimensional data set (229 parameters × 16 frequency domain data × 15 segments) was obtained for each trial. Data from two trials on each hand were combined and used to create the CM/control model. Alternatively, since CTS can occur on only one hand, data from only two trials on one hand (either the right or left) were combined and used to create the CM/non-CM model.

A support vector machine (SVM) was used to create the binary classification models [28]. SVM is one of the common machine learning algorithms used for classification and has performed well in previous studies. After the learning phase, the SVM shows a predicted label of CM with a probability score. We set a threshold and created a binary classification model to classify whether a data set was CM or not. Data from the CM and control groups were used for the CM/control model, and data from all groups were used for the CM/non-CM model. In the validation phase, 10-fold cross-validation was performed [29]. We generated a receiver operating characteristic (ROC) curve by adjusting the threshold and calculating the AUC. The point on the ROC curve closest to the upper-left corner of the graph was set as the optimal cutoff value.

Furthermore, to investigate which parts of the hand contribute to the diagnosis of CM, we also generated modified CM/control models using data from only one of the 20 bones and then similarly calculated the AUC.

Results

Comparison of Characteristics of Participants

In total, 31 participants (62 hands), 27 participants (38 hands), and 29 participants (58 hands) were recruited to the CM, CTS, and control groups, respectively. Patient demographics and characteristics are summarized in Table 2. There was no significant difference between the groups in terms of age, sex, or hand dominance.

Table 2. Characteristics of participants in the CM, CTS, and control groups.

Characteristic	Non-CM[a]		CM	*P* value	
	Control	CTS[b]		CM/control	CM/non-CM
Participants, n	29	25	31	N/A[c]	N/A
Age (years), mean (SD)	63.6 (52.1-75.0)	62.0 (49.2-74.7)	67.0 (57.0-77.0)	.23	.11
Sex (male), n	12	5	16	.59	.11
Hand dominance (right), n	29	25	30	>.99	.36
Hands, n	58	34	62	N/A	N/A
Side (right), n	29	20	31	>.99	.83
Bland classification, n	N/A		N/A	N/A	N/A
Grade 1		3			
Grade 2		0			
Grade 3		17			
Grade 4		0			
Grade 5		14			
Grade 6		4			
Primary disease, n	N/A	N/A		N/A	N/A
CSM[d]			13		
OPLL[e]			16		
CDH[f]			2		
Maximally compressed level, n	N/A	N/A		N/A	N/A
C1/2			1		
C2/3			0		
C3/4			12		
C4/5			8		
C5/6			9		
C6/7			1		

[a] CM: cervical myelopathy.

[b] CTS: carpal tunnel syndrome.

[c] N/A: not applicable.

[d] CSM: cervical spondylotic myelopathy.

[e] OPLL: ossification of the posterior longitudinal ligament.

[f] CDH: cervical disk herniation.

Binary Classification Model

The indexes of the binary classification models are listed in Table 3. The ROC curve of the control and CM/non-CM model are shown in Figure 2.

The AUC of models limited to the parameters of each bone are listed in Table 4. The AUC of the model using the parameters of the proximal phalanx of the thumb was the highest (0.86).

Table 3. Index of binary classification models.

	Sensitivity (%)	Specificity (%)	AUC[a]
CM[b]/control model	74.2	89.7	0.82
CM/non-CM model			
Total	71.0	72.8	0.74
Right hand	71.0	75.5	0.77
Left hand	74.2	79.1	0.76

[a] AUC: area under the curve.

[b] CM: cervical myelopathy.

Figure 2. Receiver operating characteristic (ROC) curve of the cervical myelopathy (CM)/control model (A) and CM/non-CM model (B). The area under the ROC curve was 0.82 and 0.74 in the CM/control model and CM/non-CM model, respectively. The red cross indicates the optimal cutoff value.

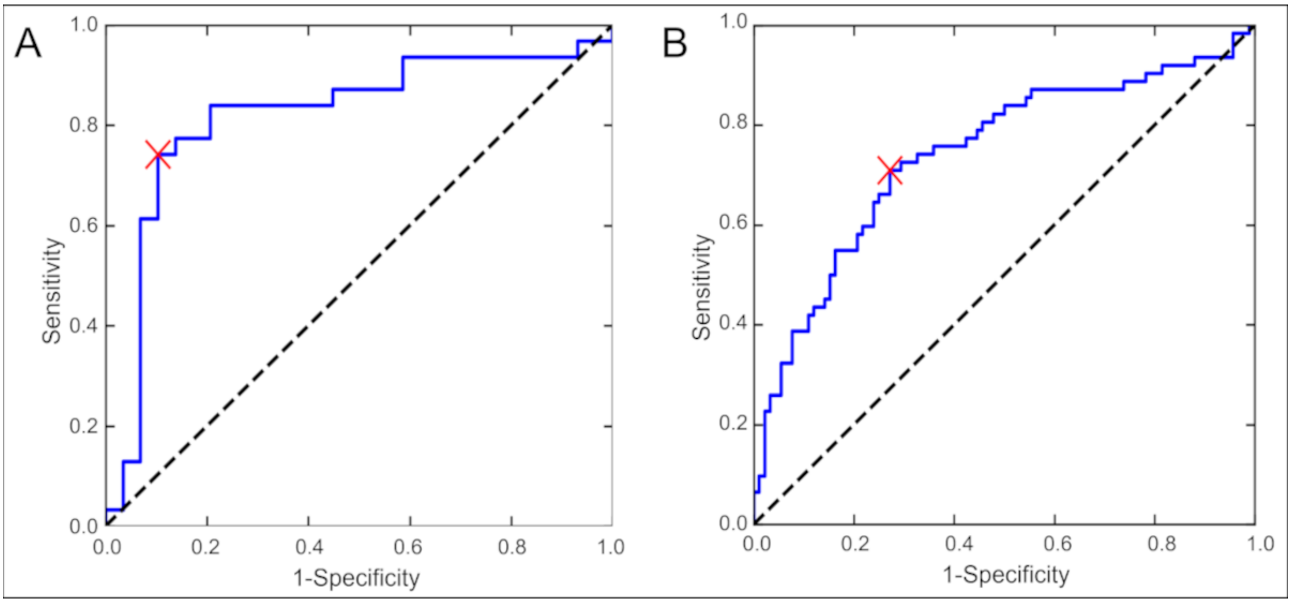


Table 4. Area under the curve of models limited to the parameters of each bone

	Thumb[a]	Index finger	Middle finger	Ring finger	Little finger
Distal phalanx	0.83	0.82	0.80	0.78	0.78
Middle phalanx	0.86	0.83	0.81	0.80	0.79
Proximal phalanx	0.84	0.82	0.84	0.83	0.83
Metacarpus	0.82	0.82	0.83	0.83	0.82

[a] Only in the thumb, middle phalanx means proximal phalanx, proximal phalanx means metacarpus, and metacarpus means carpal bones.

Discussion

Principal Results

We developed a classification model with high sensitivity and specificity to diagnose CM. However, despite increasing the parameters, major improvements in diagnostic performance of the CM/control model were not obtained in this study (74.2% sensitivity, 89.7% specificity, and 0.82 AUC) compared to the previous study (84% sensitivity, 60.7% specificity, and 0.85 AUC) [23]. Increasing only the number of parameters will result in improved diagnostic performance; therefore, it is necessary to increase the number of samples. Nevertheless, the classification model in this study is still effective as a screening method since it has a sufficiently high diagnostic performance when compared to classic tests. For example, the 10-s test showed 61%-74% sensitivity, 52%-66% specificity, and 0.71-0.77 AUC [10,30,31]; the finger escape sign showed 48%-55% sensitivity [30,31]; the deep tendon reflex change showed 15%-56% sensitivity and 96%-98% specificity [31-33]. In another previous study, the analysis and diagnoses of myelopathy hand was performed by wearing a glove with a sensor, with 87% sensitivity, 86% specificity, and 0.93 AUC [21]. Although the result of this study is inferior to the previous study, our method is superior in that it is easier to test many patients with the noncontact sensor, making it suitable for screening.

In the models limited to the parameters of each bone, the AUC of the model using the parameters of the proximal phalanx of the thumb was the highest. In addition, overall, the models using the parameters of bones of the thumb tended to have higher AUCs. This result is contrary to the finger escape sign, which indicates that the ulnar finger is more likely to be affected in CM [9]. The cause of this discrepancy may be due to the position of the sensor in this method. Because Leap Motion captures hand movement from the palmar side, the bones of the fingers other than the thumb are temporarily hidden by other bones during the grip and release movements, and occasionally not accurately captured. Alternatively, thumb movement is always tracked by Leap Motion. Moreover, another study reported that patients with CM exhibit specific changes in pinching movements with the thumb and index finger [20]. This result means that, in patients with CM, not only ulnar but also radial finger movements are significantly altered. These factors would contribute to the higher AUCs of the model using the parameters of the proximal phalanx of the thumb.

In this study, we attempted to differentiate the CM group from not only the control group, as in the previous study [23], but also the CTS (non-CM) group, and we achieved high diagnostic performance. The peak onset of CM is between the years of 40 and 60 years [1,3], but other hand disorders are also prevalent during that time. Because CTS is a common hand disorder, with a predilection for people 40 years or older [34,35], we included these patients in our study. Our system can distinguish myelopathy hand from motor disorders of the thumb that can occur in CTS [36]. While further trials are required to differentiate CM from other hand disorders, this result suggests the possibility of accurately screening for CM among a variety of hand disorders.

Several studies have also been conducted to analyze the movement of the myelopathy hand using sensors, but Leap Motion has the major advantage of simplicity. For example, motion captures can provide a detailed motion analysis, but the installation of the sensors requires skill and time of the examiners, and it is impossible to test a large number of patients in a short period of time. Alternatively, Leap Motion can be used for our test simply by connecting it to a computer if the program can be shared. Furthermore, the test can be performed by a single patient with only a simple test procedure guide. Leap Motion is also a less expensive commercial sensor, which is an advantage in that it is readily available. These advantages of Leap Motion are useful for screening large numbers of patients in a short period of time.

Limitations

This study had some limitations. First, it is possible that there were participants with potential CM in the CTS and control groups because participants in these groups did not undergo an MRI. Similarly, it is possible that there were participants with potential CTS in the CM and control groups because participants in these groups did not undergo an NCS.

Second, we did not compare subgroups by anatomical level of myelopathy and by severity of CM and CTS. There may be variation among subgroups within the same group. Third, only internal validation by 10-fold cross validation was performed and external validation was not. In future work, we will collect more samples to solve these problems.

Conclusions

We developed a screening system capable of diagnosing CM with higher sensitivity and specificity by high-dimensional analysis of finger motion and machine learning. This system can differentiate patients with CM from patients with CTS as well as healthy patients and has the potential to screen for CM in a variety of patients.

Acknowledgments

This research was supported by JST AIP-PRISM JPMJCR18Y2 and JST PRESTO JPMJPR17J4. We would like to thank Editage for English language editing.

Conflicts of Interest

None declared.

References

1. Nouri A, Tetreault L, Singh A, Karadimas SK, Fehlings MG. Degenerative cervical myelopathy: epidemiology, genetics, and pathogenesis. Spine (Phila Pa 1976) 2015 Jun 15;40(12):E675-E693. [doi: 10.1097/BRS.0000000000000913] [Medline: 25839387]
2. Rao R. Neck pain, cervical radiculopathy, and cervical myelopathy: pathophysiology, natural history, and clinical evaluation. J Bone Joint Surg Am 2002 Oct;84(10):1872-1881. [doi: 10.2106/00004623-200210000-00021] [Medline: 12377921]
3. Theodore N. Degenerative cervical spondylosis. N Engl J Med 2020 Jul 09;383(2):159-168. [doi: 10.1056/NEJMra2003558] [Medline: 32640134]
4. Tanaka J, Seki N, Tokimura F, Doi K, Inoue S. Operative results of canal-expansive laminoplasty for cervical spondylotic myelopathy in elderly patients. Spine (Phila Pa 1976) 1999 Nov 15;24(22):2308-2312. [doi: 10.1097/00007632-199911150-00004] [Medline: 10586453]

5. Tetreault L, Kopjar B, Côté P, Arnold P, Fehlings MG. A clinical prediction rule for functional outcomes in patients undergoing surgery for degenerative cervical myelopathy: analysis of an international prospective multicenter data set of 757 subjects. J Bone Joint Surg Am 2015 Dec 16;97(24):2038-2046. [doi: 10.2106/JBJS.O.00189] [Medline: 26677238]
6. Tetreault LA, Kopjar B, Vaccaro A, Yoon ST, Arnold PM, Massicotte EM, et al. A clinical prediction model to determine outcomes in patients with cervical spondylotic myelopathy undergoing surgical treatment: data from the prospective, multi-center AOSpine North America study. J Bone Joint Surg Am 2013 Sep 18;95(18):1659-1666. [doi: 10.2106/JBJS.L.01323] [Medline: 24048553]
7. Sadasivan KK, Reddy RP, Albright JA. The natural history of cervical spondylotic myelopathy. Yale J Biol Med 1993;66(3):235-242. [Medline: 8209559]
8. Kobayashi H, Kikuchi S, Otani K, Sekiguchi M, Sekiguchi Y, Konno S. Development of a self-administered questionnaire to screen patients for cervical myelopathy. BMC Musculoskelet Disord 2010 Nov 22;11:268 [FREE Full text] [doi: 10.1186/1471-2474-11-268] [Medline: 21092213]
9. Ono K, Ebara S, Fuji T, Yonenobu K, Fujiwara K, Yamashita K. Myelopathy hand. New clinical signs of cervical cord damage. J Bone Joint Surg Br 1987 Mar;69(2):215-219. [doi: 10.1302/0301-620X.69B2.3818752] [Medline: 3818752]
10. Machino M, Ando K, Kobayashi K, Morozumi M, Tanaka S, Ito K, et al. Cut off value in each gender and decade of 10-s grip and release and 10-s step test: a comparative study between 454 patients with cervical spondylotic myelopathy and 818 healthy subjects. Clin Neurol Neurosurg 2019 Sep;184:105414. [doi: 10.1016/j.clineuro.2019.105414] [Medline: 31306894]
11. Patel V, Chesmore A, Legner CM, Pandey S. Trends in workplace wearable technologies and connected ‐ worker solutions for next ‐ generation occupational safety, health, and productivity. Adv Intelligent Syst 2021 Sep 23;4(1):2100099. [doi: 10.1002/aisy.202100099]
12. Koyama T, Sato S, Toriumi M, Watanabe T, Nimura A, Okawa A, et al. A screening method using anomaly detection on a smartphone for patients with carpal tunnel syndrome: diagnostic case-control study. JMIR Mhealth Uhealth 2021 Mar 14;9(3):e26320 [FREE Full text] [doi: 10.2196/26320] [Medline: 33714936]
13. Pan D, Dhall R, Lieberman A, Petitti DB. A mobile cloud-based Parkinson's disease assessment system for home-based monitoring. JMIR Mhealth Uhealth 2015 Mar 26;3(1):e29 [FREE Full text] [doi: 10.2196/mhealth.3956] [Medline: 25830687]
14. Watanabe T, Koyama T, Yamada E, Nimura A, Fujita K, Sugiura Y. The accuracy of a screening system for carpal tunnel syndrome using hand drawing. J Clin Med 2021 Sep 27;10(19):4437 [FREE Full text] [doi: 10.3390/jcm10194437] [Medline: 34640454]
15. Burström G, Nachabe R, Persson O, Edström E, Elmi Terander A. Augmented and virtual reality instrument tracking for minimally invasive spine surgery: a feasibility and accuracy study. Spine (Phila Pa 1976) 2019 Aug 01;44(15):1097-1104. [doi: 10.1097/BRS.0000000000003006] [Medline: 30830046]
16. Sarig Bahat H, Croft K, Carter C, Hoddinott A, Sprecher E, Treleaven J. Remote kinematic training for patients with chronic neck pain: a randomised controlled trial. Eur Spine J 2018 Jun;27(6):1309-1323. [doi: 10.1007/s00586-017-5323-0] [Medline: 29018956]
17. Sarig-Bahat H, Weiss PL, Laufer Y. Cervical motion assessment using virtual reality. Spine (Phila Pa 1976) 2009 May 01;34(10):1018-1024. [doi: 10.1097/BRS.0b013e31819b3254] [Medline: 19404177]
18. Oess NP, Wanek J, Curt A. Design and evaluation of a low-cost instrumented glove for hand function assessment. J Neuroeng Rehabil 2012 Jan 17;9:2 [FREE Full text] [doi: 10.1186/1743-0003-9-2] [Medline: 22248160]
19. Omori M, Shibuya S, Nakajima T, Endoh T, Suzuki S, Irie S, et al. Hand dexterity impairment in patients with cervical myelopathy: a new quantitative assessment using a natural prehension movement. Behav Neurol 2018;2018:5138234. [doi: 10.1155/2018/5138234] [Medline: 30073036]
20. Sakai N. Finger motion analysis of the patients with cervical myelopathy. Spine (Phila Pa 1976) 2005 Dec 15;30(24):2777-2782. [doi: 10.1097/01.brs.0000190452.33258.72] [Medline: 16371902]
21. Su X, Hou C, Shen B, Zhang W, Wu D, Li Q, et al. Clinical application of a new assessment tool for myelopathy hand using virtual reality. Spine (Phila Pa 1976) 2020 Dec 15;45(24):E1645-E1652. [doi: 10.1097/BRS.0000000000003696] [Medline: 32947494]
22. Date S, Nakanishi K, Fujiwara Y, Yamada K, Kamei N, Kurumadani H, et al. Quantitative evaluation of abnormal finger movements in myelopathy hand during the grip and release test using gyro sensors. PLoS One 2021;16(10):e0258808 [FREE Full text] [doi: 10.1371/journal.pone.0258808] [Medline: 34669751]
23. Koyama T, Fujita K, Watanabe M, Kato K, Sasaki T, Yoshii T, et al. Cervical myelopathy screening with machine learning algorithm focusing on finger motion using noncontact sensor. Spine (Phila Pa 1976) 2022 Jan 15;47(2):163-171. [doi: 10.1097/BRS.0000000000004243] [Medline: 34593737]
24. Guna J, Jakus G, Pogačnik M, Tomažič S, Sodnik J. An analysis of the precision and reliability of the leap motion sensor and its suitability for static and dynamic tracking. Sensors (Basel) 2014 Feb 21;14(2):3702-3720 [FREE Full text] [doi: 10.3390/s140203702] [Medline: 24566635]
25. Weichert F, Bachmann D, Rudak B, Fisseler D. Analysis of the accuracy and robustness of the leap motion controller. Sensors (Basel) 2013 May 14;13(5):6380-6393 [FREE Full text] [doi: 10.3390/s130506380] [Medline: 23673678]

26. Bland JD. A neurophysiological grading scale for carpal tunnel syndrome. Muscle Nerve 2000 Aug;23(8):1280-1283. [doi: 10.1002/1097-4598(200008)23:8<1280::aid-mus20>3.0.co;2-y] [Medline: 10918269]
27. Harris F. On the use of windows for harmonic analysis with the discrete Fourier transform. Proc IEEE 1978 Jan;66(1):51-83. [doi: 10.1109/proc.1978.10837]
28. Noble WS. What is a support vector machine? Nat Biotechnol 2006 Dec;24(12):1565-1567. [doi: 10.1038/nbt1206-1565] [Medline: 17160063]
29. Cawley GC, Talbot NL. Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers. Pattern Recognition 2003 Nov;36(11):2585-2592. [doi: 10.1016/S0031-3203(03)00136-5]
30. Findlay GFG, Balain B, Trivedi JM, Jaffray DC. Does walking change the Romberg sign? Eur Spine J 2009 Oct;18(10):1528-1531 [FREE Full text] [doi: 10.1007/s00586-009-1008-7] [Medline: 19387702]
31. Wong TM, Leung HB, Wong WC. Correlation between magnetic resonance imaging and radiographic measurement of cervical spine in cervical myelopathic patients. J Orthop Surg (Hong Kong) 2004 Dec;12(2):239-242 [FREE Full text] [doi: 10.1177/230949900401200220] [Medline: 15621915]
32. Cook C, Brown C, Isaacs R, Roman M, Davis S, Richardson W. Clustered clinical findings for diagnosis of cervical spine myelopathy. J Man Manip Ther 2010 Dec;18(4):175-180 [FREE Full text] [doi: 10.1179/106698110X12804993427045] [Medline: 22131790]
33. Cook C, Roman M, Stewart KM, Leithe LG, Isaacs R. Reliability and diagnostic accuracy of clinical special tests for myelopathy in patients seen for cervical dysfunction. J Orthop Sports Phys Ther 2009 Mar;39(3):172-178. [doi: 10.2519/jospt.2009.2938] [Medline: 19252263]
34. Atroshi I, Gummesson C, Johnsson R, Ornstein E, Ranstam J, Rosén I. Prevalence of carpal tunnel syndrome in a general population. JAMA 1999 Jul 14;282(2):153-158. [doi: 10.1001/jama.282.2.153] [Medline: 10411196]
35. Bland JDP, Rudolfer SM. Clinical surveillance of carpal tunnel syndrome in two areas of the United Kingdom, 1991-2001. J Neurol Neurosurg Psychiatry 2003 Dec;74(12):1674-1679 [FREE Full text] [doi: 10.1136/jnnp.74.12.1674] [Medline: 14638888]
36. Kuroiwa T, Fujita K, Nimura A, Miyamoto T, Sasaki T, Okawa A. A new method of measuring the thumb pronation and palmar abduction angles during opposition movement using a three-axis gyroscope. J Orthop Surg Res 2018 Nov 16;13(1):288 [FREE Full text] [doi: 10.1186/s13018-018-0999-3] [Medline: 30445972]

Abbreviations

10-s: 10-second hand grip and release
AUC: area under the curve
CM: cervical myelopathy
CTS: carpal tunnel syndrome
MRI: magnetic resonance imaging
NCS: nerve conduction study
ROC: receiver operating characteristic
SVM: support vector machine

Edited by A Mavragani; submitted 22.07.22; peer-reviewed by M Tada, L Guo, S Pandey; comments to author 16.08.22; revised version received 05.09.22; accepted 15.09.22; published 03.10.22.

Please cite as:
Koyama T, Matsui R, Yamamoto A, Yamada E, Norose M, Ibara T, Kaburagi H, Nimura A, Sugiura Y, Saito H, Okawa A, Fujita K
High-Dimensional Analysis of Finger Motion and Screening of Cervical Myelopathy With a Noncontact Sensor: Diagnostic Case-Control Study
JMIR Biomed Eng 2022;7(2):e41327
URL: https://biomedeng.jmir.org/2022/2/e41327
doi:10.2196/41327
PMID:38875599

Original Paper

Telemonitoring of Home-Based Biking Exercise: Assessment of Wireless Interfaces

Aref Smiley[1*], PhD; Te-Yi Tsai[1*], MSc; Wanting Cui[1*], MSc; Irena Parvanova[1*], PhD; Jinyan Lyu[1*], MSc; Elena Zakashansky[1*], BSc; Taulant Xhakli[1*], BSc; Hu Cui[1*], MSc; Joseph Finkelstein[1*], MD, PhD

Center for Biomedical and Population Health Informatics, Icahn School of Medicine at Mount Sinai, New York, NY, United States

[*]all authors contributed equally

Corresponding Author:
Aref Smiley, PhD
Center for Biomedical and Population Health Informatics
Icahn School of Medicine at Mount Sinai
1770 Madison Avenue, 2nd Fl
New York, NY, 10035
United States
Phone: 1 212 659 9596
Email: aref.smiley@gmail.com

Abstract

Background: Telerehabiliation has been shown to have great potential in expanding access to rehabilitation services, enhancing patients' quality of life, and improving clinical outcomes. Stationary biking exercise can serve as an effective aerobic component of home-based physical rehabilitation programs. Remote monitoring of biking exercise provides necessary safeguards to ensure exercise adherence and safety in patients' homes. The scalability of the current remote monitoring of biking exercise solutions is impeded by the high cost that limits patient access to these services, especially among older adults with chronic health conditions.

Objective: The aim of this project was to design and test two low-cost wireless interfaces for the telemonitoring of home-based biking exercise.

Methods: We designed an interactive biking system (iBikE) that comprises a tablet PC and a low-cost bike. Two wireless interfaces to monitor the revolutions per minute (RPM) were built and tested. The first version of the iBikE system uses Bluetooth Low Energy (BLE) to send information from the iBikE to the PC tablet, and the second version uses a Wi-Fi network for communication. Both systems provide patients and their clinical teams the capability to monitor exercise progress in real time using a simple graphical representation. The bike can be used for upper or lower limb rehabilitation. We developed two tablet applications with the same graphical user interfaces between the application and the bike sensors but with different communication protocols (BLE and Wi-Fi). For testing purposes, healthy adults were asked to use an arm bike for three separate subsessions (1 minute each at a slow, medium, and fast pace) with a 1-minute resting gap. While collecting speed values from the iBikE application, we used a tachometer to continuously measure the speed of the bikes during each subsession. Collected data were later used to assess the accuracy of the measured data from the iBikE system.

Results: Collected RPM data in each subsession (slow, medium, and fast) from the iBikE and tachometer were further divided into 4 categories, including RPM in every 10-second bin (6 bins), RPM in every 20-second bin (3 bins), RPM in every 30-second bin (2 bins), and RPM in each 1-minute subsession (60 seconds, 1 bin). For each bin, the mean difference (iBikE and tachometer) was then calculated and averaged for all bins in each subsession. We saw a decreasing trend in the mean RPM difference from the 10-second to the 1-minute measurement. For the 10-second measurements during the slow and fast cycling, the mean discrepancy between the wireless interface and tachometer was 0.67 (SD 0.24) and 1.22 (SD 0.67) for the BLE iBike, and 0.66 (SD 0.48) and 0.87 (SD 0.91) for the Wi-Fi iBike system, respectively. For the 1-minute measurements during the slow and fast cycling, the mean discrepancy between the wireless interface and tachometer was 0.32 (SD 0.26) and 0.66 (SD 0.83) for the BLE iBike, and 0.21 (SD 0.21) and 0.47 (SD 0.52) for the Wi-Fi iBike system, respectively.

Conclusions: We concluded that a low-cost wireless interface provides the necessary accuracy for the telemonitoring of home-based biking exercise.

(JMIR Biomed Eng 2022;7(2):e41782) doi:10.2196/41782

KEYWORDS

telerehabilitation; wireless interface; remote cycling; home-based exercise

Introduction

Telerehabilitation overcomes the barriers of distance and time using telecommunications and enables remote delivery of health care services from clinicians to patients' homes. Advances in telerehabilitation technology has brought about a possibility of a remotely supervised rehabilitation program through real-time communication, tracking of physical activities of a patient, monitoring vital body functions, and rehab physical therapy [1-3]. Telemedicine use has grown significantly during the COVID-19 pandemic [4] when the World Health Organization encouraged physical distancing [5]. Telehealth approaches may be instrumental for supporting home-based cycling exercises in people with chronic health conditions especially when safety and adherence with exercise prescription can be monitored in real time [6]. Existing solutions are primarily designed for healthy athletes and are characterized by high cost and lack of functionality that allows health professionals to monitor cycling exercise and provide feedback to patients in a timely fashion. Limited research has been conducted on low-cost wireless interfaces for the real-time remote monitoring of cycling exercise in a telerehabilitation setting that promotes upper and lower limb rehabilitation.

Cycling exercise equipment is often used to facilitate training of the upper and lower extremities, and is widely available in rehabilitation facilities that can oversee patient exercise [7]. Cycling exercise training was shown to improve clinical outcomes in patients in hemodialysis [8], patients in recovery stage after hip fracture [9], patients with mechanical ventilation [10], patients with acute recovery stage after stroke [11], patients with chronic pulmonary disease [12], patients with chronic health conditions [13], patients with Parkinson disease [14], hemiparetic patients [15], patients with COVID-19 [16], and in-bed critically ill patients [17]. Patients' engagement in telerehabilitation could be promoted using gaming and consumer appliances [18-22], where patients get motivation to engage in enjoyable play behavior that involves useful therapy-related activities [23].

Cycling exercise equipment is widely available at homes at low cost; however, lack of remote connectivity with a team of rehabilitation professionals to monitor exercise progress in real time makes it far from being effective in practice. Providing simple real-time visualizations and numerical expressions in addition to designing an alert system, preventing exertion levels exceeding those approved by a rehabilitation team, would highly improve the remote exercise effectiveness and safety. We previously showed a high level of system acceptance by the study participants for the initial iBikE design system [24]. The data obtained from this study provided the basis for the development and testing of optimal customized telerehabilitation programs.

The goal of this study was to design and test the accuracy of two low-cost wireless interfaces for the telemonitoring of home-based cycling exercise systems. The major difference of the two systems was the communication protocol. The first system used Bluetooth Low Energy (BLE) to communicate between the iBikE and tablet PC. The second system used Wi-Fi to communicate between the iBikE and tablet PC.

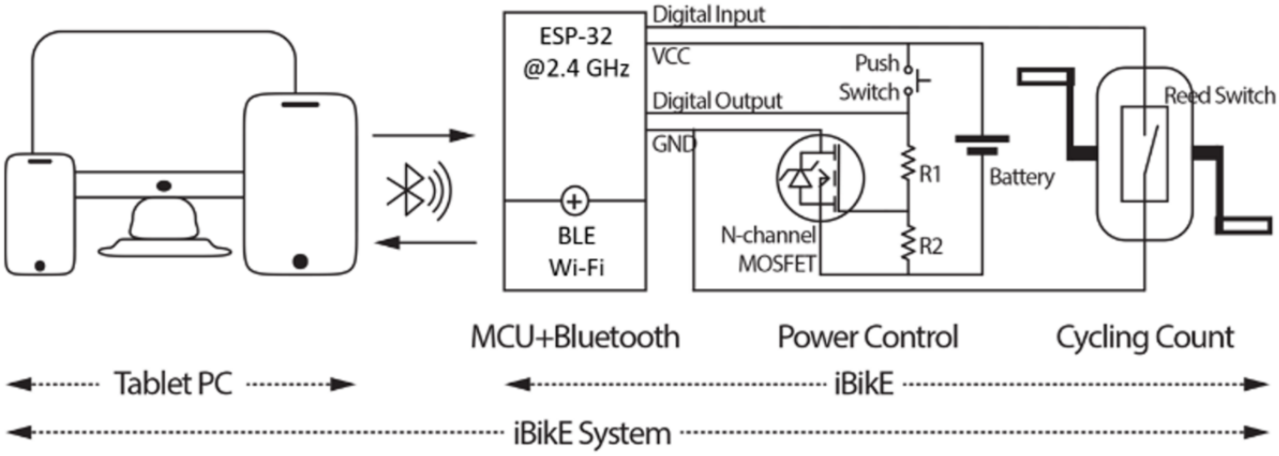
Methods

Overview

We developed two iBikE systems that use either Wi-Fi or BLE protocol to communicate between the tablet PC and iBikE. To evaluate and assess the accuracy and functionality of the systems, we tested both systems in two separate experiments. During each experiment, we used a laser tachometer to measure the bike speed. A total of 9 healthy individuals were asked to hand cycle in each experiment. Finally, collected data from the systems were compared to their representative collected data from the laser tachometer.

We used the same hardware and user interface for both designs. The iBikE system has two main parts: (1) a tablet PC application and (2) an iBikE (Figure 1). Recorded data were sent through wireless communications from built-in systems in the iBikE to the PC tablet. In addition, data were provided to telerehabilitation users when they were engaged in cycling exercises. To make it easy to use, the iBikE itself had only one physical button and the tablet PC had a touch screen button to start/stop the exercise session.

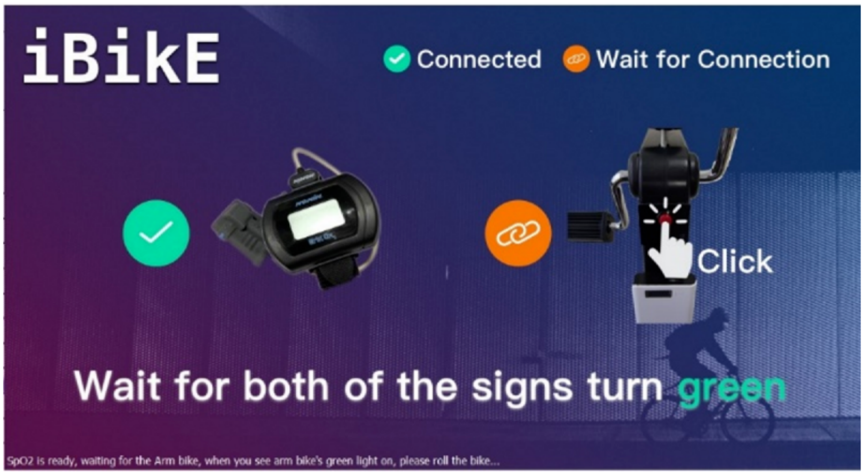
Figure 1. System design. Both developed systems have the same hardware and user interface. The major difference between the two systems is the communication protocol. They communicate through either Wi-Fi or BLE. BLE: Bluetooth Low Energy; MCU: microcontroller unit; VCC: Voltage Common Collector; GND: ground; MOSFET: metal–oxide semiconductor field-effect transistor.



Tablet PC Application

We developed a touchscreen-operated application on a tablet PC using Universal Windows Platform, which can be run on any Windows 10 operating system. We used C# programming language to develop the application. Two buttons were designed in the first page of the application, Pulse Oximeter and iBikE buttons. To display oxygen saturation and pulse rate data while the user was cycling, the user needed to push the Pulse Oximeter button to connect the wrist-worn pulse oximeter. We used the WristOx2, model 3150 [25], which connects to the tablet PC via BLE. By pressing the Pulse Oximeter image, the application first scans and then pairs the pulse oximeter through the BLE to the tablet PC. This turns its representative device button on the application to green (Figure 2). The application uses a standard universally unique identifier (UUID) to get the characteristics and their values. Successful pairing of the microcontroller unit (MCU) of the wireless interface with the tablet PC allows the user to start an exercise session and to track biking progress in real time. In addition, it allows the received data from the pulse oximeter to be displayed on the exercise page (Figure 3). Before starting the cycling session, the user had to activate the pulse oximeter by clicking on its image in the first page of the application. To start the cycling session, users needed to first push the physical button on the bike to turn on the iBikE equipment. Next, they needed to connect the tablet to the iBikE equipment by pressing its representative button on the application (Figure 2). This resulted in the connecting of the application to the iBikE through either Bluetooth or Wi-Fi and took the user to the second page of the application (Figure 3). On the second page, by starting the exercise, the real-time exercise feedback interface via the tablet PC was activated, and the user could monitor the real-time speed in revolutions per minute (RPM). In addition, the real-time pulse rate and oxygen saturation data were sent from the oximeter to the tablet and could be monitored on the second page. The user’s entire exercise data was stored on the local server and in .csv files.

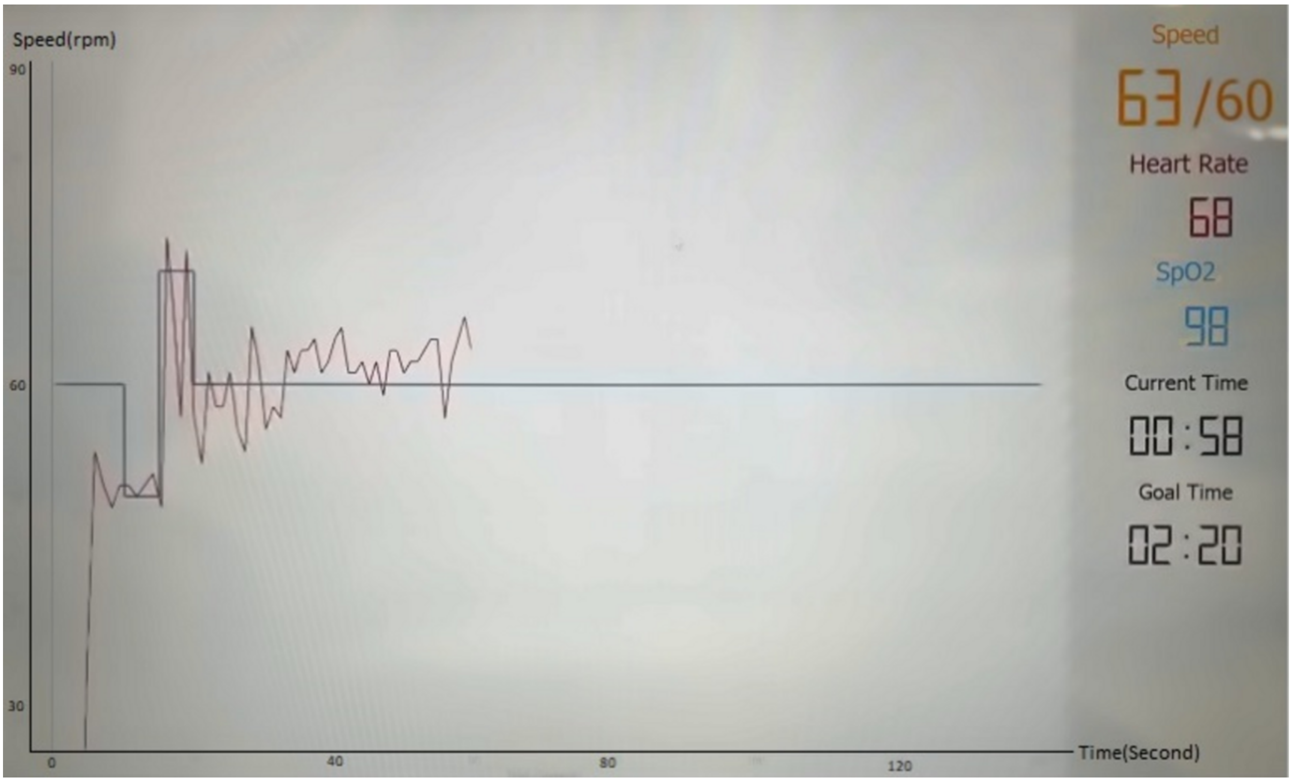
Figure 2. User interface. User needed to push a button to pair the application to the oximeter. They also needed to push the button on the right to pair the application to the iBikE equipment to start the cycling session.



Wait for both of the signs turn green

SpO2 is ready, waiting for the Arm bike, when you see arm bike's green light on, please roll the bike...

Figure 3. Real-time monitoring of the speed (rpm) versus time (second) during cycling. Information taken from the oximeter, including pulse rate and oxygen saturation, were also displayed during the cycling sessions. The image was taken with a camera and filtered for better illustration. rpm: revolutions per minute.



iBikE Equipment

There were three main modules in the iBikE exercise equipment (Figure 1): (1) control and computing module, (2) magnetic switch module (reed sensor), and (3) either Bluetooth or Wi-Fi communication module.

Control and Computing Module

In this project, we used FireBeetle ESP-32, which is a low-power consumption MCU designed for Internet of Things projects. It integrated a Dual-Core ESP-WROOM-32 module, which supports MCU and Wi-Fi and Bluetooth dual-mode communication. It also supports a 3.7V external lithium battery power supply. The iBikE first read the information provided by the reed switch sensor through the ESP32 MCU to obtain relevant cycle information and then sent the data to the tablet PC through either Bluetooth or Wi-Fi. Real-time cycling information, the intervals (milliseconds), was measured through the calculation algorithm in the MCU. The intervals were then delivered to the tablet PC via either the Wi-Fi or Bluetooth communication module. The reed switch was used to measure the time required to complete a single cycling of the iBikE.

Magnetic Switch Module

The reed sensor (magnetic switch) is a sensor that converts magnetic field changes into electrical signals. It consists of a pair of ferromagnetic flexible metal contacts, normally open, in a sealed glass envelope. By holding the magnet to the magnetic switch of the iBikE equipment, the magnetic switch detects the change in the magnetic field and closes the metal contacts, resulting in the flow of an electrical signal. By removing the magnet from the magnetic field, the magnetic switch detects the change and stops the flow of the electrical signal. In the iBikE equipment, every rotation of the pedals (one cycle) was detected by the magnetic switch. Continuous sampling of the electrical signal flow from the switch by the MCU provided cycling time intervals. The user's departure from the iBikE was detected if there was no detection of on/off changes in the sampled electrical signal for more than 90 seconds.

Bluetooth or Wi-Fi Communication Module

We developed two separate iBikE systems to transmit cycling intervals from the iBikE system to the PC tablet using either Wi-Fi or BLE communication protocol in each system. In BLE communication mode, the PC tablet was directly connected to the MCU integrated inside the iBikE equipment. Cycle intervals were first converted into two bytes: low byte (the least significant part of an integer) and high byte (the most significant part of an integer). The byte packets were then sent to the tablet PC. We defined our custom UUID to transfer values from the MCU to the tablet PC. In Wi-Fi communication mode, both the PC tablet and the MCU were connected to the same Wi-Fi network. User Datagram Protocol was used to send cycling intervals from the bike, configured as client, to the PC tablet, configured as server.

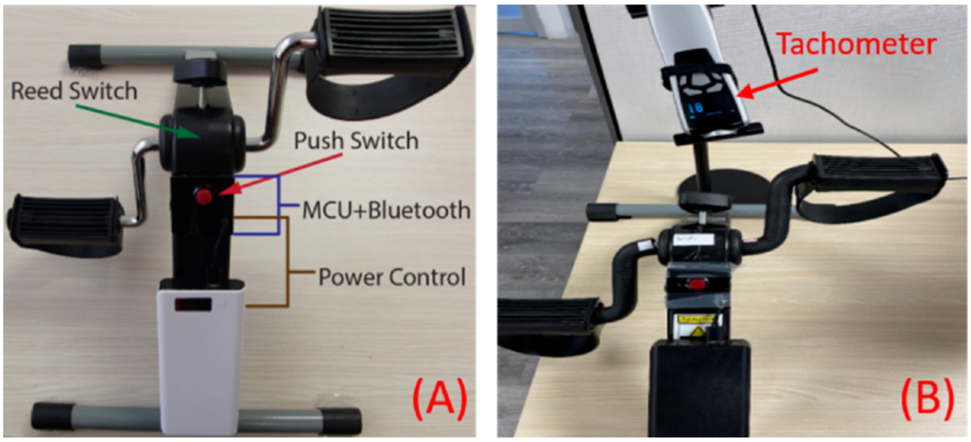
Data Collection

Both developed iBikE systems had the same operating procedure. To start the cycling session, the user was required to first press the physical red button on the iBikE exercise equipment to wake up the MCU and to activate either the Wi-Fi or Bluetooth communication with the tablet PC. The user then needed to press the button on the first page of the application

(Figure 2) to indicate the beginning of the exercise. This allowed the tablet PC to connect to the Wi-Fi/Bluetooth communication module and pair with the iBikE exercise equipment. After pairing the tablet PC to the iBikE, the prescribed cycling speeds appeared on the second page of the interface (Figure 3), where the user could exercise via the iBikE exercise equipment and receive feedback on the user's exercise.

The accuracy of each iBikE system was checked using a laser tachometer, DT-2100, Nidec-SHIMPO [26]. The tachometer was used in noncontact continuous measurement mode to detect the measured RPM in real time (Figure 4B). The tachometer samples the continuous RPM measurements with 10 Hz sampling frequency. Collected data could be visualized in its PC software in real time. At the end of each session, recorded data could be saved in an Excel (Microsoft Corporation) file. Finally, the saved data during each cycling experiment were compared to their representative collected data from the iBikE system.

Figure 4. (A) Exercise equipment. The user interface is only a physical button to make it easy to use the equipment. All the components, including the reed switch, push switch, MCU and Bluetooth Low Energy modules, and power control module are inside the iBikE equipment. (B) We used a laser tachometer, DT-2100, Nidec-SHIMPO [15], in noncontact continuous measurement mode to detect the measured revolutions per minute in real time to later compare the results with data taken from the iBikE equipment. MCU: microcontroller unit.



A single group of 9 individuals performed 3 sessions of hand cycling for each iBikE system, with a 1-minute duration for each session. They started with 1 minute of hand cycling at a slow pace, followed by 1 minute of rest. They then started with 1 minute of hand cycling at a medium pace, followed by 1 minute of rest. Finally, they started with 1 minute of hand cycling at a fast pace. The meaning of slow, medium, and fast was based on the user interpretation. Participants in the tests were 9 healthy individuals, aged between 21 and 37 years. They performed the tests on 2 different days. Data were collected from the iBikE system with the BLE communication mode on the first day, and data were collected from the iBikE system with Wi-Fi communication mode on the second day. Users completed a total of 18 sessions (54 phases). Data were collected from both the iBikE system and tachometer in parallel during each session. The iBikE system recorded completed cycling intervals and sent the data to the PC tablet. At the same time, the tachometer collected RPM values and sent them to a PC with a sampling rate of 10 Hz.

In addition to cycling information, which was sent by the MCU to the application, other information, including heart rate, SpO2 (oxygen saturation level), and Personal Activity Intelligence [27], was sent from the pulse oximeter to the application.

Ethical Considerations

As there was no risk and the participants were all authors of this paper, we did not require institutional review board approval. Data were collected between May 2022 and June 2022. No protected health information were collected, and the resulting analytical data set was fully deidentified. No compensation was provided to the study participants.

Results

The iBikE system accuracy was evaluated by comparing the collected data from the iBikE and its representative collected data from the tachometer for each session. We developed an algorithm using MATLAB R2022a (MathWorks) to compare the two collected data sets. The first step in the algorithm was to shift the collected data from the iBikE to match its representative collected data from the tachometer. This was done for all 54 (1 minute) collected data at slow, medium, and fast pace (Figures 5-7). All the figures belonged to the same individual, cycling with various paces in three sessions in 1-minute durations.

In the collected data at a slow pace, we had less samples collected from the iBikE compared to the medium- and fast-paced sessions for the same individual. This is because the iBikE sends the intervals (in milliseconds) whenever each cycle is completed. In the slow-paced session, it takes longer for the user to complete a cycle, and therefore, less samples are collected by the iBikE system for each session with the same duration (60 seconds). However, in the tachometer, the data collection sampling rate is 10 Hz in continuous mode for all sessions. Therefore, we had 600 collected samples for each 60-second session. To compare the data from the iBikE and its representative collected data from the tachometer, data from the tachometer was averaged to be the same size to its representative collected data from the iBikE (Figures 5-7).

Figure 5. Measured RPM from the iBikE system versus the tachometer in a 1-minute session at a slow pace. RPM: revolutions per minute.

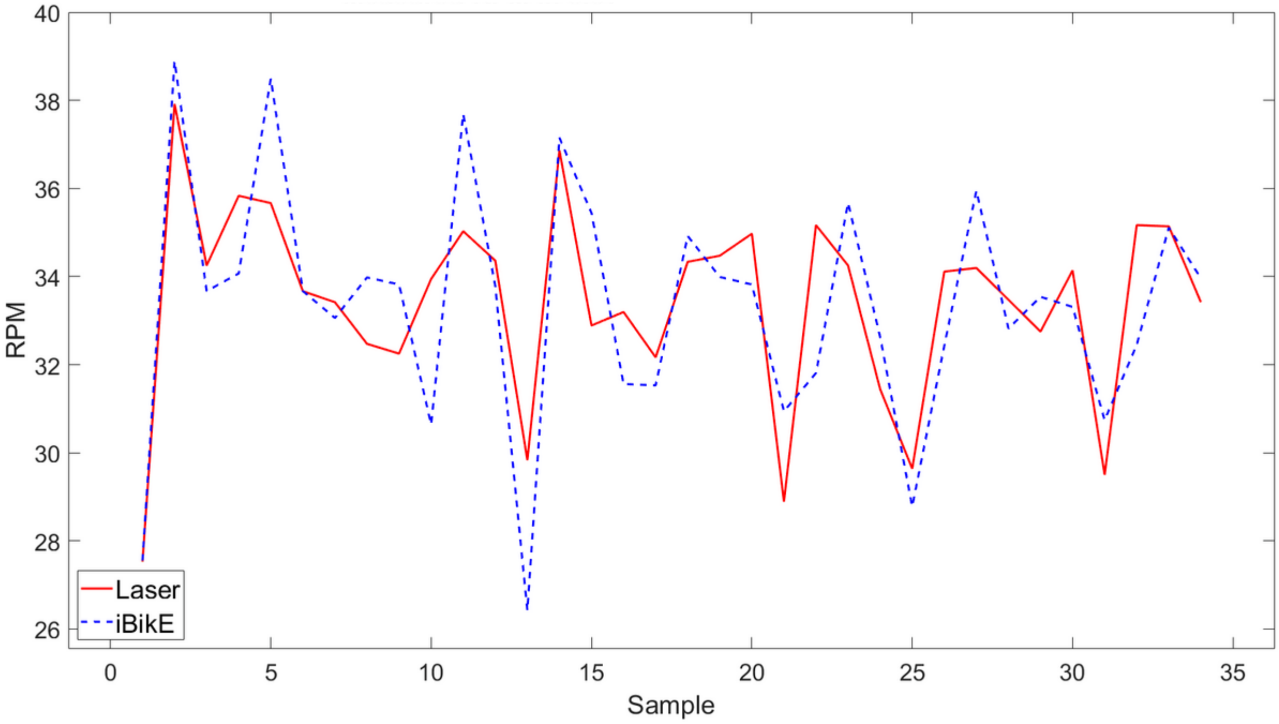


Figure 6. Measured RPM from the iBikE system versus a tachometer in 1-minute sessions at a medium pace. RPM: revolutions per minute.

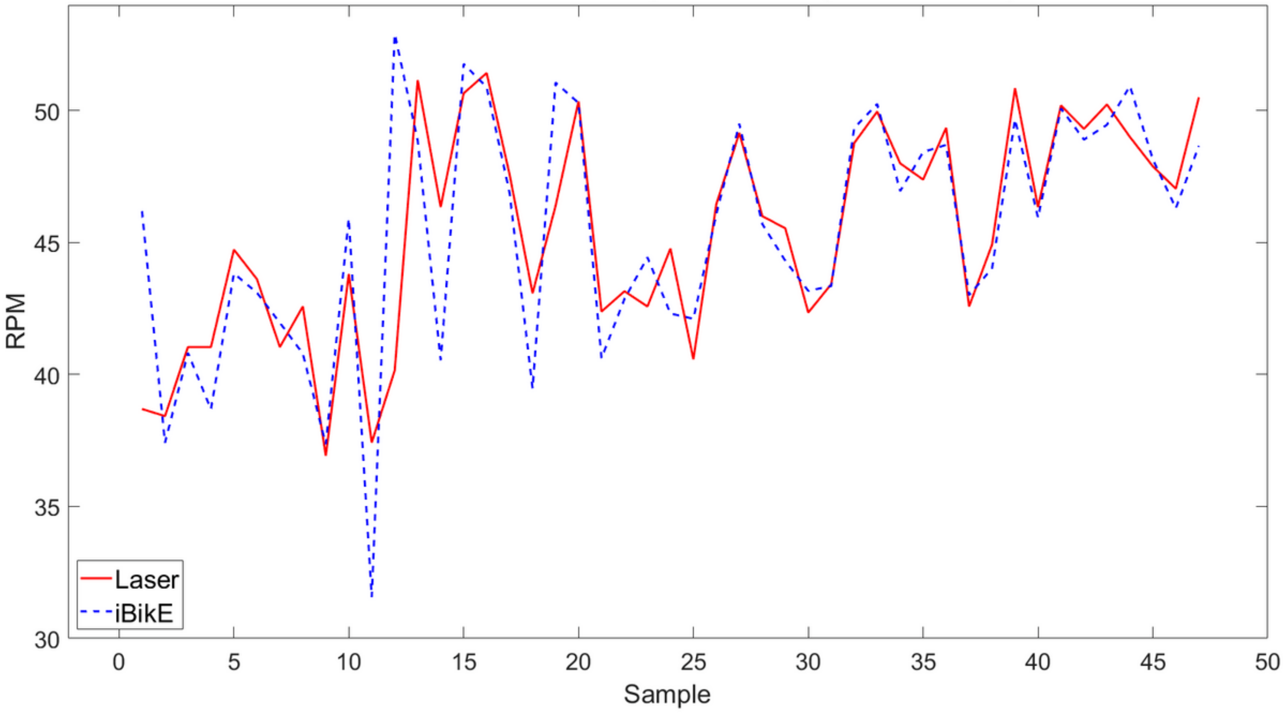
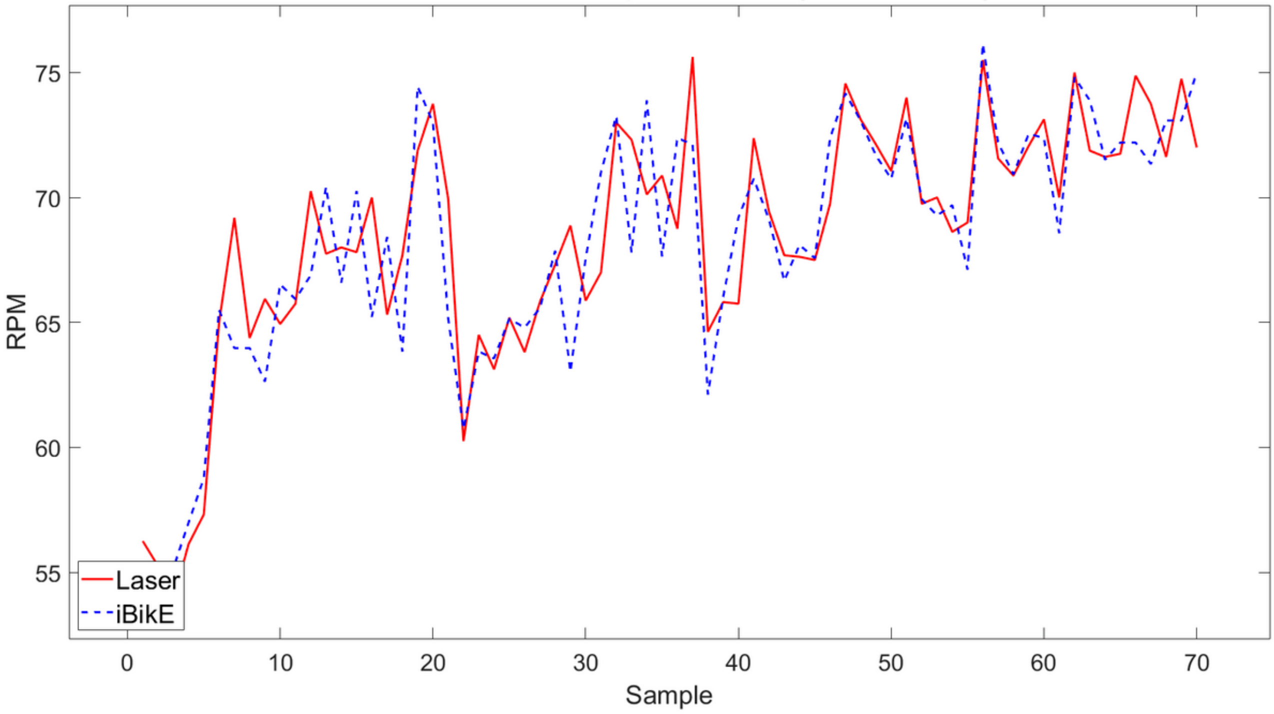


Figure 7. Measured RPM from the iBikE system versus a tachometer in 1-minute sessions at a fast pace. RPM: revolutions per minute.



Shifting collected data from the iBikE and matching its representative collected data from the tachometer allowed us to determine the starting and ending points of each session. The second step in the algorithm was to divide each session (60 seconds) into 4 subsessions: 10 seconds, 20 seconds, 30 seconds, and 60 seconds for both collected data from the tachometer and the iBikE systems. In a 10-second subsession, for example, both collected data from the iBikE and tachometer in each session (with 60-second durations) was divided into 6 bins of 10 seconds. The mean of the RPM were then calculated for every 10-second bin. The mean difference was then calculated for each bin (iBikE and its representative tachometer bin). Finally, the mean value (of the 6 bins) was calculated. Table 1 shows an example of the calculated mean for all subsections for one of the participants.

Finally, the mean and SD of all the similar subsections for all participants were calculated and are shown in Tables 2 and 3 and Figure 8.

In each session, the mean and SD of the subsections showed a downtrend from 10-second measured means to 60-second ones. The trend was the same for all the sessions for both Wi-Fi and BLE systems. The maximum calculated means of subsections for each person’s collected data were 2.96 for Wi-Fi and 2.69 for BLE, both in 10 seconds of collected data at a fast pace. The minimum mean of the subsections for all participants’ collected data was 0.2 (SD 0.3) in the 60-second subsession at medium speed. For the Wi-Fi iBikE system, this number was 0.21 (SD 0.21) in the 1-minute subsession at slow speed.

Table 1. Example of calculated mean difference for each subsection (10 seconds, 20 seconds, 30 seconds, and 60 seconds) in each session (slow, medium, fast) for both the BLE and Wi-Fi iBikE systems for one of the participants. Each section is divided into 6 bins, 3 bins, 2 bins, and 1 bin. The difference of the revolutions per minute means is then calculated for each bin. The mean of the bins in each subsection is reported.

Subsections	BLE[a] sections			Wi-Fi sections		
	Slow	Medium	Fast	Slow	Medium	Fast
10-s bins (6 bins), mean	0.69	0.87	0.95	0.53	0.53	0.7
20-s bins (3 bins), mean	0.38	1.14	0.7	0.27	0.5	0.37
30-s bins (2 bins), mean	0.42	1.13	0.52	0.13	0.1	0.39
60-s bin (1 bin), mean	0.17	0.1	0.08	0.15	0.07	0.15

[a] BLE: Bluetooth Low Energy.

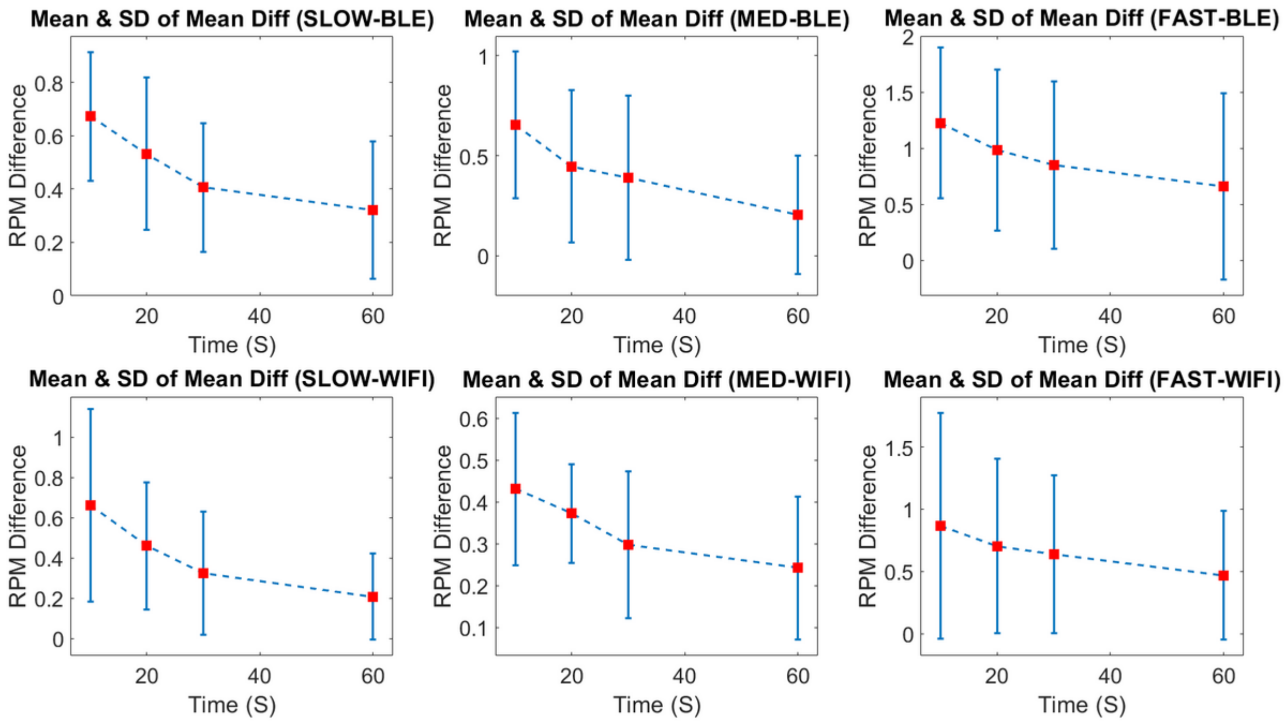
Table 2. Mean and SD of all the collected differences between the wireless interface and tachometer in each subsection with similar bins for the Bluetooth Low Energy iBikE system.

Subsections	Slow pace sessions, mean (SD)	Medium pace sessions, mean (SD)	Fast pace sessions, mean (SD)
10-s bins	0.67 (0.24)	0.65 (0.37)	1.22 (0.67)
20-s bins	0.53 (0.29)	0.45 (0.38)	0.98 (0.72)
30-s bins	0.41 (0.24)	0.39 (0.41)	0.85 (0.75)
60-s bins	0.32 (0.26)	0.20 (0.30)	0.66 (0.83)

Table 3. Mean and SD of all the collected differences between the wireless interface and tachometer in each subsection with similar bins for the Wi-Fi iBikE system.

Subsections	Slow pace sessions, mean (SD)	Medium pace sessions, mean (SD)	Fast pace sessions, mean (SD)
10-s bins	0.66 (0.48)	0.43 (0.18)	0.87 (0.91)
20-s bins	0.46 (0.32)	0.37 (0.12)	0.70 (0.70)
30-s bins	0.32 (0.31)	0.30 (0.17)	0.64 (0.63)
60-s bins	0.21 (0.21)	0.24 (0.17)	0.47 (0.52)

Figure 8. Means and SDs of the calculated mean differences in the 10-, 20-, 30-, and 60-second subsessions for all individuals for both the Wi-Fi and Bluetooth Low Energy iBikE systems. RPM: revolutions per minute.



Discussion

Principal Results

We designed and developed two systems that could be used as in-home exercise for patients requiring a telerehabilitation exercise system. We designed the system in a way that was easily accessible and required minimal experience from users to operate the system. Our new systems are low-cost and showed their capability and accuracy in communicating through either BLE or Wi-Fi. We evaluated our iBikE systems' accuracy and functionality by comparing their recorded data with the collected data from the laser tachometer. The results showed that the minimum mean RPM difference was 0.2 (SD 0.3) for 1 minute of hand cycling at a medium-paced session for the BLE iBikE system. For the Wi-Fi iBikE system, the minimum mean RPM difference was 0.21 (SD 0.21) for 1 minute of hand cycling at a slow-paced session.

We made the user interface as simple as possible for the user to work with the iBikE system and to start/stop a session. By pushing a single physical button on the iBikE equipment (Figure 3), both systems easily connect to the PC tablet. This activated the BLE/Wi-Fi connection in the MCU and then waited for the user to push a button on the tablet PC screen to confirm the connection and to start the cycling session. The same button on the iBikE equipment could be used to stop the cycling session.

Limitations

We used FireBeetle ESP-32, with integrated Dual-Core ESP-WROOM-32 module, which supports MCU and Wi-Fi and Bluetooth dual-mode communication. To accurately record the cycling intervals from the reed switch on falling edges, we used interrupt service routine (ISR) in our detection algorithm. Due to the capacitive structure of the reed switch [28] and external magnetic field effect [29], random falling edges (*false* interrupts) for every ~250 milliseconds were detected in the Wi-Fi iBikE system. For the BLE iBikE system, *false* interrupts happened every ~140 milliseconds. When the system battery got below its ~70% capacity, the effect of detecting *false* interrupts was more frequent and longer, ~180 milliseconds and ~300 milliseconds for the BLE and Wi-Fi systems, respectively. Therefore, the cycling interval detection algorithm used a hold off timer in the ISR to prevent the counter from incrementing on the *false* interrupts. This was one of the advantages of the BLE system over the Wi-Fi system, as the hold off timer was smaller (less than 200 ms) and, as a result, could detect faster RPM (up to 300) compared to the Wi-Fi system. In addition, the BLE iBikE system used less battery during similar sessions. The highest RPM value recorded during the fast-paced sessions was 142 RPM. Therefore, both systems could detect all the cycling intervals with no missing data. However, the issue should be considered for similar applications, where data need to be recorded faster with higher frequency.

To activate the Wi-Fi communication in the ESP-32 MCU, we needed to set the username and password of an active router nearby the iBikE to be connected to Wi-Fi and to communicate with the PC tablet. If the user moved to a new location with a new active router, the MCU embedded programing code needed to be updated with the new username and password. This requires an experienced person with special integrated development environment software (Arduino IDE, etc) to update the code with the new username and password. In addition, we needed to provide the IP address of the tablet PC in the programing code to communicate to the tablet PC. However, the BLE system did not have these limitations. Our suggestions for using the BLE iBikE system over the Wi-Fi system are limited to this developed system, our application's needs, and the system's data results. Both BLE and Wi-Fi technologies have their strong and weak sides. When the application requires a big data transfer with faster speed and with high security, for example, Wi-Fi is a better choice.

Conclusions and Future Work

The iBikE system consisted of a sensor for sampling fast cycles, detecting algorithms for the patient's exit from the program, programs for collecting sensor data and communicating with a tablet PC, user interfaces for displaying the iBikE and patient data and entering control variables, and follow-up records and data collection systems. A low-cost wireless interface provided the necessary accuracy for the telemonitoring of home-based biking exercises. This iBikE system is novel because it can capture and communicate real-time exercise cycling data for a rehabilitation team using a reliable low-cost wireless interface.

In previous work, we demonstrated high acceptance and positive impact of physical telerehabilitation in patients with chronic pulmonary conditions [30], multiple sclerosis [31], and geriatric syndromes [32]. The next step is to evaluate the impact of home-based cycling telerehabilitation programs that include real-time exercise monitoring and data-driven feedback from rehabilitation teams in a clinical trial setting, which includes a diverse spectrum of patients enrolled for a prolonged period of time. Short-term and long-term cycling effects should be investigated and compared to the existing rehabilitation outcomes implemented in outpatient clinics. Telerehabilitation programs will be assessed by their impact on aerobic fitness, clinical outcomes, health care use, and exercise adherence.

Acknowledgments

This project was supported in part by a grant R33HL143317 from the National Heart, Lung, and Blood Institute.

Conflicts of Interest

None declared.

References

1. Shero S, Benzo R, Cooper L, Finkelstein J, Forman DE, Gaalema DE, et al. Update on RFA increasing use of cardiac and pulmonary rehabilitation in traditional and community settings NIH-funded trials: addressing clinical trial challenges presented by the COVID-19 pandemic. J Cardiopulm Rehabil Prev 2022 Jan 01;42(1):10-14 [FREE Full text] [doi: 10.1097/HCR.0000000000000635] [Medline: 34508036]
2. Jeong I, Karpatkin H, Finkelstein J. Physical telerehabilitation improves quality of life in patients with multiple sclerosis. Stud Health Technol Inform 2021 Dec 15;284:384-388. [doi: 10.3233/SHTI210752] [Medline: 34920553]
3. Bedra M, McNabney M, Stiassny D, Nicholas J, Finkelstein J. Defining patient-centered characteristics of a telerehabilitation system for patients with COPD. Stud Health Technol Inform 2013;190:24-26. [Medline: 23823363]
4. Cui W, Finkelstein J. Impact of COVID-19 pandemic on use of telemedicine services in an academic medical center. Stud Health Technol Inform 2021 May 27;281:407-411. [doi: 10.3233/SHTI210190] [Medline: 34042775]
5. Lyu J, Cui W, Finkelstein J. Use of artificial intelligence for predicting COVID-19 outcomes: a scoping review. Stud Health Technol Inform 2022 Jan 14;289:317-320. [doi: 10.3233/SHTI210923] [Medline: 35062156]

6. Celesti A, Lay-Ekuakille A, Wan J, Fazio M, Celesti F, Romano A, et al. Information management in IoT cloud-based tele-rehabilitation as a service for smart cities: comparison of NoSQL approaches. Measurement 2020 Feb;151:107218. [doi: 10.1016/j.measurement.2019.107218]
7. Costi S, Crisafulli E, Degli Antoni F, Beneventi C, Fabbri LM, Clini EM. Effects of unsupported upper extremity exercise training in patients with COPD: a randomized clinical trial. Chest 2009 Aug;136(2):387-395. [doi: 10.1378/chest.09-0165] [Medline: 19567487]
8. Liao M, Liu W, Lin F, Huang C, Chen S, Liu C, et al. Intradialytic aerobic cycling exercise alleviates inflammation and improves endothelial progenitor cell count and bone density in hemodialysis patients. Medicine (Baltimore) 2016 Jul;95(27):e4134. [doi: 10.1097/MD.0000000000004134] [Medline: 27399127]
9. Mangione KK, Palombaro KM. Exercise prescription for a patient 3 months after hip fracture. Phys Ther 2005 Jul;85(7):676-687. [Medline: 15982174]
10. Porta R, Vitacca M, Gilè LS, Clini E, Bianchi L, Zanotti E, et al. Supported arm training in patients recently weaned from mechanical ventilation. Chest 2005 Oct;128(4):2511-2520. [doi: 10.1378/chest.128.4.2511] [Medline: 16236917]
11. Yang H, Lee C, Lin R, Hsu M, Chen C, Lin J, et al. Effect of biofeedback cycling training on functional recovery and walking ability of lower extremity in patients with stroke. Kaohsiung J Med Sci 2014 Jan;30(1):35-42 [FREE Full text] [doi: 10.1016/j.kjms.2013.07.006] [Medline: 24388057]
12. Nickel R, Troncoso F, Flores O, Gonzalez-Bartholin R, Mackay K, Diaz O, et al. Physiological response to eccentric and concentric cycling in patients with chronic obstructive pulmonary disease. Appl Physiol Nutr Metab 2020 Nov;45(11):1232-1237. [doi: 10.1139/apnm-2020-0149] [Medline: 32413271]
13. Matta T, Simão R, de Salles BF, Spineti J, Oliveira LF. Strength training's chronic effects on muscle architecture parameters of different arm sites. J Strength Cond Res 2011 Jun;25(6):1711-1717. [doi: 10.1519/JSC.0b013e3181dba162] [Medline: 21602648]
14. Ridgel AL, Peacock CA, Fickes EJ, Kim C. Active-assisted cycling improves tremor and bradykinesia in Parkinson's disease. Arch Phys Med Rehabil 2012 Nov;93(11):2049-2054. [doi: 10.1016/j.apmr.2012.05.015] [Medline: 22659536]
15. Ambrosini E, Ferrante S, Pedrocchi A, Ferrigno G, Molteni F. Cycling induced by electrical stimulation improves motor recovery in postacute hemiparetic patients: a randomized controlled trial. Stroke 2011 Apr;42(4):1068-1073. [doi: 10.1161/STROKEAHA.110.599068] [Medline: 21372309]
16. Situmorang DDB, Ifdil I, Wati CLS, Mamahit HC, Papu YM. Cycling therapy for reducing psychological problems of patients with COVID-19: as an alternative treatment after recovery. Infect Dis Clin Pract (Baltim Md) 2021 Nov;29(6):e490 [FREE Full text] [doi: 10.1097/IPC.0000000000001061]
17. Nickels MR, Aitken LM, Barnett AG, Walsham J, McPhail SM. Acceptability, safety, and feasibility of in-bed cycling with critically ill patients. Aust Crit Care 2020 May;33(3):236-243. [doi: 10.1016/j.aucc.2020.02.007] [Medline: 32317212]
18. Baranowski T, Buday R, Thompson DI, Baranowski J. Playing for real: video games and stories for health-related behavior change. Am J Prev Med 2008 Jan;34(1):74-82 [FREE Full text] [doi: 10.1016/j.amepre.2007.09.027] [Medline: 18083454]
19. Brox E, Fernandez-Luque L, Tøllefsen T. Healthy gaming - video game design to promote health. Appl Clin Inform 2011;2(2):128-142 [FREE Full text] [doi: 10.4338/ACI-2010-10-R-0060] [Medline: 23616865]
20. Wei C, Finkelstein J. Comparison of Alexa voice and audio video interfaces for home-based physical telerehabilitation. AMIA Annu Symp Proc 2022;2022:496-503 [FREE Full text] [Medline: 35854718]
21. Lieberman DA. Designing serious games for learning and health in informal and formal settings. In: Ritterfeld U, Cody M, Vorderer P, editors. Serious Games: Mechanisms and Effects. New York: Routeledge; 2009:117-130.
22. Thompson D, Baranowski T, Buday R, Baranowski J, Thompson V, Jago R, et al. Serious video games for health how behavioral science guided the development of a serious video game. Simul Gaming 2010 Aug 01;41(4):587-606 [FREE Full text] [doi: 10.1177/1046878108328087] [Medline: 20711522]
23. Liew SL, Lin DJ, Cramer SC. Interventions to improve recovery after stroke. In: Grotta JC, Broderick JP, Kasner SE, Sacco RL, Albers GW, Day AL, et al, editors. Stroke: Pathophysiology, Diagnosis, and Management. Amsterdam: Elsevier; 2021:888-899.
24. Finkelstein J, Jeong IC. Feasibility of interactive biking exercise system for telemanagement in elderly. Stud Health Technol Inform 2013;192:642-646. [Medline: 23920635]
25. WristOx2® Model 3150 with USB. Nonin. 2018. URL: https://www.nonin.com/products/3150-usb/ [accessed 2022-05-15]
26. DT-2100: combination contact / non-contact tachometer with USB output. Shimpo Instruments. URL: http://shimpoinstruments.com/product/DT-2100 [accessed 2022-05-15]
27. Nauman J, Nes BM, Zisko N, Revdal A, Myers J, Kaminsky LA, et al. Personal Activity Intelligence (PAI): a new standard in activity tracking for obtaining a healthy cardiorespiratory fitness level and low cardiovascular risk. Prog Cardiovasc Dis 2019;62(2):179-185. [doi: 10.1016/j.pcad.2019.02.006] [Medline: 30797801]
28. Jain P. Reed switch: understanding specifications. Engineers Garage. 2022. URL: https://www.engineersgarage.com/reed-switch-understanding-specifications/ [accessed 2022-05-10]
29. Teschler L. Designing with reed switches: what you need to know. Analog IC Tips. 2022. URL: https://www.analogictips.com/designing-with-reed-switches-faq/ [accessed 2022-05-10]

30. Finkelstein J, Jeong I, Doerstling M, Shen Y, Wei C, Karpatkin H. Usability of remote assessment of exercise capacity for pulmonary telerehabilitation program. Stud Health Technol Inform 2020 Nov 23;275:72-76. [doi: 10.3233/SHTI200697] [Medline: 33227743]
31. Jeong I, Karpatkin H, Stein J, Finkelstein J. Relationship between exercise duration in multimodal telerehabilitation and quality of sleep in patients with multiple sclerosis. Stud Health Technol Inform 2020 Jun 16;270:658-662. [doi: 10.3233/SHTI200242] [Medline: 32570465]
32. Finkelstein J, Wood J, Cha E. Impact of physical telerehabilitation on functional outcomes in seniors with mobility limitations. Annu Int Conf IEEE Eng Med Biol Soc 2012;2012:5827-5832. [doi: 10.1109/EMBC.2012.6347319] [Medline: 23367254]

Abbreviations

BLE: Bluetooth Low Energy
ISR: interrupt service routine
MCU: microcontroller unit
RPM: revolution per minute
UUID: universally unique identifier

Edited by A Mavragani; submitted 08.08.22; peer-reviewed by D Gaz, R Moskovitch; comments to author 13.09.22; revised version received 18.09.22; accepted 22.09.22; published 12.10.22.

Please cite as:
Smiley A, Tsai TY, Cui W, Parvanova I, Lyu J, Zakashansky E, Xhakli T, Cui H, Finkelstein J
Telemonitoring of Home-Based Biking Exercise: Assessment of Wireless Interfaces
JMIR Biomed Eng 2022;7(2):e41782
URL: https://biomedeng.jmir.org/2022/2/e41782
doi:10.2196/41782
PMID:38875588

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